ERK 2020

Portorož, Slovenija, 21. - 22. september 2020

Zbornik devetindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2020

Proceedings of the Twenty-ninth International Electrotechnical and Computer Science Conference ERK 2020

Vabljena predavanja / Invited Lectures
Elektronika / Electronics
Telekomunikacije / Telecommunications
Multimedija / Multimedia
Avtomatika in robotika / Automatic Control and Robotics
Modeliranje in simulacija / Modelling and Simulation
Močnostna elektrotehnika / Power Engineering
Merilna tehnika / Measurement - ISEMEC 2020
Akustika in elektroakustika / Acoustics and Electroacoustics
Računalništvo in informatika / Computer and Information Science
Razpoznavanje vzorcev / Pattern Recognition
Biomedicinska tehnika /Biomedical Engineering
Didaktika / Didactics
Študentski članki / Student Papers

Uredila / Edited by
Andrej Žemva, Andrej Trost

ISSN 2591-0442 (online)

IEEE
Slovenska sekcija IEEE / Slovenian Section IEEE

2591-0442 (online)

Zbornik devetindvajsete Elektotehniške in računalniške konference ERK 2020,
21. - 22. september 2020, Portorož, Slovenija

Proceedings of the Twenty-ninth Electrotechnical and Computer Science Conference ERK 2020,
21 - 22 September 2020, Portorož, Slovenia

Vabljena predavanja
Elektronika
Telekomunikacije
Avtomatika in robotika
Modeliranje in simulacija
Močnostna elektronika
Merilna tehnika - (ISEMEC)
Akustika in elektroakustika
Računalništvo in informatika / Computer and Information Science
Razpoznavanje vzorcev
Biomedicinska tehnika
Didaktika
Študentski članki

Pri organizaciji Elektrotehniške in računalniške konference so sodelovala naslednja društva:

Društvo avtomatikov Slovenije,
Slovensko društvo za merilno-procesno tehniko (ISEMEC),
SLOKO-CIGRE,
Društvo za medicinsko in biološko tehniko Slovenije,
Slovensko društvo za umetno inteligenco,
Slovensko društvo za razpoznavanje vzorcev,
Slovensko društvo za simulacije in modeliranje.



Slovenska sekcija IEEE
Fakulteta za elektrotehniko v Ljubljani

Sporočilo predsednika konference

v imenu Organizacijskega odbora imam prijetno dolžnost, da Vas pozdravim na elektrotehniški in računalniški konferenci ERK, ki je postala tradicionalno srečanje strokovnjakov, ne le na elektrotehniškem in računalniškem področju, temveč tudi na drugih področjih, kjer sta vključena elektrotehnika in računalništvo. Letos v prav posebnih razmerah poteka že 29. Mednarodna elektrotehniška in računalniška konferenca ERK 2020, 21. in 22. septembra 2020, v hotelu Bernardin, v Portorožu v Sloveniji.

Prek 110 člankov smo uvrstili v program konference in jih razdelili po številnih strokovnih področjih. Letos nas bo s svojim obiskom in vabljenim predavanjem o problematiki, ki je tako močno zaznamovala leto 2020, počastil tudi izr. prof. dr. Jure Leskovec z Univerze Stanford, ZDA. Naslov njegovega vabljenega predavanja je *Modeling the spread of Coronavirus with mobility networks: Explaining infection rates and informing reopening strategies.*

Vzporedno s konferenco ERK bo letos potekala tudi konferenca ICMI s področja strojne inteligence, ki služi kot forum za izmenjavo izkušenj, dobrih praks in priložnosti med Fakulteto za elektrotehniko Univerze v Ljubljani in japonsko tehniško univerzo Kyutech. Kolegi iz Slovenije in Japonske bodo predstavili svoje raziskovalno in projektno delo, in tako poglobili medsebojno sodelovanje z namenom vzpostavljanja novih skupnih raziskav in projektov. Konferenca letos poteka po hibridnem modelu z uporabo multimedijskih in telekomunikacijskih tehnologij.

Konferenca ERK je pogosto prvi forum, kjer mladi avtorji predstavijo rezultate svojega dela širšemu krogu. Mnogi svetovno uveljavljeni slovenski raziskovalci so imeli svoje prve nastope prav na tej konferenci. Na konferenci poteka vsako leto tudi študentsko tekmovanje Slovenske sekcije IEEE. Najboljša študentska prispevka z Univerze v Mariboru in Univerze v Ljubljani se uvrstita na nadaljnje regijsko tekmovanje IEEE R8, ki pokriva celotno področje Evrope, večji del Azije in Afrike. Na dosedanjih regijskih tekmovanjih so slovenski študentje že petkrat osvojili prvo mesto.

Program konference smo razdelili na 12 strokovnih področij, poleg vabljenega predavanja in šestih študentskih prispevkov, je v posameznih sekcijah razvrščenih 105 predstavitev. V devetindvajsetih letih je bilo predstavljenih 174 vabljenih predavanj, 283 študentskih prispevkov in 5328 rednih prispevkov.

Kljub intenzivnem delovanju v mednarodnem okolju, še naprej razvijamo slovensko tehniško besedo, zato se slovenskemu jeziku na konferenci nismo odpovedali.

Razpored referatov in vsi prispevki so dostopni na spletni strani
https://erk.fe.uni-lj.si/2020/program.php.

Andrej Žemva
Predsednik konference

Andrej Trost
Predsednik programskega odbora

Organizatorji / Organizers

- Slovenska sekcija IEEE
- Fakulteta za elektrotehniko, Univerza v Ljubljani

Strokovna društva

Pri delu posameznih konferenčnih sekcij sodelujejo naslednja strokovna društva:

- Društvo avtomatikov Slovenije,
- Slovensko društvo za merilno-procesno tehniko (ISEMEC 2020),
- SLOKO–CIGRE,
- Društvo za medicinsko in biološko tehniko Slovenije,
- Slovensko društvo za umetno inteligenco,
- Slovensko društvo za razpoznavanje vzorcev,
- Slovensko društvo za simulacijo in modeliranje.

Konferenčni odbori / Board

Predsednik konference / Conference Chairman
Andrej Žemva

Podpredsednik konference / Conference Vicechairman
Matej Zajc

Častni predsednik konference / Honorary Conference Chairman
Baldomir Zajc

Predsednik odbora za tisk / Publications Chairman
Dane Seliger

Predsednik program. odbora / Program Committee Chairman
Andrej Trost

Programski odbor / Programe Committee

Dušan Agrež	Jože Guna	Vili Podgorelec
Vanja Ambrožič	Iztok Humar	Matevž Pogačnik
Maja Atanasijević-Kunc	Franc Jager	Franc P.-Antoncich
Tadej Bajd	Marko Jagodič	Igor Pušnik
Boštjan Batagelj	Tomaž Javornik	Kurt Richter
Genevieve Baudoin	Roman Kamnik	Radovan Sernec
Sašo Blažič	Gregor Klančar	Danijel Skočaj
Zmago Brezočnik	Maharatna Koushik	Franc Solina
Franc Brglez	Igor Kuzle	Janez Sterle
Patricio Bulić	Roman Kužnar	Nermin Suljanović
Urban Burnik	Andrea M. Tonello	Igor Škrjanc
Emilio Corchado	Alenka Maček-Lebar	Sebastijan Šprager
Matjaž Debevc	Rok Mandeljc	Vitomir Štruc
Carl Debono	Blaž Meden	Fabio Tosato
Marko Delimar	Damijan Miklavčič	Aleš Ude
Simon Dobrišek	Miro Milanovič	Anton Umek
Janko Drnovšek	Marko Meža	Danijel Vončina
Žiga Emeršič	Josef Modelski	Matej Zajc
Iztok Fajfar	Aljo Mujčić	Aleš Zamuda
Peter Farkaš	David Nedeljković	Damjan Zazula
Hubert Frohlich	Nataša Nešković	Borut Zupančič
Matjaž Gams	Igor Papič	Borut Žalik
Gregor Geršak	Gregor Dolinar	Andrej Žemva
Bernard Grum	Janez Perš	

Recenzenti / Reviewers

Vanja Ambrožič
Blaž Bertalanič
Aljaž Blatnik
Sašo Blažič
Borko Boškovič
Gregor Burger
Urban Burnik
Gregor Černe
Franc Dimc
Matjaž Divjak
Jani Dugonik
Žiga Emeršič
Aljaž Francič
Helena G.-Tomc
Gregor Geršak
Timotej Gruden
Jože Guna
Aleš Hace
Aleš Holobar
Matevž Hribernik
Marija Ivanovska
Grega Jakus
David Jelenc
Roman Kamnik
Gorazd Karer
Gregor Klančar
Simon Kolmanič
Marko Kos
Andrej Košir
David Kraljić
Peter Kramar
Iztok Kramberger
Matej Kramberger
Tadej Krivec
Andrej Lavrič
Bogdan Lipuš
Alenka M.-Lebar
Blaž Meden
Marko Meža
Peter Miklavčič
Uroš Mlakar
Tomi Mlinar
Jon Muhovič
Gašper Mušič
David Nedeljković
Klemen Pečnik
Samo Penič
Matija Pirc
Peter Planinšič
Matevž Pogačnik
Franc Policardi
Boštjan Pregelj
Matevž Pustišek
Peter Rot
Urban Sedlar
Danijel Skočaj
Žiga Stržinar
Riko Šafarič
Martin Šavc
Jaka Šircelj
Andrej Štern
Vitomir Štruc
Simon Tomažič
Andrej Trost
Anton Umek
Matej Zajc
Janez Zaletelj
Aleš Zamuda

SEKC./SECT. VP
Vabljena predavanja / Invited Papers 1

Modeling the spread of Coronavirus with mobility networks: Explaining infection rates and informing reopening strategies
Jure Leskovec 2

SEKC./SECT. EL
Elektronika / Electronics 3

Merilnik faznega premika z nastavljivo ločljivostjo
Mitja Solar, Gregor Nikolič 4

Real-Time Image Processing – Performance to Power Ratio of Flash-Based FPGA vs. GPU
Uroš Hudomalj, Muhammad Subhan Hameed, Christopher Mandla, Markus Plattner 8

Prenos pristopov uporabljanih pri razvoju programske opreme v EDA okolja
Mitja Nemec 12

Analiza močnostnih modulov za učinkovito krmiljenje elektromotorjev
Aleksander Sešek, Tadej Skuber, Janez Trontelj 16

SEKC./SECT. TC
Telekomunikacije / Telecommunications 20

Razvoj UWB radarja
Marko Malajner, Erich Leitgeb, Peter Planinšič, Dušan Gleich 21

Preizkušanje radarja v milimetrskem frekvenčnem področju za uporabo na morju
Mladen Radovanović, Boštjan Batagelj 25

Spremljanje satelitov GPS v vidnem polju na osnovi modula NEO-6M
Blaž Bertalanič, Aljaž Blatnik, Boštjan Batagelj 29

Komunikacijsko satelitsko omrežje Starlink
Žiga Andrejc, Boštjan Batagelj 33

Generator THz optičnega signala z DFB lasersko diodo
Blaž Pongrac, Dušan Gleich 37

Izboljšanje ojačenja BoR-monopolnih anten z algoritmom Diferencialne Evolucije
Marko Radović, Peter Kitak, Gorazd Lešnjak, Peter Planinšič 41

Satelitska podpora vplutju v pristanišče in privezovanju
Franc Dimc, Blaž Luin, Marko Perkovič 45

Multimodal Electronic Access Control System Using Digital Identity for Enabling Authorized Physical Access to Premises
Matej Rabzelj, Tjaša Jereb, Aljaž Martinčič, Andrej Kos 49

Zagotavljanje varnosti v celičnih omrežjih
Jernej Mušič, Boštjan Batagelj 53

Senzorski sistem za ocenjevanje agilnosti
Matevž Hribernik, Erik Keš, Anton Umek, Anton Kos 58

5G - ali res predstavlja tveganje za zdravje?
Tomi Mlinar 62

Modeliranje razširjanja radijskih valov v naravnih jamah
Andrej Hrovat, Michel D. Bedford, Tomaž Javornik 66

Pametno vinogradništvo: načrtovanje pilota v Šmarjah pri Kopru z uporabo LoRaWAN tehnologije
Aleksander Hrastič, Vid Čermelj, Emilija Stojmenova Duh 70

Določanje dejanskega efektivnega lomnega količnika svetlovoda v silicijevem optičnem čipu
Andrej Lavrič, Janez Krč, Boštjan Batagelj 74

Spremljanje hitrosti in načina gibanja uporabnika multimedijske storitve z visokofrekvenčnimi radijskimi signali
Peter Nimac, Levac Urška, Nina Gartner , Jure Janez Markovič, Aleš Simončič, Rok Vidmar, Amadej Vidic, Jakob Gazič, Andrej Košir, Matija Svetina, Urban Burnik 78

Nizkocenovna avtomatizacija visokofrekvenčnih meritev
Peter Miklavčič, Boštjan Batagelj 82

Sistemi za elektronsko bojevanje
Gal Likar, Boštjan Batagelj 86

SEKC./SECT. MM
Multimedija / Multimedia 90

Vrednotenje uporabniške izkušnje inteligentnih tipk
Gašper Kastelic 91

Raznolikost voznikovih odzivov na zahtevo za prevzem vodenja pogojno avtomatiziranega vozila
Timotej Gruden, Grega Jakus 95

Interaktivna razstavna vitrina – orodje za muzejske zbirke prihodnosti
Simon Kolmanič, Niko Lukač, Selma Rizvić, Borut Žalik 99

Platforma za razpoznavo 3D gest s pomočjo električnega polja
Aljaž Blatnik 103

Opis delokroga za produkcijo 360-stopinjskega glasbenega videospota
Veronika Malačič, Marko Cafnik, Helena Gabrijelčič Tomc 107

Urejanje in dostava video posnetkov z oblačno arhitekturo mikrostoritev
Uroš Zoretič, Grega Jakus 111

Primerjava HEVC in VP9
David Valič, Marko Meža, Urban Burnik 115

Sensitivity analysis for face detection
Romi Koželj, Žiga Emeršič, Luka Šajn 119

Razvoj filtra za obogatenje resničnosti
Miha Malenšek, Blaž Pridgar, Narvika Bovcon 123

Influence of interaction designs of in-vehicle infotainment systems on drivers' preferences
Tomaž Čegovnik, Kristina Stojmenova, Sašo Tomažič, Jaka Sodnik 127

SEKC./SECT. AR
Avtomatika in robotika / Automatic Control and Robotics 131

Vodenje rotacijskega invertiranega nihala
Aljaž Blažič, Gorazd Karer 132

Avtomatizacija namiznega nogometa
Matevž Bošnak, Andrej Zdešar, Simon Tomažič, Goran Andonovski, Gregor Klančar, Sašo Blažič, Igor Škrjanc 136

Izdelava laboratorijskega demonstracijskega sistema (Mitsubishi)
Matej Milavec, Goran Andonovski 140

Trajectory Planning By Applying Optimal Velocity Profile Algorithm on Bernstein-Bézier Motion Primitives
Martina Loknar, Gregor Klančar, Sašo Blažič 144

Vpliv kinematičnih indeksov manipulabilnosti na sposobnost izvedbe trajektorije orodja z robotom UR5
Saša Stradovnik, Aleš Hace 148

Optimizing robot motion synthesis from digital images using neural networks
Matija Mavsar 152

Learning of the Velocity Profile for Quality Inspection Motion Using PoWER
Zvezdan Lončarević, Rok Pahič, Andrej Gams 156

Turingov test zaznave gibalne sposobnosti pri fizičnem sodelovanju med človekom in inteligentnim robotskim agentom
Rebeka Kropivšek Leskovar, Tadej Petrič 160

Ocenjevanje dvoročne vadbe z robotom
Ana Mandeljc, Eva Čebašek, Marko Munih, Janez Podobnik, Matjaž Mihelj 164

Merilni sistem za analizo dinamičnega trka sodelujočega robota v skladu z ISO/DIS 21260
Sebastjan Šlajpah, Mario Klenovšek, Marko Munih, Matjaž Mihelj 168

Vodenje sodelujoče mobilne robotske celice
Peter Kmecl, Matjaž Mihelj, Marko Munih, Janez Podobnik 172

Sekc./Sect. SM
Modeliranje in simulacija / Modeling and Simulation 176

Analytical Modelling of Achieving Energy Efficiency with Green Walls
Dejan Tasić, Amor Chowdhury, Dalibor Igrec 177

Vpliv viskoznosti nosilne tekočine na izgube magnetne tekočine v izmeničnem magnetnem polju
Jakob Vizjak, Miloš Beković, Anton Hamler 181

Model globokih gaussovskih procesov za identifikacijo procesa Bouc-Wen
Tadej Krivec, Juš Kocijan 185

Samorazvijajoča se segmentacija globinske slike stereo kamere za notranja strukturirana okolja
Miloš Antić, Andrej Zdešar, Igor Škrjanc 189

Poligon z miniaturnimi avtomatsko vodenimi vozički za učenje in razvoj avtonomnih mobilnih sistemov
Andrej Zdešar, Matevž Bošnak, Gregor Klančar 196

Solution space of optimal heat pump schedules
David Kraljić, Miha Troha 200

Sekc./Sect. ME
Močnostna elektrotehnika / Power Engineering 204

Electric field strength inside a residential house due to a high voltage overhead line proximity
Amar Zejnilović, Breda Cestnik, Boštjan Blažič 205

Dinamična u-t karakteristika varistorjev
Anton Zhezhov, Mislav Trbušić, Anton Hamler 211

Načrtovanje laboratorijskega udarnega napetostnega generatorja
Mislav Trbušić, Marko Jesenik, Mladen Trlep, Anton Hamler 215

Več-vejno delovanje paralelnih močnostnih filtrov
Mitja Nemec 219

Iskanje in odpravljanje elektromagnetnih motenj pri štirikvadrantnem pogonu
Mitja Nemec, Andraž Rihar 223

Dvojni DCT regulator za regulacijo harmonskih komponent
Denis Sušin 227

Optimizacija izgub električnega pogona z asinhronskim motorjem
Žiga Fabijan, Rastko Fišer, Henrik Lavrič, Klemen Drobnič 231

Valovitost toka v 3-faznem 2-nivojskem razsmerniku
David Kovačič, Rastko Fišer, Klemen Drobnič 235

Odpravljanje negativne sofazne napetosti pri unipolarno napajanem odštevalnem vezju z operacijskim ojačevalnikom
Marko Petkovšek, Peter Zajec 239

Sekc./Sect. MT
Merilna tehnika / Measurement 243

Sledljivost in točnost čelnih termometrov
Igor Pušnik 244

Automatic ski-jump distance measurement with convolutional neural networks and computer vision
Matjaž Kukar, David Nabergoj 248

Optimizacija ojačevalnika za praktični šumni termometer
Rok Tavčar, Samo Beguš, Jovan Bojkovski 252

Avtomatizacija sistema za umerjanje merilnikov napetostnega razmerja
Ana Mandeljc, Miha Hiti, Gaber Begeš 256

Orodje za oddaljeno testiranje spletnih uporabniških vmesnikov
Jure Trilar, Vid Čermelj, Emilija Stojmenova Duh 260

Električna merilna aparatura za penetracijo kaplje tekočine v porozni podlagi
Marion Thebault, Franc Policardi-Antoncich 264

Primerjava termovizijskih kamer za merjenje telesne temperature pri ljudeh
Vid Mlačnik, Igor Pušnik 268

LOEE 4.0 – oddaljeno merjenje vektorskega polja z uporabo 3D kartezičnega stroja
Kristjan Vuk Baliž, Janez Zaletelj, Marjan Dolinšek, Iztok Humar 272

SEKC./SECT. AE
Akustika in elektroakustika / Acoustics and Electroacoustics 276

Zahteve za odmevnice in pristop k projektiranju odmevnice akustičnega laboratorija InnoRenew CoE
Rok Prislan 277

Acoustical project of the new Banca Popolare di Sondrio in Berbenno Service Center - I
Marcello Brugola, Franc Policardi-Antoncich 280

Uporaba prostorskega zvoka v avtomatiziranih vozilih
Harsh Sanghavi, Kristina Stojmenova, Chihab Nadri, Sangjin Ko, Myounghoon Jeon, Jaka Sodnik . 284

Detekcija infrazvoka z zvočnikom v funkciji mikrofona
Samo Beguš, Gregor Geršak 288

Vodno žigosanje audio vsebin z uporabo diskretne valčne transformacije (DWT)
Nejc Bevk, Urban Burnik, Marko Meža 292

Suspended False Ceiling acoustic elements theoretical and experimental investigation
Giovanni Amadasi, Franc Policardi-Antoncich 296

SEKC./SECT. CS
Računalništvo in informatika / Computer and Information Science 300

Ekstrakcija ključnih besed filmov iz podnapisov
Jan Popič, Timi Ornik, Nejc Planer, Borko Bošković, Janez Brest 301

Diferencialna evolucija za učenje agenta umetne inteligence pri igranju splošnih videoiger
Aleš Zamuda, Matjaž Vörös 305

Uporaba metod strojnega učenja za analizo vpliva lipoproteinov (a) na bolezni srca in ožilja
Tajda Bogovič, Peter Kokol, Tadej Završnik, Jernej Završnik, Helena Blažun Vošner, David Šuran 309

Pregled nekaterih metod na področju analize časovnih vrst
Žiga Stržinar, Boštjan Pregelj, Igor Škrjanc 313

Načrtovanje in izdelava prototipa solarnega sledilnika
Domen Gošek, Amor Chowdhury, Karl Gotlih, Riko Šafarič 317

Sistem za zajemanje obdelavo in obveščanje uporabnikov o kritičnih situacijah v njihovi okolici
Dalibor Igrec, Uroš Sadek, Amor Chowdhury 322

Spletna aplikacija za dostop do medicinskih slik v standardu DICOM
Žan Regoršek, Samo Penič, Miha Fošnarič, Janez Žibert 326

Sistem prikazovanja eksponatov muzeja
Jan Breznar, Bogdan Lipuš, Simon Kolmanič 331

Mehanizmi izmenjave zaupanja in ugleda
David Jelenc 335

SEKC./SECT. PR
Razpoznavanje vzorcev / Pattern Recognition 339

Štetje objektov na slikah z uporabo genetskega algoritma
Gregor Babnik, Luka Šajn 340

Semi-automated correction of MOBIUS eye region annotations
Ožbej Golob, Peter Peer, Matej Vitek 344

Measuring the induction of affect using facial expression analysis technology: a pilot study
Marko Meža, Andrej Košir, Gregor Strle 348

Semantic Face Editing with GAN Latent Code Optimization
Martin Pernuš, Vitomir Štruc, Simon Dobrišek 351

O klasifikaciji slik v ne-enolično določljive razrede
Jon Natanael Muhovič, Domen Tabernik, Danijel Skočaj 355

Co-segmentation for visual object tracking
Luka Čehovin Zajc 359

Using Google Colaboratory for Image Classification
Martina ć, Damir Demirović, Amira Šerifović-Trbalić 363

A review and comparison of time series similarity measures
Maša Kljun, Matija Teršek, Erik Štrumbelj 367

SEKC./SECT. BM
Biomedicinska tehnika / Biomedical Engineering 371

Pasti zajema in obdelave znotrajsrčnih elektrogramov s primeri iz klinične prakse
Jernej Štublar, Matevž Jan, David Žižek, Damijan Miklavčič, Tomaž Jarm 372

Izdelava fantoma za ultrazvočno medicinsko diagnostiko s postopkom 3D tiskanja
Nejc Klanjšček, Sašo Arnuga, Miha Fošnarič, Samo Penič 376

Sklopljena numerična simulacija temperaturnega profila linearne pretočne komore med dovajanjem električnih pulzov
Peter Lombergar, Samo Mahnič-Kalamiza, Karel Flisar, Damijan Miklavčič 380

Electrical Impedance Spectroscopy insights into plant tissues treated by Pulsed Electric Fields
Jessica Genovese, Samo Mahnič-Kalamiza, Matej Kranjc, Pietro Rocculi, Damijan Miklavčič ... 384

Numerično modeliranje porazdelitve pH sprememb v tkivu med dovajanjem elektroporacijskih pulzov
Rok Šmerc, Samo Mahnič-Kalamiza, Damijan Miklavčič 388

SEKC./SECT. DI
Didaktika / Didactics 392

ICT engineering study programs to meet modern society needs: Erasmus+ project BENEFIT
Matej Zajc, Aljo Mujčić, Andrea Tonello, Vlado Delić, Mladen Koprivica, Snježana Rimac Drlje, Urban Burnik, Nermin Suljanović 393

Quality management of Erasmus+ BENEFIT project
Urban Burnik, Francesco Marcuzzi, Dušanka Bošković, Drago Žagar, Nataša P. Maleš-Ilić, Darko Huljenić, Matej Zajc 397

BENEFIT Industry information portal - a service for better cooperation between universities and companies in the teaching process
Snježana Rimac-Drlje, Drago Žagar 401

Teaching Methodologies Adopted at the BENEFIT Project for ICT Engineering Education Before and During Covid-19
Vlado Delić, Aljo Mujčić, Zdenka Babić, Dušanka Bošković, Nataša Maleš-Ilić, Mladen Koprivica, Drago Žagar, Matej Zajc, Nermin Suljanović, Andrea Tonello 405

Role of M2M communications laboratory formed within BENEFIT project in education processes at FEE-UNI
Dejan Milić, Nataša Maleš-Ilić, Dejan Ćirić, Aleksandar Atanasković 409

Experimental Performance Evaluation of IEEE 802.15.4g Applications Using OpenMote-B Devices
Milica Lekić, Gordana Gardašević 413

Stimulating Entrepreneurship in Teaching Human Computer Interaction
Dušanka Bošković, Nihad Borovina, Merima Zukic 417

An Overview of Open Source Environments for 5G Internet of Things Experimentation
Srdjan Sobot, Nikola Gavric, Zivko Bojovic, Dejan Vukobratovic 421

Multi-subject Project Based Learning: Two Semester Pilot Study
Goran B. Marković, Jelena D. Ćertić, Marija Božić, Mladen Koprivica 425

Sodobne učne strategije in tehnologije: lažna obrnjena učilnica s poizvedovalnim učenjem in 3D tiskom
Špela Kosec, Tanja Rozman, Janez Jamšek 429

Virtualni laboratorij za poučevanje digitalnih sistemov
Janez Perš, Andrej Trost 433
Učenje na daljavo: spletno učenje v osnovnošolskem tehniškem izobraževanju - pedagoška praksa
Bernarda Urankar, Janez Jamšek 438
SEKC./SECT. ST
Študentski članki / Student Competition IEEE Slovenia 443
Time series classification using time-frequency analysis and Convolutional Neural Networks
Domen Kavran 444
Avtomatska segmentacija Golgijevih aparatov v volumetričnih podatkih z grobo označenim zlatim standardom
Eva Boneš 448
Predvidevanje vrednosti delnice in problem pretiranega prilagajanja modela
Miha Breznik 452
Estimation of the derating PV power due to inverter clipping using Artificial Neural Network
Jakob Bevc 456
Vector Space Representation of Movies and Music Albums Using Node Embedding Methods
Vid Keršič 460
Razvoj IoT aplikacije za pametni dom
Jaša Vid Meh Peer 464

Vabljena predavanja

Invited Lectures

Modeling the spread of Coronavirus with mobility networks: Explaining infection rates and informing reopening strategies

Jure Leskovec

Stanford University, USA

Abstract. Developing computational models of how viruses spread among a population of people is extremely challenging. In this talk I will demonstrate how fine-grained epidemiological modeling of the spread of Coronavirus -- capturing who is infected at which locations -- can aid the development of policy responses that account for heterogeneous risks of different locations as well as the disparities in infections among different demographic groups. We use US cell phone data to capture the hourly movements of millions of people and model the spread of Coronavirus from March 1 - May 2, 2020 among a population of 98 million people in 10 of the largest US metropolitan areas. We show that even a relatively simple epidemiological model can accurately capture the case trajectory despite dramatic changes in population behavior due to the virus. We also estimate the impacts of fine-grained reopening plans: we predict that a small minority of "superspreader" location account for a large majority of infections, and that reopening some locations (like restaurants) pose especially large risks. We also explain why infection rates among disadvantaged racial and socioeconomic groups are higher. Overall, our model supports fine-grained analyses that can inform more effective and equitable policy responses to the Coronavirus.

Reference: Mobility network modeling explains higher SARS-CoV-2 infection rates among disadvantaged groups and informs reopening strategies.

Jure Leskovec is an associate professor of Computer Science at Stanford University, the Chief Scientist at Pinterest, and an Investigator at the Chan Zuckerberg Biohub. He was the co-founder of a machine learning startup Kosei, which was later acquired by Pinterest. Leskovec's research area is machine learning and data science for complex, richly-labeled relational structures, graphs, and networks for systems at all scales, from interactions of proteins in a cell to interactions between humans in a society. Applications include commonsense reasoning, recommender systems, social network analysis, computational social science, and computational biology with an emphasis on drug discovery. This research has won several awards including a Lagrange Prize, Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship, and numerous best paper and test of time awards. It has also been featured in popular press outlets such as the New York Times and the Wall Street Journal. Leskovec received his bachelor's degree in computer science from University of Ljubljana, Slovenia, PhD in machine learning from Carnegie Mellon University and postdoctoral training at Cornell University. You can follow him on Twitter at @jure.

Elektronika

Electronics

Merilnik faznega premika z nastavljivo ločljivostjo

Mitja Solar, Gregor Nikolič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, 2000 Maribor

Adjustable resolution phase angle meter

Abstract. A phase shift meter between two sinusoidal signals of the same frequency is presented, which has an adjustable high resolution, does not require additional clock generators for measurement, is simple, affordable and usable in the frequency range from 1 kHz to 1 MHz. The circuit is integrable in a combined analog-digital integrated circuit.

1 Uvod

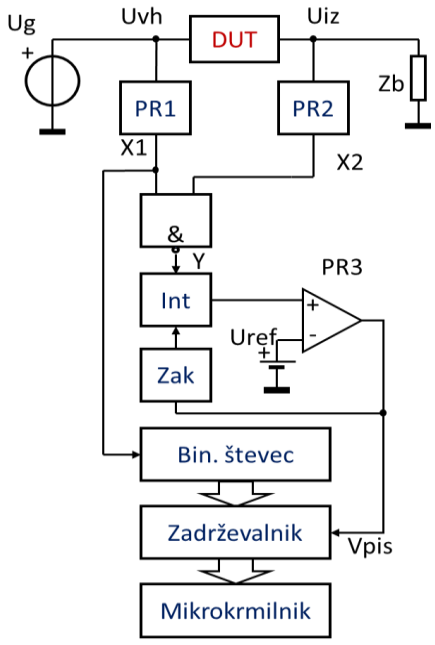
Za meritve faznega premika se, razen specialnih merilnikov faznega premika, uporabljajo osciloskop, kjer pa je merilna negotovost v rangu merilne negotovosti izmerjenega časovnega intervala med prehodom vzbujevalnega in merjenega signala, ali pa v rangu merilne negotovosti izmerjenih amplitud v X-Y načinu delovanja osciloskopa. Kadar gradimo vezje za merjenje faznega premika lahko uporabimo – obročne diodne demodulatorje, - analogne množilnike, - detektorje prehoda skozi nič, z ekskluzivnimi-ali vrati, ki pa ne dosežejo željene ločljivosti. Prav tako lahko uporabimo vezje PLL [3]. Druga možnost je uporaba posebnih integriranih vezji, ki merijo tako fazni premik kot amplitudo signalov [4,5]. Vse prej naštete naprave oziroma elementi imajo omejitve bodisi v razponu uporabe, kompleksnostjo ali pa je uporaba le te popolnoma namenska. V članku predstavljamo zasnovo merilnika faznega premika z nastavljivo ločljivostjo v kombinaciji analogno-digitalnega vezja, ki je integrabilno v integriranem vezju.

2 Materiali in metode

Osnovna zasnova merilnika je bila izvedba, na način meritve faznega premika med dvema sinusnima signaloma iste frekvence, z možnostjo nastavitve visoke ločljivosti, pri čemer ne potrebuje za izvajanje meritev dodatnih generatorjev takta. Zasnova je preprosta, cenovno sprejemljiva in uporabna v frekvenčnem območju od 1 kHz do 1 MHz. Delovanje merilnika temelji na cikličnem merjenju prehodov vzbujevalnega in merjenega signala sinusne oblike skozi vrednost nič, integriranju faznih premikov med njima in s štetjem period potrebnih da napetost iz integratorja doseže izbran nivo. Izvedena je bila tudi analiza pogreškov merilnika faznega premika.

Osnovne zahteve za delovanje vezja: detektiranje faznih premikov med od -180° do +180°, amplituda signalov je med 0,05 in 1 V, frekvenčno območje za vhodne signale f_{min} = 1 kHz f_{max} = 1 MHz, odziv vezja je odvisen od zahtevane ločljivosti.

Po raziskavi različnih možnih izvedb faznih detektorjev smo zasnovali fazni detektor predstavljen na sliki 1.



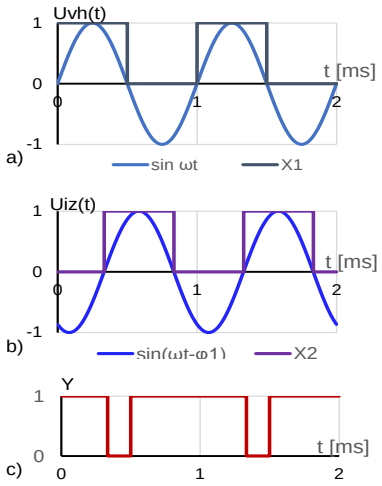
Slika 1 Zasnovan detektor faznega premika

Pri čemer je DUT merjenec, primerjalnika PR1 in PR2 preoblikujeta signal sinusne oblike v pravokotna signala X1 in X2. Z NEIN vrati izberemo interval integriranja integratorja Int. Ko napetost integratorja

po več periodah vhodnega signala doseže referenčni signal na vhodu primerjalnika PR3, primerjalnik preklopi v visoko stanje in sproži vpis iz binarnega števca v zadrževalnik, z majhno zakasnitvijo Zak pa z elektronskim stikalom izprazni kondenzator integratorja.

Signal iz generatorja sinusne oblike Ug pripeljemo na vhod merjenca in na primerjalnik. Na izhodu primerjalnika PR1 dobimo napetost pravokotne oblike X1. Na izhodu merjenca imamo drugi primerjalnik PR2 in na izhodu dobimo napetost X2. Med obema signaloma je fazni premik -φ oziroma časovni premik signala za -t_φ Na sliki 2 so signali predstavljeni za frekvenco generatorja f_g = *1 kHz*.

Periodo signala označimo s $T = \frac{1}{f_g} = 1\ ms$, zakasnitev izhodnega signala s $t_\varphi = 0{,}35\ ms$. Integrabilni interval je določen s $\left(\frac{T}{2} + t_\varphi\right) = 0{,}85\ ms$ oziroma fazni premik $(180° + \varphi) = 180° + 126° = 306°$. Interval je določen zaradi znanega omejenega trajanja časa meritve, ki znaša med $1/2 \cdot N_0 \cdot T_0$, kjer je fazni premik v območju $\varphi = \pm 180°$, in $N_0 \cdot T_0$, kjer je N_0 število period vhodnega signala pri faznem premiku $\varphi = 0°$.



Slika 2 Časovni poteki: a) vhodnega signala Uvh (t), b) izhodnega signala Uiz (t) in c) signala Y iz logičnega vezja

Integrator je zasnovan s tokovnim generatorjem, zaporedno vezanimi prehodnimi vrati in kondenzatorjem ter z invertirajočim Millerjevim integratorjem ali pa je izveden z neinvertirajočim Deboo integratorjem [6,7] ali s transkonduktančnim ojačevalnikom, ki polni kondenzator.

V primeru tokovno napajanega kondenzatorja je v eni periodi vhodnega signala T na kondenzatorju napetost

$$u_{C0}(t) = \frac{1}{C}\int_0^T i(t)\,dt = \frac{1}{C}\int_0^{t_\varphi + T/2} I_0\,dt = \frac{I_0}{C}\left(\frac{T}{2} + t_\varphi\right) \quad (1)$$

V N periodah bo napetost na kondenzatorju $u_C(t) = N \cdot u_{C0}(t)$ dosegla referenčno vrednost napetosti na primerjalniku U_{ref}. Ker nas zanima odvisnost od faznega premika $\varphi = \omega t_\varphi$, vpeljemo v enačbo $T = \frac{1}{f_0}$ in $t_\varphi = \frac{1}{f_0}\frac{\varphi[°]}{360°}$ ter dobimo

$$u_C(t) = N\,u_{C0}(t) = N\frac{I_0}{2C}\frac{1}{f_0}\left(1 + \frac{\varphi[°]}{180°}\right) \quad (2)$$

Iz enačbe 2 lahko izrazimo število period za merjen fazni premik

$$N = \frac{U_{ref}\,2C\,f_0}{I_0}\,\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} = N_0\,\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} \quad (3)$$

in fazni premik izražen s številom period

$$\varphi[°] = \left(\frac{U_{ref}\,2C\,f_0}{I_0\,N} - 1\right)180° = \left(\frac{N_0}{N} - 1\right)180° \quad (4)$$

Če želimo doseči ločljivost $\Delta\varphi = 0{,}1°$, nas zanima koliko period signala generatorja je potrebno za dosego te ločljivosti. Najprej določimo koeficient občutljivosti faznega premika

$$\frac{dN}{d\varphi} = \frac{U_{ref}\,2C\,f_0}{I_0}\,\frac{\frac{1}{180°}}{\left(1+\frac{\varphi[°]}{180°}\right)^2} = -N_0\,\frac{\frac{1}{180°}}{\left(1+\frac{\varphi[°]}{180°}\right)^2} \quad (5)$$

Spremembo števila period za željeno ločljivost

$$\Delta N = \frac{dN}{d\varphi}\Delta\varphi = N_0\,\frac{\frac{0{,}1°}{180°}}{\left(1+\frac{180°}{180°}\right)^2} = N_0 \cdot 0{,}000278 \quad (6)$$

V primeru največjega faznega premika $\varphi = 180°$ pri ločljivosti $\Delta\varphi = 0{,}1°$ dobimo minimalno spremembo števila period $\Delta N = 1$, zato je $N_0 = 3600$.

3 Analiza pogreškov

Frekvenca generatorja določa število merjenih period N_0 in N. Spremembe frekvence med meritvijo ne pričakujemo, zato ni vključena v izračunu skupnega merilnega pogreška.

Koeficienti občutljivosti, ki vplivajo na število period N in N_0 (enačba 2 in 3) so:
- kapacitivnost kondenzatorja C s koeficientom

občutljivosti kapacitivnosti

$$\frac{dN}{dC} = \frac{U_{ref}\,2\,f_0}{I_0}\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} = \frac{N_0}{C}\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} \quad (7)$$

- polnilni tok I_0 s koeficientom občutljivosti toka

$$\frac{dN}{dI_0} = -\frac{U_{ref}\,2C\,f_0}{{I_0}^2}\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} = -\frac{N_0}{I_0}\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} \quad (8)$$

- referenčna vrednost preklopa primerjalnika U_{ref} s koeficientom občutljivosti referenčne napetosti

$$\frac{dN}{dU_{ref}} = \frac{2C\,f_0}{I_0}\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} = \frac{N_0}{U_{ref}}\frac{1}{\left(1+\frac{\varphi[°]}{180°}\right)} \quad (9)$$

Točnost vhodnih primerjalnikov, ki sinusni signal spremenijo v pravokotni signal s koeficientom občutljivosti premika faznega kota je ključnega pomena in jih določimo z enačbo 11.

$$\frac{dN}{d\varphi_c} = \frac{U_{ref}\,2C\,f_0}{I_0}\frac{\frac{1}{180°}}{\left(1+\frac{\varphi[°]}{180°}\right)^2} = -N_0\frac{\frac{1}{180°}}{\left(1+\frac{\varphi[°]}{180°}\right)^2} \quad (10)$$

Z upoštevanjem, da so posamezni izvori pogreškov med seboj ne-korelirani, dobimo najslabši primer merilnega pogreška števila period ΔN.

$$\Delta N = \frac{dN}{dC}\Delta C + \frac{dN}{dI_0}\Delta I_0 + \frac{dN}{dU_{ref}}\Delta U_{ref} + \frac{dN}{d\varphi_c}\Delta\varphi_c \quad (11)$$

Drugi vir pogreškov pri izračunu faznega premika je operacija deljenja N_0/N (enačba 4), katera se izračuna v aritmetiki mikrokrmilnika s plavajočo vejico, zato je ta vir pogreška zanemarljiv.

Skupni merilni pogrešek je tako določen z:

$$\Delta N = \pm\frac{N_0}{\left(1+\frac{\varphi[°]}{180°}\right)}\left(\frac{\Delta C}{C}+\frac{\Delta I_0}{I_0}+\frac{\Delta U_{ref}}{U_{ref}}+\frac{\Delta\varphi_c}{180°}\right) \quad (12)$$

Pogrešek faznega premika pa z:

$$\Delta\varphi[°] = \pm 180°\left(\left|\frac{d\varphi[°]}{dN_0}\right|\Delta N_0 + \left|\frac{d\varphi[°]}{dN}\right|\Delta N\right) =$$

$$180°\left(\frac{1}{N}\Delta N_0 + \left|-\frac{N_0}{N^2}\right|\Delta N\right) \quad (13)$$

4 Rezultati in diskusija

Najprej smo v EXCELu pripravili tabelo 1 z izbranimi vrednostmi in njihovimi pogreški.

Tabela 1. Vrednosti komponent in njihovi pogreški

C [µF]	0,1	ΔC [nF]	0,1
I_0 [µA]	134	ΔI_0 [µA]	0,34
Uref [V]	1,5	ΔU_{ref} [mV]	0,6
N_0	2227	ΔN	3

V tabeli 2 so izračunane vrednosti pogreškov za frekvenco 1 MHz, za fazne premike od 0 do 180°.

V tabeli 2 je z $\Delta\varphi$ označen merilni pogrešek izračunan z uporabo enačbe 14, z $\Delta\varphi_{sqr}$ pa srednji kvadratni pogrešek izračunan s korenom iz vsote kvadratov posamičnih pogreškov.

Tabela 2. Izračunan merilni pogrešek faznega premika

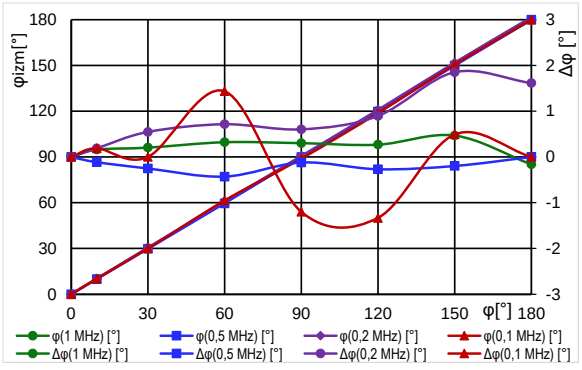
n	φ[°]	N	ΔN_0	ΔN	ΔUref	ΔC	ΔI_0	$\Delta\varphi_C$	Δφ	$\Delta\varphi_{sqr}$
1	0	2227	0,24	0,24	0,14	0,04	0,90	0,08	1,64	0,98
2	10	2110	0,26	0,27	0,15	0,04	0,95	0,09	1,75	1,03
3	30	1909	0,28	0,33	0,17	0,04	1,05	0,11	1,98	1,15
4	60	1670	0,32	0,43	0,19	0,05	1,20	0,14	2,33	1,34
5	90	1485	0,36	0,55	0,22	0,05	1,35	0,18	2,71	1,53
6	120	1336	0,40	0,67	0,24	0,06	1,50	0,22	3,10	1,72
7	150	1215	0,44	0,81	0,26	0,07	1,65	0,27	3,51	1,93
8	180	1114	0,48	0,97	0,29	0,07	1,80	0,32	3,93	2,14

Vidimo, da imajo največji vpliv, na merilni pogrešek faznega premika, točnost števca (ΔN_0 in ΔN), ki je povezana z linearnostjo integratorja, konstantnost polnilnega toka (ΔI_0) in preciznost vhodnih primerjalnikov ($\Delta\varphi_c$), sprememba kapacitivnost (ΔC) in točnost referenčne napetosti primerjalnikov (ΔU_{ref}) pa ob izbiri kvalitetnih elementov vplivajo manj.

Vezje z Deboo-jevim integratorjem smo načrtali in simulirali s SPICE analognim simulatorjem Tina-TI [9] in je predstavljeno na sliki 5.

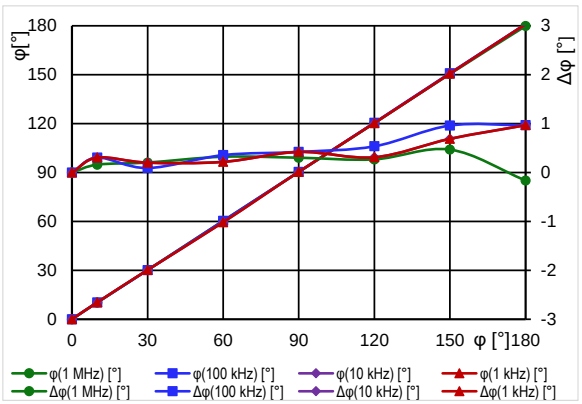
Na sliki 3 so predstavljeni rezultati simulacij za frekvence 1, 10 in 100 kHz in 1MHz za vezje iz slike 5. Pri tem je število period $N_0 = 2227$ pri 1MHz, 1114 pri 500kHz, 448 pri 200kHz in 224 pri 100kHz.

Vidimo, da je merilni pogrešek faznega premika pri frekvencah 1MHz, 500kHz in 200kHz $E = \pm 0{,}5°$, pri frekvenci 100kHz pa se zaradi majhnega števila preštetih period poveča na $E = \pm 2°$. Zato smo predvideli, da se glede na vhodno frekvenco spremenijo vrednosti kondenzatorja v integratorju. Za ta primer smo pri frekvencah od 1MHz do 1kHz povečevali kapacitivnost za 10-krat na dekado.



Slika 3. Prenosna karakteristika merilnika faznega premika in merilni pogreški faznih premikov pri frekvencah od 100kHz do 1MHz

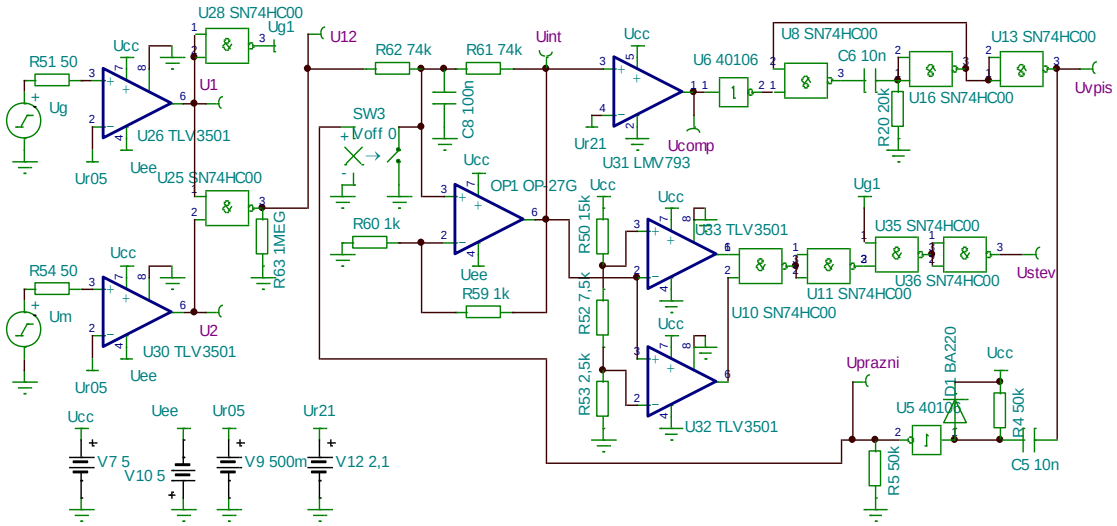
Na sliki 4 so rezultati simulacij za frekvence 1, 10, 100kHz in 1MHz.



Slika 4. Prenosna karakteristika merilnika faznega premika in merilni pogreški faznih premikov pri frekvencah od 1kHz do 1MHz

Iz simuliranih rezultatov vidimo, da se pogreški povečajo pri večjih faznih premikih (nad 120°), a pri največjem številu period $N_0 = 2227$ ne presežejo 1°. Pri visokih frekvencah (nad 100kHz) so pogreški prav tako znotraj 1°, bi pa bilo potrebno povečat kapacitivnost kondenzatorja C8 iz slike 5 za 10-krat na 1μF.

Vezje smo tudi preizkusili na eksperimentalni ploščici.



Slika 5. Vezje merilnika faznega premika

5 Zaključek

Skonstruirali smo merilnik faznega premika med merjenima signaloma, pri katerem se da, z izbiro integracijske konstante, izbrati želeno ločljivost merilnika tudi pri frekvencah nad 100kHz. Rešitev je relativno enostavna, omogoča pa izvedbo: - z diskretnimi integriranimi komponentami, - z analognim vhodnim vezjem in mikrokrmilnikom in - v kombiniranem integriranem vezju.

Literatura

[1] F. Casper and P. Kowina, RF Measurement Concepts, Published by CERN in the Proceedings of the CAS-CERN Accelerator School: Advanced Accelerator Physics, Trondheim, Norway, 19–29 August 2013, edited by W. Herr, CERN-2014-009 (CERN, Geneva, 2014).

[2] Min Zhang , Hai Wang, Hongbo Qin, Wei Zhao and Yan Liu, Phase Difference Measurement Method Based on Progressive Phase Shift, Electronics 2018, 7, 86.

[3] Peter O'Shea, Phase Measurement, in John. G. Webster, The measurement, instrumentation, and sensors, handbook, 1999, CRC Press LLC, Springer, IEEE Press.

[4] Analog Devices, LF-2.7GHz RF/IF gain and phase detector, AD8302, Analog Devices, Inc., 2002

[5] John Cowles, Barrie Gilbert, Accurate Gain/Phase Measurement at Radio Frequencies up to 2.5 GHz, Analog Devices, Analog Dialogue 35-05, 2001

[6] Maxim Integrated, Consider the *"Deboo"* single supply integrator, Application note 1155, 29 Aug, 2002

[7] Texas Instruments , AN-1515 A Comprehensive Study of the Howland Current Pump, SNOA474A-January 2008-Revised April 2013

[8] Joan Peuteman, Jean-Jacques Vandenbussche, Peter Lee, Development af a Low-Cost Accurate Phase Measurenment System, International Conference of Technologies and Innovation, ICTIA 2014, Sousse, Tunesia.

[9] Texas Instruments, Getting Started with TINA-TI, SBOU052A–August 2007–Revised August 2008.

Real-Time Image Processing – Performance to Power Ratio of Flash-Based FPGA vs. GPU

Uroš Hudomalj[1], Muhammad Subhan Hameed[1], Christopher Mandla[1], Markus Plattner[1]

[1]*Max Planck Institute for Extraterrestrial Physics, Gießenbachstraße 1, 85748 Garching bei München, Germany*
E-mail: hudomalj@mpe.mpg.de

Abstract. *This paper describes a low-power solution for upgrading an existing imaging system to achieve real-time processing of images. The presented solution is to add an FPGA based system between the camera and the existing receiver. The FPGA is configured to capture images incoming from the camera, process them in real-time, and transmit the processed images further to the end receiver. The solution is demonstrated on an existing imaging system that uses a camera with the Camera Link interface (I/F). The system is upgraded using Microsemi's Smartfusion2 Advanced Development Kit which includes a flash-based FPGA. The processing capabilities of the developed system are evaluated in terms of achievable processing data rate and power consumption for two widely used algorithms, namely for image filtering and image averaging. The system's performance is compared to a high performance GPU, the RTX 2080 Ti by Nvidia. The benefits and limitations of FPGA and GPU based real-time image processing systems are discussed.*

1 Introduction

Nowadays most of the imaging applications require real-time processing of the captured images for them to be relevant, useful and to prevent loss of image frames [1]. To process images incoming from a camera in real-time, each frame has to be processed before the next one arrives [2]. If this processing criterion is not met, frames could be dropped, thus causing the loss of information.

The real-time processing constraints can only be satisfied with sufficient processing data rates of the imaging system. These processing requirements are always increasing because we strive to improve our applications by developing more sophisticated algorithms as well as by capturing larger images at increasing frame rates. Due to increasing processing requirements, it is getting more and more challenging to process the captured images in real-time. Therefore, the most important requirement for real-time imaging systems is the achievable processing data rate. Another important characteristic of an imaging system is also its power consumption. This is crucial, especially in embedded applications [3]. The embedded applications are mostly battery powered. Hence, the imaging systems should use as little power as possible. However, the real-time processing constraints still have to be met.

In traditional imaging applications, multiple general purpose processors or a GPU are used for image processing. To improve an existing imaging application, it is often necessary to change the hardware of the system because the existing one is insufficient to satisfy the new real-time constraints required by the new application. Modifying the system is costly, especially if applications are upgraded frequently. An upgrade to a new, more powerful imaging system might even be infeasible to implement for compact embedded applications as they are restricted by weight, size and power consumption. A more effective and less expensive solution is to add to the system an additional, modifiable component with dedicated functionality of real-time image processing to comply with the changing processing needs. Field Programmable Gate Arrays (FPGAs) are the most appropriate choice for such purposes because they can achieve high processing data rates due to parallel processing. They can also be easily reconfigured when the application changes [2].

This paper presents a low-power solution for upgrading an existing imaging system to achieve real-time processing. This is done by inserting an FPGA based system between the camera and the end receiver. The FPGA acquires the image sent by the camera via an interface (I/F), processes it and then transmits the processed image further to the end receiver via the same or a different I/F. The real-time image processing capabilities of the developed system are compared with processing capabilities of a high performance GPU, the RTX 2080 Ti by Nvidia [4], for two widely used preprocessing algorithms, namely image filtering and image averaging. The implementations of the algorithms in the FPGA and the GPU are evaluated in terms of achievable processing data rates and their power consumptions. The differences between the two platforms for real-time image processing are discussed by highlighting their benefits as well as limitations.

2 System Overview

The developed low-power FPGA based system for real-time processing included into an existing imaging system is depicted in Fig. 1. The existing imaging system uses an industrial camera GO-2400M PMCL by Jai and a PC with a PCIe frame grabber as the image receiver. The developed FPGA system is based on Microsemi's Smartfusion2 Advanced Development Kit [5]. The FPGA is flash-based, meaning it has a lower power

consumption than more frequently used SRAM based FPGAs. Both the camera and the frame grabber support the Camera Link I/F. Therefore, the development board is extended with two FPGA Mezzanine Cards (FMCs) to connect the camera via Camera Link I/F to the FPGA and the FPGA via the Camera Link I/F to the frame grabber of the PC. For receiving the images captured by the camera and for transmitting the processed images further to the PC, Camera Link receiver and transmitter are implemented in the FPGA according to [6]. The used camera captures grayscale images of a maximum size of 1936 by 1216 pixels with bit depths of 8 bits. Each clock cycle the camera transmits 8 pixels in parallel, row by row. The used Camera Link implementation in the FPGA supports maximum interface frequency of 37.125 MHz, resulting in a frame rate up to 118 fps, which corresponds to a data rate of 2.23 Gbps.

3 Algorithm Implementation

The following subsections describe implementations of image filtering and image averaging in the developed FPGA system and in a GPU for performance comparison. Implementations of algorithms in both devices take advantage of their parallel processing capabilities.

The image filtering is in both devices implemented as a two dimensional convolution in spatial domain between the filter mask and the image. The image averaging implements the running average filter.

3.1 FPGA Implementations

In order to process the images in real-time, the FPGA implementations of both algorithms process also 8 pixels in parallel each clock cycle. The two algorithms differ in their memory requirements. To filter an image, only pixels from a small neighbourhood are needed. Therefore, only a certain number of rows corresponding to the filter size are needed to be stored in memory. On the other hand, to average multiple images, they all have to be stored in memory. The internal SRAM memory of the FPGA is too small to store such large amounts of data. Therefore, external DDR memory needs to be used. Memory access time limits the maximum achievable processing data rate. To maximize the throughput of the DDR memory, a circuit for accessing it with multiple overlapping bursts in a sequence is designed according to [6], achieving a throughput of 5.32 Gbps.

3.1.1 Image Filtering

To filter 8 pixels in parallel, 8 convolutions need to be calculated in parallel as well. To reduce the number of resources needed, the calculations are done at an increased frequency corresponding to the filter size. Implementation is done as described in [7].

The FPGA does not include hardware multipliers and adders for floating-point operations. Therefore, the filter coefficient are stored with 16 bit fixed floating-point precision. The intermediate calculations are done with fixed floating-point precision with a minimum number of bits, which do not influence the 8 bit output precision.

3.1.2 Image Averaging

As mentioned, 8 pixels are processed in parallel each clock cycle, meaning for image averaging that the running average filter (1) is applied to 8 pixels in parallel.

$$y(n) = y(n-1) + \frac{1}{L}[x(n) - x(n-L)] \qquad (1)$$

If image averaging is implemented according to (1), four memory accesses are needed each clock cycle for each pixel, i.e. to store the current averaged output pixel *y(n)*, to read the previous output pixel *y(n-1)*, to store the incoming pixel *x(n)* and to read its L-th previous value *x(n-L)*. At a pixel data rate of 2.23 Gbps, a minimal throughput of 8.92 Gbps of the DDR memory is needed, which is more than available. A solution is to simplify the filter by assuming $x(n-L) \approx y(n-1)$, resulting in:

$$y(n) = (1 - \frac{1}{L})y(n-1) + \frac{1}{L}x(n) \qquad (2)$$

which requires only two memory accesses. Therefore, a throughput of 4.46 Gbps is needed, which can be achieved in the used FPGA. The simplification is valid if pixels do not change much through time.

For efficient hardware implementation, it is further assumed that the number of images to average L is a power of 2, replacing the division with a simple bit shift.

3.2 GPU Implementations

The GPU implementations are written in CUDA and run on Nvidia's RTX 2080 Ti. The GPU has a total of 68 streaming multiprocessors (SM), a memory capacity of 11 GB and a base clock speed of 1350 MHz [4].

Each implementation of an algorithm on a GPU consists of three main steps: memory transfer of kernel inputs from the CPU to the GPU, execution of one or more kernels on the GPU, and memory transfer of kernel outputs from the GPU to the CPU. A kernel is executed on multiple GPU threads in parallel. The kernels are implemented to use image tiling so that tiles of the predetermined length of an input image are stored in the kernel shared memory in order to reduce memory access times. The kernel is launched with a predetermined GPU thread block and grid size. Upon launch, the thread blocks are queued to run on the GPU's SM.

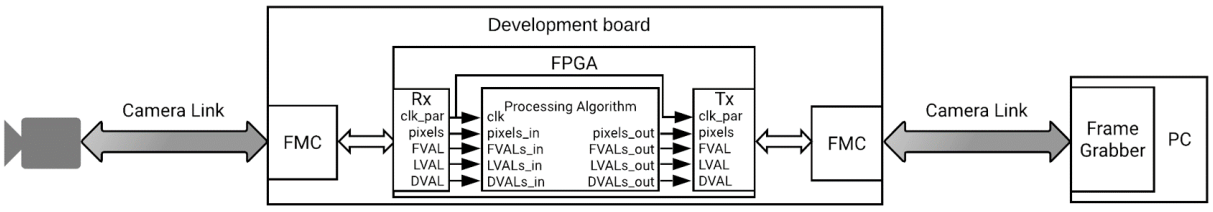


Fig. 1. Overview of developed system included into existing imaging system.

3.2.1 Image Filtering

Image filtering is implemented in two steps. The input image is first zero padded on each of its sides by a length equal to half the filter dimension. The zero padded image is then convolved with a filter kernel, producing the filtered image. The filter kernel is initialized beforehand. Therefore, the initialization does not affect the performance. Both the zero padding and convolution steps are implemented in separate CUDA kernels. As already mentioned, the code within each kernel is executed on multiple GPU threads in parallel. The separate kernels are executed on the default CUDA stream non concurrently.

3.2.2 Image Averaging

The image averaging is implemented in a single kernel according to (2), meaning the input images *x(n)* are transferred from the CPU to the GPU sequentially to compute the running average. The calculated averaged image *y(n)* is stored in the GPU after each pass for use in the next average calculation as *y(n-1)* while also being transferred to the CPU as an output result.

4 Methods of Performance Evaluation

The implementations of algorithms in the FPGA and the GPU are evaluated in terms of achievable processing data rates and power consumption as two of the most important features of an imaging system. With higher power consumption, usually higher processing data rates are achieved. Therefore, for comparing the efficiency of the processing, the ratio between the achievable processing data rate and the power consumption is considered as well.

The image filtering is evaluated for filter sizes of 3x3 to 15x15. The filters' coefficients are chosen as Gaussian low-pass filters with standard deviation equal to one-sixth of the filter size. The image averaging is evaluated for calculating the running average of 8 images (*L=8*).

4.1 Methods of FPGA Evaluation

For the FPGA, the achievable processing data rates are calculated based on the measurements of the frame rate acquired by the PC from Fig. 1 and the image size used, namely 1936 by 1216 pixels at 8 bit resolution.

The power consumption of the FPGA is estimated with the Microsemi's tool for power estimation [8], which calculates the consumption based on the resource usage of the FPGA and the operational frequency. The resource usage is acquired from the output of the synthesis tool of Libero. For the overall power consumption, the power consumption of the DDR memory needs to be added when it is used. The power consumption of the DDR memory is estimated based on Micron's power estimator tool [9] for the DDR memory used [5].

4.2 Methods of GPU Evaluation

For the GPU, the processing data rate is calculated based on the image size and the measured processing frame rate. The frame rate is acquired by timing the execution of the implemented algorithm. To get a precise time measurement, the average of a 1000 executions in a sequence is taken. The whole execution time, including the memory transfers, is taken into account. The CUDA Runtime API [10] is used for timing the algorithm execution. For measuring the power consumption of each algorithm, Nvidia's NVML API [11] is used. The API can periodically poll the GPU and log the power drawn by during runtime. For each implementation, the logging is initiated before the start of the loop which calls the implemented algorithm iteratively until 1000 executions of the algorithm have been achieved. An average of the logged power measurements is taken to obtain an average power consumption of the implemented algorithm.

5 FPGA and GPU Evaluation

The comparison of the FPGA and the GPU power consumption and achievable data rates for image filtering with different filter sizes is shown in Fig. 2. From the figure it can be seen that the GPU achieves larger processing data rates than the FPGA. These are decreasing linearly with increasing filter size. On the other hand, the processing data rates of the FPGA remain constant. This is because the processing is done at the same base frequency as the pixels are being transmitted, which is constant. However, due to convolution multiplications being done at a frequency corresponding to the filter size, only filters up to the size of 9x9 can be implemented in the FPGA. With larger filters the maximum operational frequency of 400 MHz [5] would be exceeded. Despite this, theoretical achievable values are shown in the figure for completeness (marked with a light pattern). Moreover, from the figure it can be seen the larger data rates of the GPU come at the cost of larger power consumption than in the FPGA. For both platforms, the consumption is increasing with filter size.

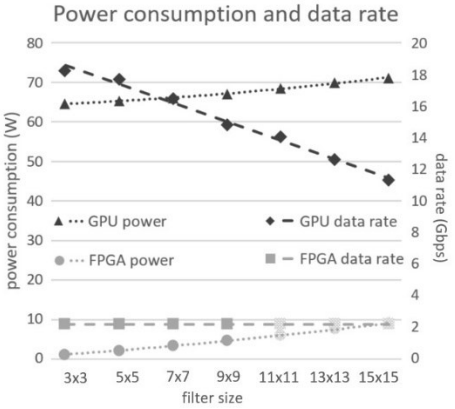


Fig. 2. Power consumption and data rate comparison between FPGA and GPU implementations of image filtering for different filter sizes.

In order to get a representation of the efficiency of both platforms, the ratio between achievable data rates and power consumption for image filtering is presented in Fig. 3. The figure shows that the ratio for both cases decreases with filter size, meaning that both platforms are less efficient for larger filter sizes. However, the decrease for the FPGA is inversely proportional to filter size as compared to the GPU, which decreases linearly. The result is that the ratio of FPGA to GPU efficiency falls from a factor of 6 to 1.5.

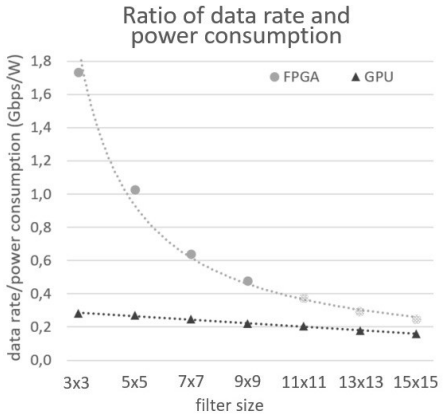


Fig. 3. Ratio of data rate and power consumption for FPGA and GPU implementations of image filtering for different filter sizes.

The evaluation for the image averaging is summarized in Table 1. The data rates of the GPU are up to 12 times higher than that of the FPGA. However, the FPGA power consumption is more than 60 times lower.

Table 1. Evaluation of Image Averaging Algorithm

Platform	Data rate [Gbps]	Power [W]	Data rate/power [Gbps/W]
FPGA	2.23	0.97	2.29
GPU	27.56	64.9	0.42

6 Discussion

The evaluation shows that the developed FPGA system is more efficient in terms of achievable data rate per watt compared to using a GPU for the processing. Hence, it is an appropriate choice for low-power applications. Nevertheless, the FPGA based system has more severe limitations compared to a GPU. These limitations are: frequency limitations, available resources, DDR memory throughput, and lower precision.

The maximum frequency of the FPGA limits the filter size for image filtering with the employed implementation. A different implementation could be used, where all the multiplications are done in parallel at the base processing frequency. However, more resources would be used, limiting the complexity of algorithms that can be implemented. The limitation in the available resources also affects the precision of the processed result. If more resources are available, intermediate calculations and filter coefficients could be stored with increased precision. However, more resources also mean increased power consumption. For practically all applications, the precision achieved by the used implementation is satisfactory.

Moreover, the DDR memory throughput of the FPGA board limits the complexity of algorithms that can be implemented as demonstrated by the implementation of the image averaging. Due to the limitations in memory throughput, the algorithms might need to be simplified for implementation.

The same limitations as discussed for the FPGA apply for the GPU. However, the operating frequency is higher; it includes more DDR memory with increased throughput and dedicated resources for image processing (e.g. floating-point multipliers). This increased performance comes at the cost of increased power consumption. Therefore, careful consideration of both platforms is necessary when designing real-time imaging application.

7 Conclusion

This paper described an FPGA based system as a low-power solution for upgrading an existing imaging system to achieve real-time processing of images. The developed system can be inserted between the camera and the existing receiver. The presented solution has been demonstrated on an existing imaging system that uses a camera with the Camera Link I/F. The processing capabilities of the developed system were evaluated in terms of achievable processing data rate and power consumption, and compared to the performance of a GPU RTX 2080 Ti.

The evaluation of the system's performance and its comparison with a GPU showed that the developed low-power FPGA based system is a valid approach when upgrading existing imaging systems to achieve real-time processing of images. Nevertheless, for applications of high complexity, the evaluation indicated that the choice between using an FPGA and a GPU is not straightforward. Therefore, an investigation of FPGA and GPU performance for high complexity applications shall be conducted in the future.

References

[1] B. Brown, "Camera Connections," *Vision Systems Design*, 01-Apr-2008. [Online]. Available: www.vision-systems.com/non-factory/environment-agriculture/article/16739233/camera-connections. [Accessed: 22-Sep-2019].

[2] C. Li, S. Balla-Arabe, and F. Yang-Song, *Architecture-Aware Optimization Strategies in Real-time Image Processing*. Great Britain and USA: John Wiley & Sons, Inc., 2017.

[3] M. Shafique and J. Henkel, *Hardware/Software Architectures for Low-Power Embedded Multimedia Systems*. Springer New York, 2011.

[4] Nvidia Corporation, "GeForce RTX 2080 Ti." [Online]. Available: www.nvidia.com/de-de/geforce/graphics-cards/rtx-2080-ti/. [Accessed: 25-Jan-2020].

[5] Microsemi, "UG0557: User Guide SmartFusion2 SoC FPGA Advanced Development Kit." 2016.

[6] U. Hudomalj, "Real-Time Processing of Image Streams," Maser's Thesis, University of Ljubljana, 2019.

[7] U. Hudomalj, C. Mandla, and M. Plattner, "FPGA Implementations for Real Time Processing of High Frame Rate and High Resolution Image Streams," in *IEEE International Conference on Computing, Electronics & Communications Engineering 2020*. to be published.

[8] Microsemi, "Microsemi Power Estimator (MPE) - v10d: SmartFusion2 and IGLOO2." 2015.

[9] Micron Technology, Inc., "System Power Calculators." [Online]. Available: www.micron.com/support/tools-and-utilities/power-calc. [Accessed: 26-Jan-2020].

[10] Nvidia Corporation, "CUDA Runtime API" 2019.

[11] Nvidia Corporation, "NVML: Reference Guide vR440". 2019.

Prenos pristopov uporabljanih pri razvoju programske opreme v EDA okolja

Mitja Nemec

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: mitja.nemec@fe.uni-lj.si

Software development practices in EDA environment

Abstract. *This paper presents how the processes which are well known from software development practices are slowly penetrating into design of printed circuit boards. The biggest impacts come from source code management tools which form a foundation for more advanced tools and services. Open source also has an effect by exposing EDA environment to developers with different background than electrical engineering which generates fresh and completely unheard of solutions.*

1 Uvod

Orodja za računalniško podprto načrtovanje električnih vezij (Electronic Design Automation – EDA oziroma Electronic Computer-Aided Design - ECAD) se v grobem delijo na orodja za:

- načrtovanje integriranih vezij
- načrtovanje tiskanih vezij

V tem članku se bomo osredotočili na orodja za načrtovanje tiskanih vezij in osnovna funkcionalnost ki jih ta orodja omogočajo je zajem električne sheme in postavitev elementov na vezje ter avtomatsko ali ročno ustvarjanje bakrenih povezav na podlagi električne sheme. Za načrtano tiskano vezje se nato generira datoteke, ki jih proizvajalci rabijo za njegovo izdelavo (gerber, drill, …). V večini orodja omogočajo tudi izdelavo 3D modela tiskanega vezja skupaj s komponentami kar omogoča integracijo z mehaničnimi CAD programi.

Orodja tudi omogočajo vzpostavitev in vzdrževanje knjižnic s simboli in odtisi komponent, naprednejša orodja pa lahko knjižnico simbolov povežejo s podatkovno bazo kar omogoči dodatne funkcionalnosti (sledenje zalogam, pregled dobavljivosti, izdelava stroškovnika, …).

Razvoj EDA orodij se je začel v osemdesetih letih prejšnjega stoletja, ki so tekom desetletji prerasla v zelo zmogljiva a kompleksne orodja [1]. Smer razvoja diktirajo njihovi uporabniki, saj podjetja, ki jih razvijajo, lahko dobijo prednost na trgu le, če se njihovim uporabnikom nove funkcionalnosti, ki jih podjetje ponudi, zdijo vredne investicije. Tako lahko v grobem trdimo, da inženirji elektrotehnike v večjem delu usmerjajo razvoj EDA orodij.

V zadnjem desetletju pa je opazen napredek teh orodij tudi na področjih ki so bliže računalništvu in informatiki in v nadaljevanju si bomo pogledali v katero smer lahko pričakujemo, da se bodo ta orodja razvijala v prihodnosti.

Pri odprtokodnih EDA orodjih je smer razvoja javno objavljena [2] in tudi uporabniki teh programov velikokrat objavijo na kak način jih uporabljajo [3]. Tako smo lahko pri odprtokodnih EDA orodij tudi dobro seznanjeni z samim procesom načrtovanja električnih vezij.

Zaradi zaprte narave zaprtokodnih EDA orodij bistveno težje sodimo o smeri njihovega razvoja in tudi uporabniki zaprtokodnih orodij ne razkrivajo na kak način jih uporabljajo. Tako je proces načrtovanja vezij pri katerih se uporablja zaprtokodna EDA orodja javnosti neznan.

Tako velja podati opozorilo, da je podana raziskava, zaradi teh dejstev, zelo verjetno pristranska.

2 Vpliv odprte kode

V devetdesetih letih prejšnjega stoletja so se pojavila prva odprtokodna EDA orodja. Pri odprtokodnih EDA okoljih pa pri razvoju sodelujejo tako inženirji elektrotehnike, ki poznajo tudi področje programiranja, kot tudi inženirji računalništva, ki poznajo področje elektronike. In prav slednji so domači s pristopi s katerimi se inženirji elektrotehnike tipično nikoli ne srečajo.

Odprte specifikacije in predvsem odprt in dokumentiran format zapisa posameznih datotek (shema, …) omogoča, da se lahko kdorkoli loti razvoja samostojnih aplikacij, ki se povežejo z EDA orodjem in tako razširijo nabor funkcionalnosti.

Tako se v/okoli odprtokodnih EDA okoljih pogosto pojavijo rešitve, ki izvirajo s področja računalništva in informatike. Nekatere od njih so uporabne in jih kasneje povzamejo tudi druga EDA orodja.

3 Preneseni pristopi

3.1 Vpeljava standardov

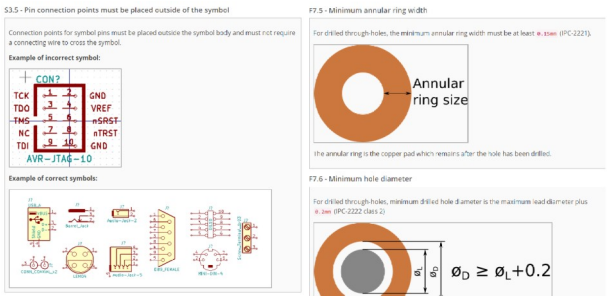
Pri razvoju programske opreme je pregled kode [4]–[6] del procesa. In za učinkovitost pregledov kode je treba kodo pisati na podlagi standardov, ki določajo slog, obliko in nabor funkcionalnosti jezika ki se smejo oziroma ki se ne smejo uporabljati [7], [8].

Na področju načrtovanja je ERC (electrical rules check) funkcionalnost na voljo že nekaj časa vendar pa le ta preverja samo električno smiselnost sheme (kratek stik med dvema napajalnima linijama, manjkajoče povezava …). Praktično enako stanje je na področju načrtovanja električnih vezij. DRC (design rules check) funkcionalnost preverja ali je vezje primerno za

izdelavo (dovolj velik razmik med bakrenimi povezavami, širina bakrenih povezav, …). Nekoliko bolje je na področju izdelave odtisov komponent. V zadnjih dveh desetletjih sta se dodobra uveljavila standarda IPC7351 in IPC-2222, ki nekoliko bolj podrobno opisujeta poleg električnih tudi ostale slogovne parametre odtisa (ime, dokumentacijski sloji, …). Z leti sta se standarda tako dopolnila, da danes obstaja že več orodij s katerimi lahko programsko generiramo odtise v celoti in ti ustrezajo standardu.

V grobem pa dlje od napotkov oziroma namigov [9], [10] še nismo prišli. Razlogov za to je verjetno več, eden izmed njih pa je zagotovo dejstvo, da je tako risanje sheme kot samega vezja v osnovi ročno grafično opravilo in računalnik tudi danes težko preverja ali je grafika po standardu ali ne. In če se grafika ne more avtomatsko generirati ali pa vsaj avtomatizirano preverjati, potem se je takih standardov težko držati. Vendar se tudi na tem področju kažejo premiki.

Eden od primerov je standard za vnos simbolov in odtisov v knjižnico [11]. Ta standard precej podrobno predpisuje kako naj bo izgled simbola (kje naj bodo napajalne nožice, kje naj bodo signalne nožice, kje naj bo tekst, kako velik naj bo tekst, poimenovanje nožic, …) in kakšen naj bo odtis komponente (lokacija nožice št. 1, rotacija komponente, velikost in položaj teksta, kateri tekst na katerem sloju, …). Ker se precej teh pravil lahko preverja z računalnikom se lahko to avtomatizira, kar pripomore k splošni sprejetosti standarda.



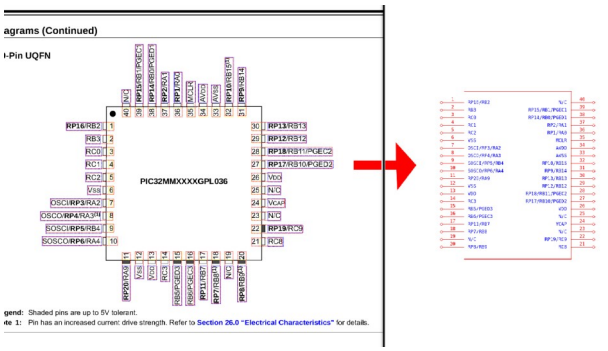
Slika 1: Izsek iz standarda KLC (KiCad Library Convention)

3.2 Programsko generiranje simbolov, odtisov, 3D modelov

Težnja po avtomatizaciji ponavljajočih procesov je v samem bistvu računalništva in informatike. To se najbolj kaže pri vseh orodjih ki se še vedno pojavljajo in so v pomoč pri ustvarjanju simbolov, odtisov in 3D modelov komponent.

Prvi primer je orodje, s katerim programsko iz PDF datoteke, ki opisuje komponento, generiramo simbol zapisan v formatu, ki ga uporablja EDA orodje [12]. Dodatno je mogoče nastaviti stil kakšen naj bo generiran simbol.

Drugi primer, je orodje s katerim lahko generiramo odtis induktivnosti [13]. Orodje omogoča da določimo število slojev, število ovojev in dimenzije odtisa ter nam generira odtis, s FEM orodjem izračuna induktivnost in upornost pri različnih frekvencah ter po potrebi tudi iterativno prilagodi dimenzije odtisa, da dosežemo želeno induktivnost.



Slika 2: Pretvorba simbola iz PDF datoteke v zapis, ki ga uporablja EDA orodje

Pri 3D modelih je osnova za avtomatizacijo CadQuerry knjižnica [14]. Z njeno pomočjo lahko s skriptami zapisanimi z jezikom python generiramo množico 3D modelov [15].

Poleg očitnega prihranka na času, ki ga omogoča avtomatizacija pri izdelavi simbolov, odtisov in 3D modelov, je dodatna prednost tudi zmanjšana verjetnost za napake zaradi človeškega faktorja.

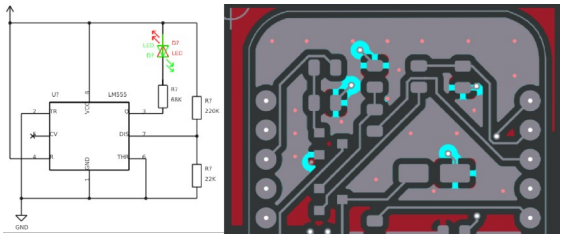
3.3 Pretvorbe med različnimi zapisi

Odprti formati zapisov datotek EDA okolji omogočajo pretvorbo med različnimi podatkovnimi tipi, ki tipično niso direktno podprta. To omogoča direktno povezavo različnih orodij [16], [17] brez uporabe vmesnih bolj standardnih zapisov (.step, dwg, .dxf, ...). Hkrati pa take pretvorbe omogočajo dostop do tehnologije izdelave tiskanih vezij področjem, ki temu sicer niso izpostavljena (umetnost, ...) [18].

3.4 Orodja za sledenje spremembam

Osnovna funkcionalnost, ki jo orodja za nadzor kode (SCM – source control management) nudijo je shranjevanje in sledenje različnim verzijam kode. Le to postane še bolj uporabno z orodij, ki omogočajo primerjavo dveh verzij (diff). Na področju računalništva imajo ta orodja že dolgo zgodovino in v tem času so šla čez več evolucijskih korakov. V zadnjem času orodje »git«, ki omogoča distribuiran nadzor kode, postaja de-facto standard na tem področju.

Ker orodja za nazor kode lahko delujejo nad datotekami v katerem koli zapisu so se pojavila različna orodja, ki omogočajo primerjavo različnih verzij [19]–[22].

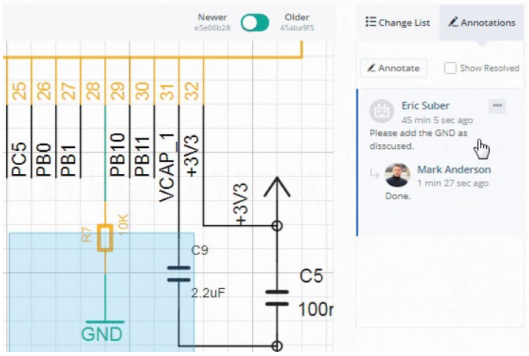


Slika 3: Grafični prikaz razlik na shemi na tiskanem vezju

3.5 Orodja in storitve za kolaboracijo

Podpora orodjem za nadzor kode pa je omogočila tudi razvoj novih storitev. Tako so se tudi za načrtovanje tiskanih vezij pojavile storitve, ki nudijo gostovanje

repozitorijev kamor se shranjujejo različne verzije. Poleg osnovnega gostovanja, te storitve podpirajo tudi pregled nad repozitorijem preko brskalnika in funkcionalnosti, ki omogočajo podpirajo sodelovanje (»pull request, veje, komentarji sprememb, …)[23], [24]. S »Pull request«-om sprožimo prošnjo za pregled in vključitev sprememb v repozitorij. Ko jih eden ali več sodelavcev pregleda in potrdi kot ustrezne, se spremembe dejansko vključijo v repozitorij. Pri tem procesu je v pomoč namenska spletna stran, na kateri lahko avtor in ostali diskutirajo o predlaganih spremembah. (slika 4).



Slika 4: Primer komentarja na spremembo sheme

Koncept vej pa je najbolj uporaben pri načrtovanju vezja. Pri zahtevnejših vezjih lahko preizkusimo več možnih razporeditev komponent ter poteka povezav in pri tem vsako možnost shranimo v svojo vejo. Tako lahko hitro prikličemo različne verzije in jih tudi med seboj primerjamo.

Nekatera EDA orodja podpirajo tudi hrambo knjižnic v repozitorijih SCM orodij. Le to omogoča lažje sodelovanje preko procesa, ki ga ta orodja podpirajo. V LRTME se pri EDA programu KiCad poslužujemo tega pristopa. Tako ima vsak študent in zaposleni svoj lokalen repozitorij s knjižnicami gradnikov. V njega lahko uporabnik vnaša nove gradnike kot tudi vnaša spremembe iz glavnega repozitorija. V glavni repozitorij pa se spremembe vnaša samo preko »pull request« mehanizma. Tako so v glavnem repozitoriju prisotni samo gradniki, ki so bili pregledani in potrjeni.

Pomembno področje kolaboracije je tudi med ECAD in MCAD orodij. V večini primerov to poteka preko datotek zapisanih v standardnih formatih (.step, idx) [25], vendar pa je ta kolaboracija zelo osnovna in v večini primerov enosmerna (iz ECAD v MCAD). Za nadgrajene možnosti ECAD-MCAD kolaboracije (dvosmerna kolaboracija, …) se pojavljajo povezave ECAD in MCAD okolji [26], ki pa so pri zaprtokodnih rešitvah tipično omejene na točno določene programe (tipično so proizvajalci lastniško povezani).

3.6 Vpeljava CI/CD pristopov

Z razmahom SCM orodij in storitev, ki temeljijo na njih, so se vzpostavili tudi CI/CD (continuous integration / continuous delivery) procesi v razvoju programske opreme[27]. Pri tem procesu se ob vsakem zapisu v repozitorij, ki se nahaja na strežniku požene avtomatiziran proces v katerem se lahko preverijo zadnje spremembe in po potrebi tudi pošljejo v uporabo.

Prvi tak primer v EDA okoljih je avtomatsko testiranje novih simbolov in odtisov v knjižnici ki je shranjena v repozitoriju. Tu se vsak simbol ali odtis preveri ali ustreza predpisanemu standardu (KLC). V kolikor avtomatski test potrdi skladnost se o tem obvesti upravljalce knjižnice, ki ob dodatnem ročnem pregledu potrdijo in simbol ali odtis se doda v knjižnico.

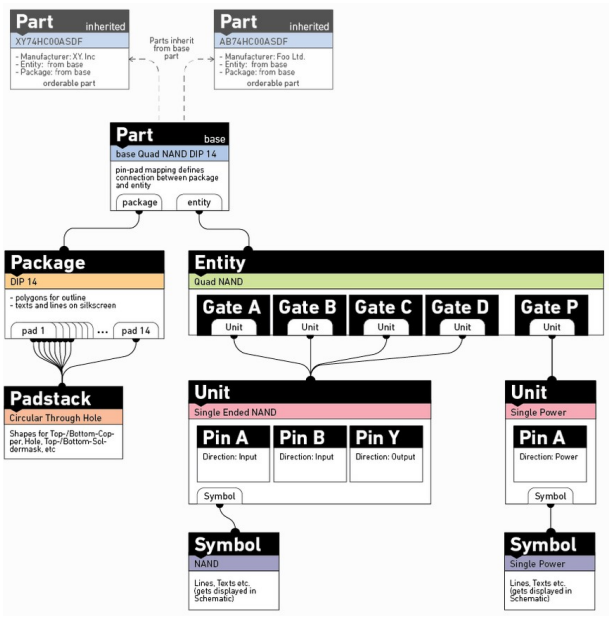
Drug primer je avtomatski test prekrivanj v 3D prostoru. Ko se zadnja verzija tiskanega vezja pošlje v repozitorij se požene test, ki preveri ali med 3D modeli komponent in ohišja prihaja do prekrivanja.

V pripravi so tudi rešitve pri katerih bi se zadnja verzija tiskanega vezja avtomatsko preverila z PDN (power delivery network) analizo, ki vključuje FEM analizo vendar pa je za avtomatiziranje tega procesa treba že na nivoju sheme vnesti dovolj specifikacij.

4 Popolnoma novi pristopi

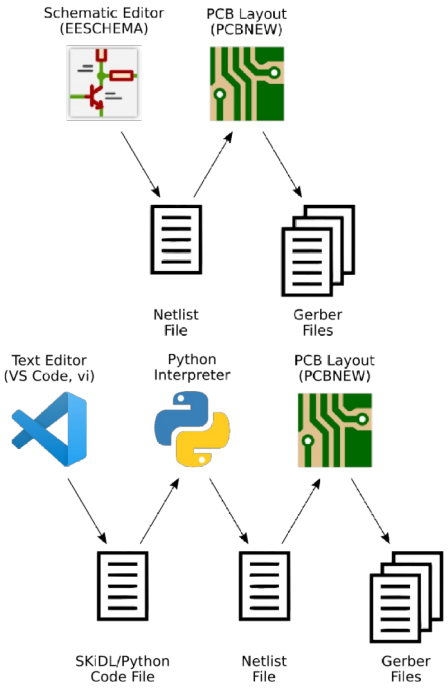
Do sedaj omenjeni pristopi predstavljajo bolj ali manj inkrementalne razvoj EDA orodij. Pojavljajo pa se tudi nekoliko bolj nekonvencionalne rešitve.

Prvi tak primer predstavlja EDA orodje [28], ki popolnoma na novo zastavi upravljanje z osnovnimi gradniki. V večini obstoječih EDA orodjih prav upravljanje z osnovni gradniki (simboli, odtisi, 3D modeli) predstavlja večino dela. Na podlagi izkušenj iz vseh teh orodij predlagan način gradnike še nekoliko detajlno razbije (slika 5). Pri simbolih večina EDA okolji podpira simbole ki so sestavljeni iz več podsimbolov. Vendar pa podsimbolov ne moremo ponovno uporabiti pri različnih simbolih. V tem primeru pa so posamezni gradniki (simbol, podsimbol, nožica, ...) neodvisni. Tako lahko na primer različni simboli uporabijo isti podsimbol (e.g. 7400 in 742G00 uporabljata isti podsimbol za NAND vrata).



Slika 5: Primer gradnikov in povezav med njimi

Še bolj eksotična rešitev pa je projekt »SKiDL« [29]. Tu avtorji predlagajo zamenjavo za prvi korak načrtovanja vezja to je vnos sheme. Namesto grafične sheme predlagajo da se medsebojne povezave med elementi vpišejo v obliki kode. Prednosti takega vnosa so očitne: bistveno bistveno lažje pregledovanje sprememb med različnimi verzijami in lažja ponovna uporaba saj se pogosto uporabljene povezave (od najbolj enostavnih kot je na primer napetostni delilnik do kompleksnih kot je na primer večkanalni DC/DC pretvornik z vso potrebno periferijo) bistveno lažje modularizirajo.



Slika 6: Klasičen proces ravzoja tiskanega vezja (zgoraj) in proces kjer se povezave med elementi zapiše s kodo (spodaj)

5 Zaključek

Kot smo lahko videli razvoj EDA orodij poteka v več smereh, dejstvo pa je, da bodo procesi, ki so dobro znani iz področja razvoja programske opreme, prodrli tudi v proces načrtovanja električnih vezij.

Pričujoči članek zagotovo ni pokril vseh področji na katerih lahko pričakujem napredek pri razvoju EDA orodij vendar pa z omejenimi informacijami, ki so na voljo, bolj celovite slike ni možno vzpostaviti.

Zagotovo pa lahko trdimo, da ko eno izmed EDA orodij vpelje nov pristop, ostala EDA orodja prej ali slej vpeljejo primerljivo (ali enako) funkcionalnost.

Zahvala

Delo je bilo sofinancirano iz programa ARRS »Pretvorniki električne energije in regulirani pogoni« P2-0258 (B).

Literatura

[1] "Electronic design automation," *Wikipedia*. Mar. 23, 2020, Accessed: Jul. 20, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Electronic_design_automation&oldid=946896501.

[2] "KiCad Roadmap · GitLab," *GitLab*. https://gitlab.com/groups/kicad/-/roadmap (accessed Jul. 20, 2020).

[3] "KiCon 2019 Talks," *KiCon 2019*. https://kicad-kicon.com/talks/ (accessed Jul. 20, 2020).

[4] T. Berling and T. Thelin, "A case study of reading techniques in a software company," in *Proceedings. 2004 International Symposium on Empirical Software Engineering, 2004. ISESE '04.*, Aug. 2004, pp. 229–238, doi: 10.1109/ISESE.2004.1334910.

[5] K. E. Wiegers, *Peer reviews in software: A practical guide*. Addison-Wesley Boston, 2002.

[6] A. F. Ackerman, L. S. Buchwald, and F. H. Lewski, "Software inspections: an effective verification process," *IEEE Softw.*, vol. 6, no. 3, pp. 31–36, May 1989, doi: 10.1109/52.28121.

[7] "MISRA C," *Wikipedia*. Jun. 25, 2020, Accessed: Jul. 20, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=MISRA_C&oldid=964441704.

[8] "C Coding Standard: Rules for Developing Safe, Reliable, and Secure Systems." SEI CERT, 2016.

[9] D. L. Jones, "PCB Design Tutorial." 2004.

[10] T. J. Sobering, "Guidelines for Drawing Schematics," p. 13, 2008.

[11] "KiCad Library Convention." https://kicad-pcb.org/libraries/klc/ (accessed Jul. 20, 2020).

[12] *uConfig*. Robotips, 2020.

[13] in3otd, *spiki*. 2020.

[14] *cadquery*. CadQuery, 2020.

[15] easyw, *kicad 3d models in freecad*. 2020.

[16] T. Lepoix, *Qucs-RFlayout*. 2020.

[17] Jānis, *pcbmodelgen*. 2020.

[18] B. P. Iyok, *svg2shenzhen*. 2020.

[19] "AllSpice: Hardware Development Tools for Electrical Engineers," *AllSpice Diff Tool*. https://www.allspice.io (accessed Jul. 20, 2020).

[20] "Eeshow." https://neo900.org/stuff/eeshow/ (accessed Jul. 20, 2020).

[21] John, *KiCad-Diff*. 2020.

[22] J.-N. Avila, *plotkicadsch*. https://github.com/jnavila/plotkicadsch, 2020.

[23] "CADLAB.io | Visual collaboration and version control platform for your PCB." https://cadlab.io/ (accessed Jul. 20, 2020).

[24] "InventHub." https://inventhub.io/ (accessed Jul. 20, 2020).

[25] "ECAD/MCAD Collaboration with IDX Standard | prostep ivip." https://www.ecad-mcad.org/ (accessed Jul. 20, 2020).

[26] easyw, *kicadStepUpMod*. 2020.

[27] "CI/CD," *Wikipedia*. Jul. 13, 2020, Accessed: Jul. 20, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=CI/CD&oldid=967493732.

[28] "horizon-eda.org." https://horizon-eda.org/ (accessed Jul. 20, 2020).

[29] XESS Corporation, "SKiDL," *SKiDL*. https://xesscorp.github.io/skidl/docs/_site/ (accessed Jul. 20, 2020).

Analiza močnostnih modulov za učinkovito krmiljenje elektromotorjev

Aleksander Sešek, Tadej Skuber, Janez Trontelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana
E-pošta: aleksander.sesek@fe.uni-lj.si

Power module analysis for efficient electric motors driving

Abstract. *The paper presents power module analysis for efficient electric motor driving optimization. The optimization bases on correct choice of module substrate, type and number of power elements, their arrangement and thermally efficient design. The main differences in power losses in the case of different number of power transistors on ceramic substrate is presented in the paper, comparing simulation results and measurements on fabricated modules.*

1 Uvod

Elektromotorji so dandanes sestavni del vse več osebnih vozil, kmetijskih pripomočkov, gradbene mehanizacije, ipd [1]–[3]. Nepogrešljivi so tudi na področjih kjer veljajo višje zahteve za njihovo natančno delovanje, porabo in povzročanje EMC motenj [4]–[6]. V skladu z vedno višjimi zahtevami po nizki porabi, uporabi novih materialov in inovativnih tehnologij so učinkovitejši in konkurenčnejši elektromotorski pogoni ključni za spodbujanje okolju prijaznejših oblik prometa in industrije [7]–[9]. Učinkovitejši elektromotorji pa za učinkovito delovanje potrebujejo še dodatne komponente. Ena glavnih komponent pri uporabi električnih pogonov je njihov natančni nadzor in krmiljenje. To sestavljata krmilna enota z mikroprocesorjem [10], [11] in močnostni modul [12], ki elektromotorju ustrezno dovaja električno energijo in s tem skrbi za pravilno delovanje po zahtevah upravitelja motorja – bodisi je to industrijska aplikacija ali električni avtomobil.

Za pravilno krmiljenje elektromotorja torej poskrbi krmilna elektronika, ki temelji na mikroprocesorju kateri s pomočjo signalov zunanjih senzoričnih komponent (krmilna palica, stopalka, računalniški program, končno stikalo, ipd) krmili izhodno stopnjo krmilnega sistema, imenovano močnostni modul.

Pri izdelavi močnostnih modulov se srečamo z naprednimi proizvodnimi procesi z visokimi zahtevami glede čistosti, zahtevnega nadzora tehnološkega procesa, bondiranja, spajkanja, površinske obdelave namestitvenih površin za močnostne elemente ter zaščite elektronski modulov pred vplivi okolja.

Bistvo kontrole izdelave in izbire optimalne sestave močnostnega modula je izboljšanje zmogljivosti krmiljenja na več ravneh, kot so dvig izkoristka elektromotorja, zmanjšanje količine vgrajenih elementov, možnost večje preobremenitve tako elektromotorja, kot tudi krmilne elektronike. Tako optimalna sestava močnostnega modula omogoča uporabo manjših elektromotorjev, kar zmanjša porabo električne energije, zavzame manj prostora in zniža ceno končnega izdelka, prav tako pa omogoča tudi boljšo detekcijo napak in preobremenitve.

V laboratoriju za mikroelektroniko, Fakultete za elektrotehniko UL (LMFE), smo v preteklosti razvili številne module, ki omogočajo učinkovito krmiljenje različnih, za določeno aplikacijo uporabljenih elektro motorjev [13], [14]. Ker so bili močnosti moduli načrtovani za različne nazivne tokove in napetosti, niso uporabni za vse motorske sisteme. Prav tako izkoristek modula in motorja za različne sisteme enakih napetostno/tokovnih parametrov ni optimalen.

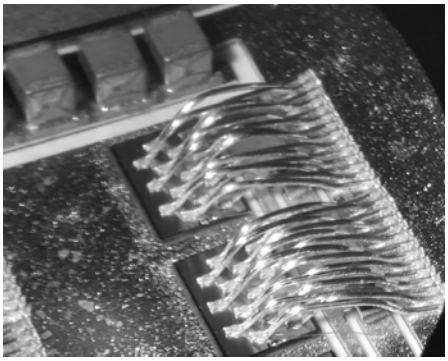
Razvoj močnostne elektronike temelji na substratu, ki je največkrat keramičen [15] in se uporablja pri kompaktnih in visokotemperaturnih aplikacijah. Substrati temeljijo na enoplastni tehnologiji DCB (ang. »direct copper bonding«). Na substratu, na katerem je nanešen baker, srebro ali zlato, se glede na tokovne zahteve namesti polprevodniške krmilne elemente – tranzistorje različnih materialov in tehnologij. Novejši tranzistorji so iz galijevega nitrida (GaN) [16] in silicijevega karbida (SiC) [17], ki prenesejo višje temperature in napetostne obremenitve, vendar zahtevajo kompleksno krmiljenje, tehnologija njihove izdelave pa je kompleksnejša. Zato se v osnovnih razvojnih sistemih uporabljajo najbolj dostopni tranzistorji MOSFET, ki imajo nizke izgube, tehnologija njihovega krmiljenja pa je enostavnejša in že zelo razvita [18].

V LMFE smo izvedli simulacije z uporabo različnega števila močnostnih elementov in ugotavljali optimalno izbiro za določen električni motor. Načrtali in izdelali smo tudi prototipni močnostni modul z močnostnimi tranzistorji, na katerem smo izvedli osnovne meritve segrevanja. V članku so najprej predstavljeni rezultati simulacij za module, pri katerih so bili uporabljeni 1, 2, 3 in 4 močnostni tranzistorji in uporabljen vodni hladilni sistem. Predstavljeni so tudi simulacijski rezultati segrevanja teh modulov, ki je posledica trošenja moči zaradi izgub. V ločenem podpoglavju so predstavljene tudi meritve modula z močnostnimi tranzistorji, kjer je poudarek na slabem odvajanju toplote zaradi nekvalitetnega hladilnega sistema.

2 Močnostni modul in izgube

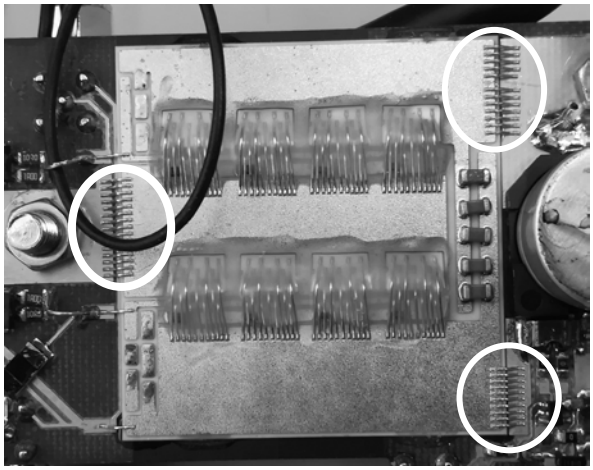
Poudarek pri predstavljenih simulacijah je na prikazu spremembe velikosti izgub pri uporabi različnega števila močnostnih elementov. Izbira števila močnostnih

elementov je nekakšen kompromis med nazivno vrednostjo toka, ki jo močnostni element lahko prenese, velikostjo termičnih izgub na elementu (ki je posledica treh parametrov, ki bodo prestavljeni v nadaljevanju) ter seveda cene modula. Velikost in oblika modula je v prvi vrsti prilagojena razpoložljivemu prostoru ob motorju, hladilniku ter načinu priklopa, ter seveda številu tranzistorjev na fazo pogona. To število se giblje od 1 pa tja do 12 močnostnih tranzistorjev – 6 za vsako smer toka. V LMFE razvijamo module, ki delujejo na principu mostiča torej smer toka razumemo kot prevodno pot od pozitivnega napajanja do motorske faze ter od faze pa do negativnega napajanja oziroma ozemljitve.



Slika 1. Močnostni tranzistorji na modulu povezanimi z aluminijasto bondirno žico - bondom

Na sliki 1 vidimo spodnji del modula in sicer je ponor tranzistorjev prispajkan [19] na bakreni fazni vodnik na keramičnem modulu in povezan na bakreni vodnik negativnega napajanja z aluminijastimi bondirnimi žicami. Posebej je povezan tudi krmilni priključek tranzistorja, vendar s tanjšo bondirno žico. Tak sistem – baker-spajka-tranzistor-bond-baker predstavlja osnovo za izračun izgub na modulu. Poleg predstavljene tokovne poti pa moramo povezati tudi modul na glavni napajalni priključek, kar prav tako storimo z močnostnimi bondi, kot je prikazano na sliki 2.

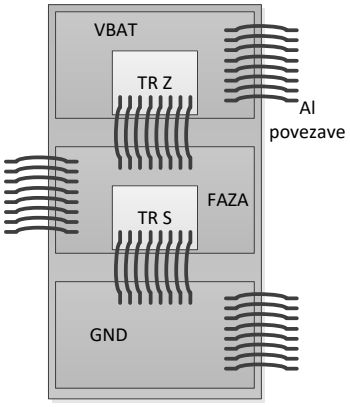


Slika 2. Močnostne povezave na napajalne priključke in fazni vodnik (označeno s krogci)

Na sliki 2 lahko vidimo zaščitno plast na močnostnih modulih, ki preprečuje korodiranje, oksidacijo in vnos delcev med povezave, hkrati pa poslabša hlajenje elementa.

2.1 Simulacije modulov

Simulacije za oceno izgub so bile narejene z uporabo 24V tranzistorjev, z maksimalnim dovoljenim tokom 100A ter zelo učinkovitim vodnim hladilnim sistemom. Sistem smo razdelili na 5 skupin in sicer zgornja bakrena povezava VBAT, srednja fazna povezava FAZA, spodnja povezava GND ter dva tranzistorja, TR Z ter TR S. Sistem je prikazan v blok diagramu na sliki 3.



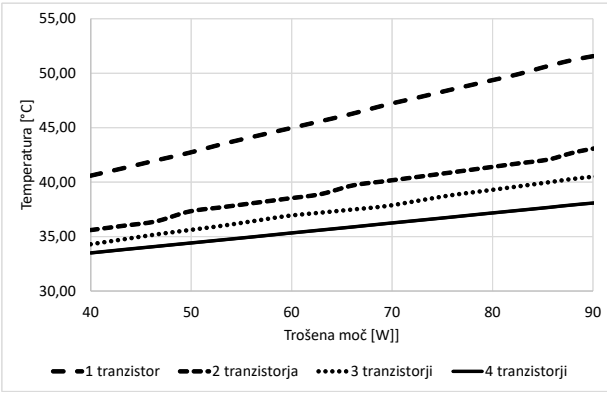
Slika 3. Blok diagram uporabljen za oceno izgub

Za izgube VBAT štejejo izgube na bakreni povezavi VBAT, vključno z izgubami na bondih za dovod baterijskega napajanja. Za izgube na priključku FAZA se upoštevajo izgube na bakreni povezavi, bondih na motorni priključek ter na tranzistorskih TR Z bondih. Izgube na GND vključujejo bakreno povezavo GND ter dve seriji bondov na negativni priključek baterije ter na tranzistor TR S. Rezultati simulacij s tokom 100A po fazi so zbrani v tabeli 1.

Tabela 1: Izgube po sklopih

Št. tranzistorjev		1	2	3	4
VBAT izgube [W]		0,71	0,76	0,75	0,80
FAZA izgube [W]		2,51	1,64	1,53	1,30
GND izgube [W]		2,24	1,61	1,32	1,26
TR Z izgube [W]	diodne	6,56	3,82	2,64	2,2
	stikalne	1,90	1,28	1,5	1,28
	kanal	6,36	3,18	2,07	1,6
	skupaj	**14,83**	**8,28**	**6,24**	**5,08**
TR S izgube	diodne	6,51	3,92	2,58	2,16
	stikalne	1,90	1,5	1,62	1,44
	kanal	6,29	3,12	2,04	1,64
	skupaj	**14,70**	**8,54**	**6,24**	**5,28**

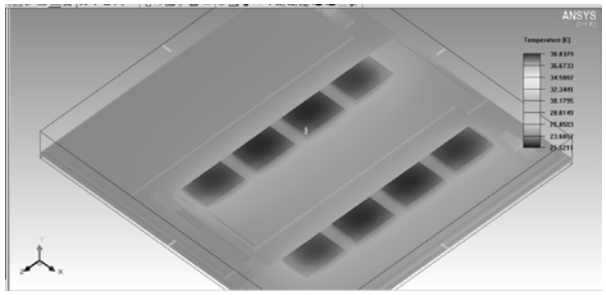
V tabeli 1 so izgube po sklopih definirane za različno število tranzistorjev, kjer so tranzistorske izgube nadalje razdeljene na izgube zaradi preklopov (stikalne), izgube zaradi upornosti kanala (kanal) ter zaradi toka preko diode, kadar je tranzistor zaprt – induktivno breme (diodne). Iz tabel je razvidno, da se z uporabo več tranzistorjev izgube manjšajo, ker se tok porazdeli po tranzistorjih in bondih. Raziskave, ki smo jih opravili v preteklosti, pa so pokazale, da je uporaba petih tranzistorjev ali več nesmiselna, ker je cenovno in prostorsko neugodna, prav tako pa ne doprinese več veliko k skupnem zmanjšanju izgub.
Predstavljene izgube se seveda odražajo kot gretje modula. Zato je ena od pomembnejših nalog načrtovalca sistema poskrbeti za odvajanje toplote z ustreznim hladilnim sistemom. Za predstavljene rezultate smo uporabili zelo učinkovit hladilnik z vodnim hlajenjem. Na sliki 4 so prikazane krivulje gretja modula ob uporabi različnega števila tranzistorjev.



Slika 4. Temperaturne krivulje različnih modulov

Kot vidimo, se moduli pri istem vhodnem toku segrejejo precej bolj če uporabljamo en tranzistor ali pa štiri. Razmerje izgub se ujema s prejšnjimi simulacijami. Z uporabo štirih tranzistorjev dosežemo zmanjšanje temperature za 19 stopinj, še vedno pa je pomembno dobro odvajanje toplote preko hladilnika.

Simulacije so bile opravljene v programskem okolju Ansys. Na sliki 5 je predstavljena termična slika modula v podprogramu Icepack.

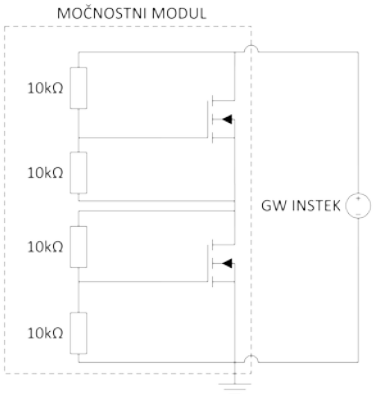


Slika 5. Simulacija segrevanja modula s štirimi tranzistorji v okolju Ansys Icepack

V simulaciji smo modul obremenili z 200A. Pričakovana moč, ki naj bi se trošila na hladilniku, je bila 100W. Hladilno telo, na katero je modul nameščen, je bilo ohlajeno na 30°C. Kot vidimo, se močnostni tranzistorji segrejejo na 39°C. Prav tako smo pri simulacijah ugotovili, da se kvocient razmerja temperatur 1:1,3 ohranja. Torej bi se v primeru, da ne uspemo ohranjati 30°C na hladilniku, ampak 50°C, kar je bolj realno, temperatura močnostnih elementov dvignila na 65°C. Nizka razlika v temperaturi je posledica dobrega odvajanja toplote.
Da bi potrdili termične izgube na spoju substrat-spajka-tranzistor, smo izdelali prototip modula, ki je predstavljan v naslednjem poglavju.

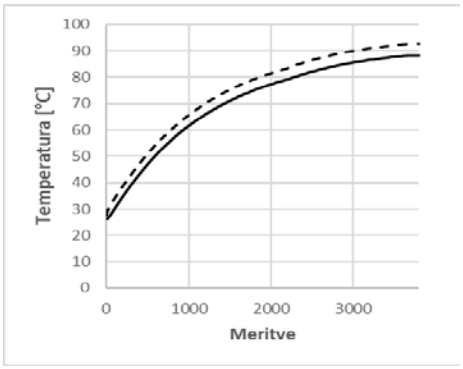
2.2 Meritve modula

Za meritve termičnih izgub smo tranzistorje priključili, kot kaže shema na sliki 6.



Slika 6. Shema testne priključitve modula

Modul je bil nameščen na majhno hladilno telo brez prisilnega vpihovanja, zato smo pričakovali močno segrevanje, prav tako pa so tranzistorji na sliki 6 vedno odprti in predstavljajo le ohmsko breme (stikalnih in diodnih izgub tu ni). Modulu smo dovajali dovolj toka, da se je na njem trošilo 20W moči. Krivulja segrevanja je prikazana na sliki 7.



Slika 7. Temperaturni krivulji za tranzistor in za hladilnik

Meritve, prikazane na sliki 7, so bile opravljene vsako sekundo, v času ene ure. Kot vidimo, se tako modul kot hladilnik segrevata in na koncu dosežeta stacionarno stanje, in sicer je tranzistor segret na 92.7°C, hladilnik pa na 88.3°C. Razlika 4.4°C nam pove, da se toplota preko substrata in spajke slabše odvaja, kot bi želeli, predvsem je problem v stiku substrat - hladilnik, kjer se toplota prevaja preko tankega sloja termične paste, ki ni idealna, prav tako se površini popolnoma ne prilegata. Hladilnik uporabljen v tem primeru, je premajhen in nima dodatnega hlajenja.

3 Povzetek

Članek je povzel rezultate simulacij močnostnih modulov, kjer smo želeli predstaviti termične razmere pri uporabi različnega števila močnostnih elementov ter uporabe učinkovitega hlajenja. Meritve na prototipnem modulu s slabim hlajenjem so pokazale, da se temperatura zlahka dvigne čez 90°C že pri izgubah 20W, medtem ko se po simulacijah sodeč lahko temperaturo z učinkovitim hlajenjem zadrži precej nižje. Naslednji korak je testiranje modula z enakim hladilnikom kot je simulirani, da se potrdi simulacijske rezultate. Prispevek je del večjega projekta izdelave ekspertnega sistema, ki bo omogočal izbiro idealnega sestava za krmiljenje električnih motorjev.

4 Zahvala

Avtorji se zahvaljujejo Ministrstvu za izobraževanje, znanost in šport za sofinanciranje projekta: »Učinkovitejši elektromotorji z razvojem ekspertnega sistema in novih tehnologij (MOTZART)«, v okviru operativnega programa za izvajanja evropske kohezijske politike v obdobju 2014-2020.

Literatura

[1] ‘Electric Motors for Electric Vehicles 2013-2023: Forecasts, Technologies, Players: IDTechEx’, Mar. 04, 2013. https://www.idtechex.com/en/research-report/electric-motors-for-electric-vehicles-2013-2023-forecasts-technologies-players/344 (accessed Sep. 03, 2020).

[2] M. Bloom, ‘The Brushless DC Motor and Its Use in Electric Cars’. http://buildipedia.com/aec-pros/design-news/the-brushless-dc-motor-and-its-use-in-electric-cars (accessed Sep. 03, 2020).

[3] E. Adams, ‘The Secrets of Electric Cars and Their Motors: It’s Not All About the Battery, Folks’, *The Drive*. https://www.thedrive.com/tech/17505/the-secrets-of-electric-cars-and-their-motors-its-not-all-about-the-battery-folks (accessed Sep. 03, 2020).

[4] ‘Brushless dc Motors in Medical Applications’, *Machine Design*, Jun. 02, 2009. https://www.machinedesign.com/mechanical-motion-systems/article/21829410/brushless-dc-motors-in-medical-applications (accessed Sep. 03, 2020).

[5] ‘Motors and Drives for Advanced Prosthetics’. https://www.electrocraft.com/motors-for/medical/advanced-prosthetics/#product-cats (accessed Sep. 03, 2020).

[6] ‘Motors and Drives for Surgical Robots - ElectroCraft’. https://www.electrocraft.com/motors-for/medical/surgical-robots/#product-cats (accessed Sep. 03, 2020).

[7] ‘Advantages of DC motors in industrial applications’, *Control Global*. https://www.controlglobal.com/blogs/guest-blogs/advantages-of-dc-motors-in-industrial-applications/ (accessed Sep. 03, 2020).

[8] J. A. Tudor and D. M. Platt, ‘The application of electric motors in the pulp and paper industry’, *Electrical Engineering*, vol. 71, no. 1, pp. 31–35, Jan. 1952, doi: 10.1109/EE.1952.6437885.

[9] B. R. Shover, ‘The industrial application of the electric motor, as illustrated in the gary plant of the Indiana steel company’, *Proceedings of the American Institute of Electrical Engineers*, vol. 28, no. 3, pp. 185–230, Mar. 1909, doi: 10.1109/PAIEE.1909.6660243.

[10] J. Podržaj, A. Sešek, and J. Trontelj, ‘Intelligent power MOSFET driver ASIC’, in *2012 Proceedings of the 35th International Convention MIPRO*, May 2012, pp. 107–111.

[11] R. Vrtovec and J. Trontelj, ‘Advanced gate control system for power MOSFET switching losses reduction with complete switching sequence control’, *Informacije MIDEM*, vol. 46, no. 4, p. 12, 2016.

[12] J. Podržaj and J. Trontelj, ‘Optimized selection of materials and components for power module realization’, *Informacije MIDEM*, vol. 40, no. 3, pp. 178–182, 2010.

[13] I.-I. informacijskih znanosti Maribor, ‘Motor driver power stage/module simulation; Elektronski vir’. https://plus.si.cobiss.net/opac7/bib/12217940 (accessed Aug. 26, 2020).

[14] ‘Design consideration for power modules of electro-motor drives’. http://www.dlib.si/details/URN:NBN:SI:doc-847E8821 (accessed Sep. 03, 2020).

[15] ‘Optimized selection of materials and components for power module realization’. http://www.dlib.si/details/URN:NBN:SI:doc-RBD3NT0D (accessed Sep. 03, 2020).

[16] I. T. AG, ‘GaN - HEMT,Gallium Nitride Transistor - Infineon Technologies’. https://www.infineon.com/cms/en/product/power/gan-hemt-gallium-nitride-transistor/ (accessed Aug. 27, 2020).

[17] I. T. AG, ‘Silicon Carbide, MOSFET Discretes - Infineon Technologies’. https://www.infineon.com/cms/en/product/power/mosfet/silicon-carbide/discretes/ (accessed Aug. 27, 2020).

[18] I. T. AG, ‘OptiMOS and IRFET - Latest Family Selection Guide - Infineon Technologies’. https://www.infineon.com/cms/en/product/power/mosfet/12v-300v-n-channel-power-mosfet/optimos-and-strongirfet-latest-family-selection-guide/ (accessed Aug. 27, 2020).

[19] A. Sesek, O. Chambers, and J. Trontelj, ‘Study on the Die-Attach Voids Distribution With X-Ray and Image Processing Techniques’, *J. Electron. Packag*, vol. 141, no. 2, Jun. 2019, doi: 10.1115/1.4042804.

Telekomunikacije

Telecommunications

Razvoj UWB radarja

Marko Malajner[1], Erich Leitgeb[2], Peter Planinšič[1], Dušan Gleich[1]

[1]*UM FERI, Koroška cesta 46, 2000 Maribor*
[2]*Institut für Hochfrequenztechnik, Graz University of Technology, A-8010 Graz, Inffeldgasse 12*
E-pošta: marko.malajner@um.si

Abstract

This paper proposes an improved design of a pulse-based radar. An improved design of a pulse generator is presented using step recovery diodes and a signal mixer for the received signal. Two-step recovery diodes produce pulses of 120 ps in duration. A pulse generator is improved by removing the negative power supply, resulting in a reduced number of electronic pulses. A sampling mixer at the receiver's site receives the generated signal and stretches it from picoseconds into microseconds. The improved pulse generator is also used in the sampling mixer as a strobe pulse generator, which makes the sampling mixer much simpler. The stretched signal is then sampled by a low sample rate using an analog to digital converter. The proposed radar design achieves up to 8 GHz bandwidth and an equivalent receiving sample rate of about 100 GSa/s. The radar is controlled using a software-defined radio called Red Pitaya, which is also used for data acquisition. The proposed radar design uses widely available commercial components, which makes radar design widely available with low cost implementation.

1 Uvod

Radarji z zelo širokim radiofrekvenčnim (RF) pasom (ang. UWB - Ultra Wide Band) oddajajo in sprejemajo signale s frekvenčno širino od 500 MHz do nekaj GHz. UWB radarje odlikuje dobra prostorska ločljivost za uporabo v mnogih aplikacijah kot je geologija, nadzor življenskih funkcij, iskanje podzemnih min, itd [1]. UWB radarje lahko v osnovi delimo v dve skupini: (i) radarji, ki temeljijo na generiranju kratkih RF pulzov (časovno-domenski radarji) in (ii) radarji, ki temeljijo na frekvenčnih modulacijah (frekvenčno-domenski radarji). Slednje lahko delimo na radarje, katerim se stopnično spreminja frekvenca od minimalne do maksimalne (ang. SFCW - Stepped Frequency Continuous Wave) [2, 3] in pa radarje, katerim se frekvenca spreminja zvezno (ang. FMCW - Frequency Modulated Continuous Wave) [4]. Radarji v frekvenčni domeni imajo bolj kompliciran arhitekturo in signalno procesiranje v primerjavi z radarji v časovni domeni.

Radar v časovni domeni (pulzni radar) je sestavljen iz pulznega generatorja, detektorja pulzov, RF ojačevalnikov in anten ter enote za obdelavo signalov [5]. Na oddajni strani mora pulzni generator proizvajati pulze, ki so krajši od nanosekunde. To lahko dosežemo z uporabo plazovnih transistorjev [6], tunelskih diod [7], nelinearnih prenosnih linij [8] ali z uporabo stopnično obnovljivih diod (ang. SRD - Step Recovery Diode) [9]. Večina pulznih generatorjev uporablja SR diode, ki lahko generirajo pulse v rangu 100 ps.

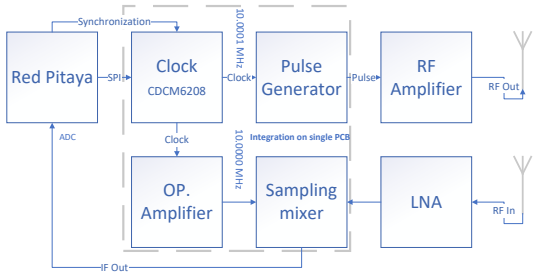
Sprejemnik pulznega radarja mora sprejeti in digitalizirati RF signale s pasovno širino nekaj GHz. Sprejemnik je bolj zapleteno narejen kot oddajnik. Za zajemanje RF signalov lahko uporabimo visoko zmogljive anologno-digitalne pretvornike (ADC). Takšen ADC mora imeti stopnjo vzorčenja več kot 10 GSa/s in je seveda zelo drag. Poleg visoke cene, takšni ADC-ji proizvedejo ogromno količino podatkov, ki jih je potrebno obdelati. Prednost pa je, da lahko delujejo v realnem času. Namesto neposrednega zajemanja z ADC-ji, lahko uporabimo t.i. ekvivaletni čas tipanja (ang. EST . Equivalent Sample Time) [10]. EST pretvorniki delujejo tako, da v vsaki periodi ponavljajočega se signala, zajamejo del tega signala in ga po določenem številu period rekonstruirajo. To pomeni, da ne zajamejo celotnega signala hkrati, ampak po delih v več periodah. Pogoj je, da mora biti zajeti signal ponavljajoč v določenem številu period. Čas zajemanja signala se poviša na račun enostavnejšega in cenejšega vezja za zajem RF signala. Z drugimi besedami: EST pretvorniki časovno raztegnejo zajeti signal, ki ga lahko potem digitaliziramo z manj zmogljivimi in cenejšimi AD pretvorniki. V tuji literaturi najdemo EST pretvornike pod imenom sampling head ali sampling mixer (vzorčevalni mešalnik) [11].

V tem članku bomo opisali načrtovanje in izdelavo celotnega predlaganega UWB radarja, ki ga odlikuje visoka zmogljivost, enostavnost, komercialno dobavlljive komponente in seveda nizka cena. V poglavju 2 bomo opisali predlagano implementacijo pulznega generatorja in sprejemnika, v poglavju 3 bomo opisali integracijo vseh komponent radarja in razvoj programske opreme. V zadnjem poglavju bomo pa podali nekaj izsledkov in zaključkov implementacije radarja.

2 Načrtovanje UWB radarja

Slika 1 prikazuje blokovno shemo predlaganega radarja. Blokci znotraj prekinjajoče črte so integrirani na eni tiskanini kot je opisano v poglavju 3. Radar je sestavljen

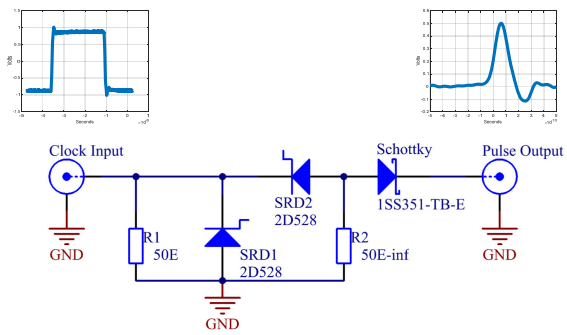
modularno, kar pomeni, da lahko priključimo željene izhodne in vhodne RF ojačevalnike in antene. Prav tako lahko izberemo poljuben AD pretvornik in mikrokrmnilnik za nadziranje radarja. V načrtu predlaganega radarja smo uporabili mikrokračnulanik Red Pitaya z integriranimi AD in DA pretvorniki [12].



Slika 1: Blokovna shema predlaganega radarja.

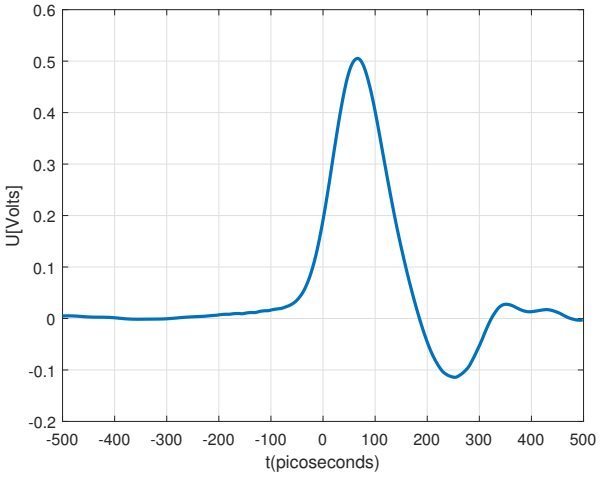
2.1 Načrtovanje pulznega generatorja

Pulzni generator smo načrtovali z uporabo SR diod. Za generiranje kratkih pulzov lahko uporabimo specifično lastnost SR diode, da lahko shrani naboj, ko tok teče v prevodni smeri. Ko tok spremeni smer iz prevodne v zaporno, SR dioda od razlike s klasično diodo, ne zapre PN spoja v trenutku, temveč ostane še določen čas prevodna. Ko zmanjka naboja, ki drži diodo prevodno, pa se zgodi takojšnje zaprtje diode. Ta vrsta diode lahko preide iz prevodnega v neprevodno stanje v manj kot 50 ps. To lastnost lahko uporabimo za generiranje zelo kratkih pulzov. Osnovo za načrtovanje generatorja smo povzeli po [9]. Generiranje pulzov smo poenostavili z uporabo generatorja preciznega pravokotnega urinega signala [13]. Shema predlaganega generatorja je na sliki 2. Generator sestavljata dve paralelno vezani SR diodi. Izhodna Schottky dioda je namenjena oblikovanju izhodnega pulza. Upor na vhodu služi 50 Ohmski impedančni prilagoditvi, vrednost upora R_2 lahko spreminjamo in s tem vplivamo na širino izhodnega pulza.



Slika 2: Generator kratkih pulzov

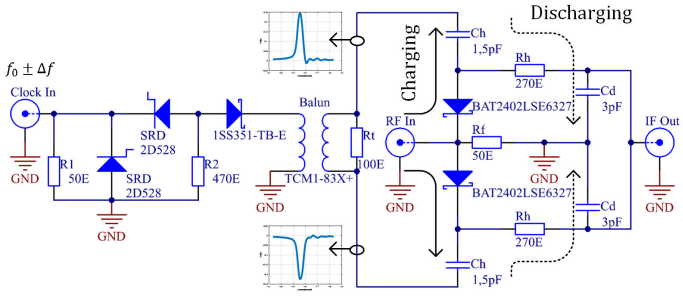
Slika 3 prikazuje izmerjen pulz na izhodu generatorja. Širina pulza je 120 ps, kar zasede pasovno širino od DC do 8 GHz.



Slika 3: Izmerjen pulz na izhodu generatorja

2.2 Načrtovanje vzorčevalnega mešalnika

Pulzni generator je lažje načrtovati in izdelati kot pa vzorčevalni mešalnik (na kratko: vzorčevalnik). Vzorčevalnik mora biti sposoben zajeti širokopasovni signal, v našem primeru od DC do 8 GHz. Kot smo omenili v uvodu, so direktni analogno-digitalni pretvorniki cenovno neprimerni za uporabo v radarjih. Zaradi tega, predlagamo uporabo vzorčevalnika, ki deluje na principu ekvivalentnega časa tipanja. V literaturi obstaja mnogo različnih izvedb takšnega vzorčevalnika. Kot osnovo, smo vzeli princip vzorčevalnika na dveh vzorčevalnih diodah [11] in ga nadgradili za naše potrebe.



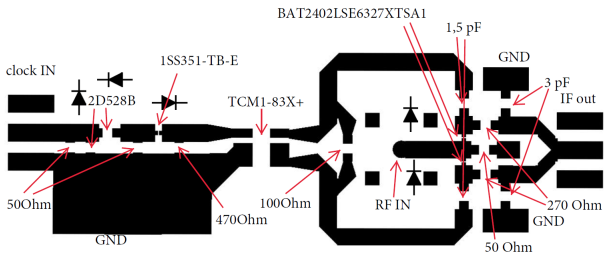
Slika 4: Shema predlaganega vzorčevalnega mešalnika

Slika 4 prikazuje shemo predlaganega vzorčevalnega mešalnika. Vhodni RF signal vzorčevalnik zajema z odpiranjem in zapiranjem hitrih Schottky diod (BAT2402). Odpiranje diod se vrši s stroboskopskimi pulzi, ki jih generiramo z generatorjem opisanim v prejšnjem poglavju. Pulzi morajo imeti zadostno veliko amplitudo, da odprejo

diode, v tem primeru najmanj 0,3 V, kolikor znaša bariera uporabljenih diod. Generiran pulz je potrebno razdeliti na dva pulza, ki imata nasprotno polariteto, da odpirata obe diodi in stem zajemamo pozitivne in negativne RF signale. Pulz razdelimo z uporabo baluna. Generetor pulzov prožimo z urinim signalom na vhodu. Frekvenca urinega signala je določena glede na frekvenco pojavnosti pulzov (ang.: PRF-Pulse Repetition Frequency) na oddajni strani. Vzorčevalnik mora sekvenčno zajemati delčke signala v vsaki periodi, kar pomeni, da mora biti sinhroniziran z oddajnikom. Frekvenca ure na vzorčevalniku ($f_0 \pm \Delta f$) je za nekaj deset ali sto Hz višja ali nižja od PRF frekvence oddajnika, kar pomeni, da vzorčevalna frekvenca skenira čez frekvenco oddajnika. Faktor raztezanja signala α lahko določimo kot: $^{f_0}/_{\Delta f}$. Če izberemo $f_0 = 10MHz$ in $\Delta f = 100Hz$, dobimo faktor raztezanja 100000. Z drugimi besedami: tako nastavljen vzorčevalnik raztegne signal iz 200 ps na 20 μs.

Ko sta diodi odprti, se vhodni RF signal shrani v kondenzatorjih Ch. Ob zaprtju diod pa se shranjen signal prazni skozi upor Rh in polni kondenzator Cd. Raztegnjen signal zajamemo na obeh Cd kondenzatorjih in ga lahko digitaliziramo z uporabo manj zmogljivih AD pretvornikov.

Zaradi zelo širokega spektra signala in pa majhnih amplitud, je potrebno veliko pozornosti posvetiti načrtovanju tiskanega vezja. Pri tem smo si pomagali s simulacijskimi orodji za načrtovanje RF vezij. Rezultat simulacije je tiskano vezje na sliki 5.



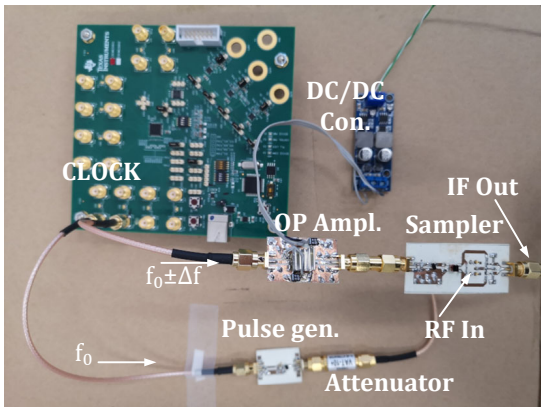
Slika 5: Tiskano vezje vzorčevalnika

Tiskano vezje je izdelano na laminatu ROGERS RO-4350B z dielektrično konstanto 3,66 in debelino substrata 0,762 mm. Izdelan vzorčevalnik, pulzni generator in generator ure CDC6208 je prikazan na sliki 6. Za izvedbo meritev smo sklenili izhod generatorja z vhodom v vzorčevalnik preko atenuatorja.

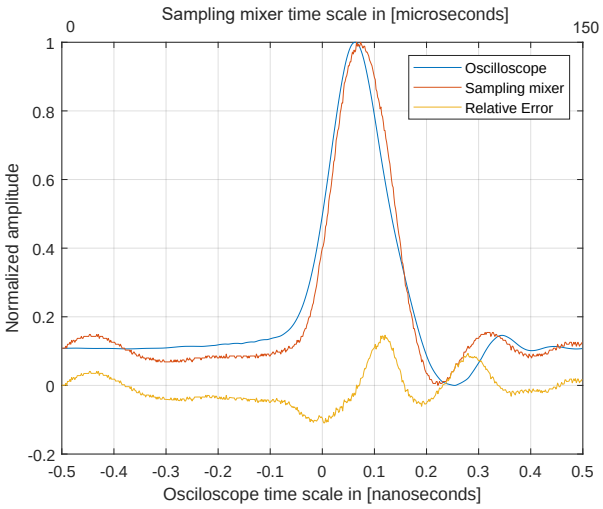
Na sliki 7 je primerjava generiranega pulza zajeta z 40 GSa/s osciloskopom in pulza raztegnjenega z vzorčevalnikom in kasneje zajetega z 2 GSa/s osciloskopom. Rumena krivulja predstavlja relativni pogrešek, ki je v povprečju manjši od 1 %.

3 Integracija komponent radarja in razvoj programske opreme

Prvi prototip radarja je bil razvit modularno, zaradi lažjega razvoja in izvedbo meritev. Ko so bili vsi sklopi razviti, smo se lotili integracije radarja na eno tiskano vezje.



Slika 6: Tiskano vezje vzorčevalnika

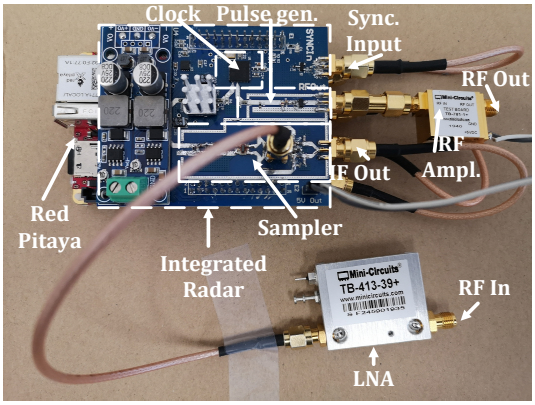


Slika 7: Primerjava zajetih pulzov z osciloskopom in vzorčevalnikom. Rumena krivulja prikazuje relativni pogrešek.

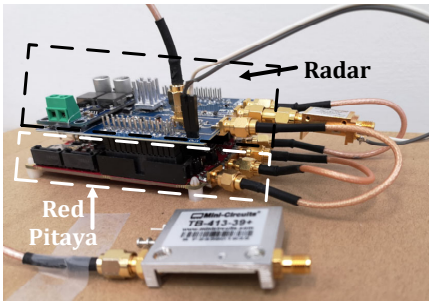
Tiskano vezje je narejeno tako, da se lahko natakne na konektorje od Red Pitaye. Red Pitaya služi kot procesna in krmilna enota za radar, omogoča tudi povezljivost preko etherneta za nadaljno obdelavo signalov. Na tiskanem vezju so integrirane komponente, ki so znotraj črtkane črte na sliki 1. Morebitni izhodni in vhodni RF ojačevalniki ter antene se priključijo na radar preko SMA konektorjev. Sliki 8 in 9 prikazujeta integriran radar priključen na Red Pitayo.

3.1 Razvoj programske opreme in grafičnega vmesnika

Razvoj programske opreme lahko razdelimo na dva dela: (i) program, ki se izvaja na Red Pitayi in skrbi za generiranje in zajemanje signalov, sinhronizacijo sprejemnika in oddajnika ter za komunikacijo preko etherneta, (ii) drugi del programske opreme se izvaja na osebnem računalniku ali laptopu in skrbi za vizualizacijo prejetih podatkov. Red Pitaya nima svojega grafičnega vmesnika, zato pošiljamo podatke preko etherneta na drugo napravo, kjer

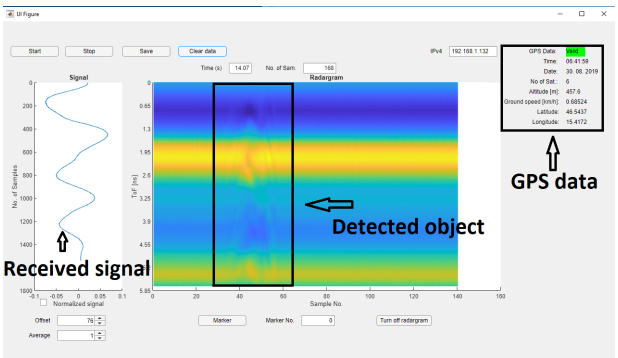


Slika 8: Radar (modra ploščica) nataknjen na Red Pitayo



Slika 9: Stranski pogled na celoten radar

lahko vizualiziramo podatke. Za vizualizacijo radarskih podatkov smo uporabili programski paket Matlab. V Matlabu izrisujemo v realnem času sprejete signala radarja v načinu B-scan (Slika 10).



Slika 10: Grafični vmesnik radarja. Na levi strani je prikazan zajet signal na sprejemniku, v sredini so zaporedni signali združeni v radarsko sliko (B-scan). Na desni strani se prikazujejo GPS podatki, če je na radar priključen GPS modul.

4 Zaključek

V članku smo opisali razvoj in izdelavo cenovno ugodnega UWB pulznega radarja s pasovno širino od DC do 8 GHz. Pulzni generator in vzorčevalnik smo nadgradili in uporabili komponente, ki so komercialno dobavljive. Vse skupaj smo integrirali na eno tiskano vezje, ki se lahko priključi na Red Pitayo. Takšen radarski sistem ima možnost poljubnih nadgradenj, strojnih in programskih ter se lahko uporabi v mnogih aplikacijah.

Zahvala

Del prispevka je nastajal v času mobilnosti v okviru projekta CEEPUS CIII-BG-1103-04-1920-M-139006 na TÜ Graz, julij 2020.

Literatura

[1] L. Li, A. E. Tan, K. Jhamb, and K. Rambabu. Buried object characterization using ultra-wideband ground penetrating radar. *IEEE Transactions on Microwave Theory and Techniques*, 60(8):2654–2664, Aug 2012.

[2] I. Nicolaescu, P. van Genderen, K. W. Van Dongen, J. van Heijenoort, and P. Hakkaart. Stepped frequency continuous wave radar-data preprocessing. In *Proceedings of the 2nd International Workshop onAdvanced Ground Penetrating Radar, 2003.*, pages 177–182, 2003.

[3] D. Šipoš and D. Gleich. A lightweight and low-power uav-borne ground penetrating radar design for landmine detection. *Sensors*, 8(20), April 2020.

[4] G.L. Charvat. *Small and Short-Range Radar Systems*. CRC Press Inc.: Boca Raton, FL, USA, 2014.

[5] L. Liu, X. Xia, S. Ye, J. Shao, and G. Fang. Development of a novel, compact, balanced, micropower impulse radar for nondestructive applications. *IEEE Sensors Journal*, 15(2):855–863, Feb 2015.

[6] Nikolai Beev, Jonas Keller, and Tanja E. Mehlstäubler. Note: An avalanche transistor-based nanosecond pulse generator with 25 mhz repetition rate. *Review of Scientific Instruments*, 88(12):126105, 2017.

[7] A. Matiss, A. Poloczek, A. Stohr, W. Brockerhoff, W. Prost, and F. . Tegude. Sub-nanosecond pulse generation using resonant tunneling diodes for impulse radio. In *2007 IEEE International Conference on Ultra-Wideband*, pages 354–359, Sep. 2007.

[8] J. M. Johnson, J. M. Parson, D. V. Reale, A. A. Neuber, J. J. Mankowski, and J. C. Dickens. 10 kv, 44 ns pulse generator for 1 khz trigatron reprate operation of nltl. In *2014 IEEE International Power Modulator and High Voltage Conference (IPMHVC)*, pages 108–110, June 2014.

[9] L. Zou, S. Gupta, and C. Caloz. A simple picosecond pulse generator based on a pair of step recovery diodes. *IEEE Microwave and Wireless Components Letters*, 27(5):467–469, May 2017.

[10] J. Han and C. Nguyen. Development of a tunable multiband uwb radar sensor and its applications to subsurface sensing. *IEEE Sensors Journal*, 7(1):51–58, Jan 2007.

[11] C. Zhang, A. E. Fathy, and M. Mahfouz. Performance enhancement of a sub-sampling circuit for ultra-wideband signal processing. *IEEE Microwave and Wireless Components Letters*, 17(12):873–875, Dec 2007.

[12] Red pitaya. 2020.

[13] Cdcm6208 2:8 ultra low power, low jitter clock generator — ti.com. 2020.

Preizkušanje radarja v milimetrskem frekvenčnem področju za uporabo na morju

Mladen Radovanović, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: mlaki@yahoo.com

Millimeter-Wave Radar Testing for Use at Sea

Abstract. Medium/short range system on a chip (SoC radar) TI AWR1843 77-81 GHz has been evaluated onboard cruise vessel M/S Silver Muse for proof of author concepts aiming in increasing safety of life at sea. Among other tests radar has been used to measure vessel water line shape as input to dynamic vessel model monitoring occurrence of parametric roll phenomena, detecting man overboard falls, tank levels monitoring, detection of people presence inside public spaces, measurement of car position in elevator shaft and main engine vibration monitoring.

In this paper, following radar description, measurement of waterline shape as input to vessel parametric roll observer will be presented and man over board detection test.

1 Uvod

Številne pomorske nesreče z izgubo življenj in/ali materialnimi posledicami, kličejo po načrtovanju in uporabi sodobnih tehnoloških rešitev in sistemov, ki pravočasno opozorijo posadko na nevarnosti. Kot primerna rešitev se v ta namen lahko uporabljajo radarski sistemi. [1] Po uspešni uveljavitvi radarjev v milimetrskem frekvenčnem področju v avtomobilski industriji [2], postopnem prodiranju v telekomunikacije [3] in celo uporabo v satelitskih sistemih [4], je nastopil čas za njihovo uveljavitev tudi na drugih področjih, kot je na primer pomorstvo, ki je predmet tega prispevka.

Cenovno dostopen radarski sistem na čipu v milimetrskem frekvenčnem področju, ponuja možnosti načrtovanja in izdelave različnih sistemov, ki dvignejo varnost ladje in posadke na morju. [5] Z vse večjim razmahom razvoja avtonomnih plovil, postajajo radarski sistemi na čipu, ki se lahko integrirajo v bodoča avtonomna plovila, primerna rešitev tudi na tem področju.

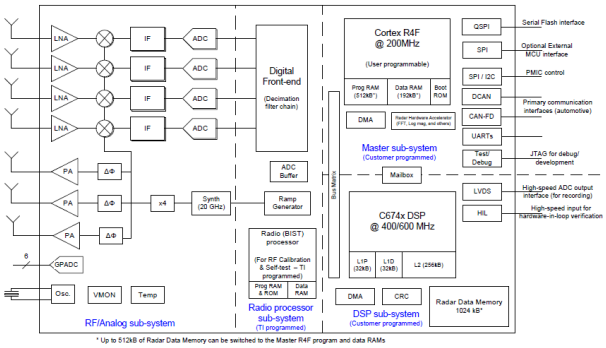
V prispevku je sprva predstavljeno delovanje radarja AWR 1843 in njegove lastnosti. Nato se članek osredotoča na uporabo na morju, pri čemer je opisan primer uporabe pri zaznavanju pojava parametrskega in ekstremnega nihanja ladje ter detekciji padca potnika v morje.

2 Delovanje radarja

AWR 1843 na čipu je integrirani radarski sistem z uporabo frekvenčno moduliranega kontinuiranega valovanja (angl. frequency-modulated continuous wave – FMCW), katerega blokovni načrt je prikazan na sliki 1. Deluje v območju od 77 do 81 GHz in ima 3 vzporedne oddajne (angl. transmitter – Tx) kanale, pri čemer ima vsak neodvisno kontrolo amplitude in faze signala. Na sprejemni strani ima 4 vzporedne sprejemne (angl. receiver – Rx) kanale s kompleksno (I,Q) arhitekturo, ki omogoča kalibracije, oziroma digitalno kompenzacijo nepravilnosti v radarju, katere so posledica nepopolne izdelave anten, umazanije na ohišju radarja, vpliva staranja komponent in podobno.

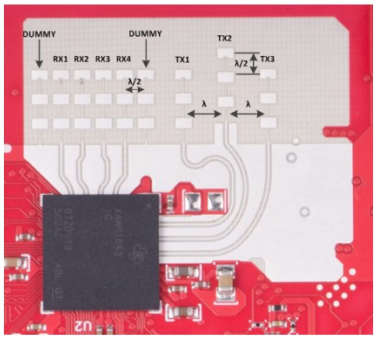
4Rx/3Tx arhitektura omogoča detekcijo objektov v dveh in treh dimenzijah. Radar pošilja položaj in hitrost zaznanih objektov prek serijskega vodila po obdelavi sprejetih signalov v signalnem procesorju (angl. digital signal processor – DSP).

Z direktnim prenosom po vzorčenju prek vodila LVDS (angl. low-voltage differential signaling) v DCA1000EVM zajemno kartico, je možna obdelava radarskih odmevov v računalniku. V programskem okolju Matlab je možno načrtovati in preveriti algoritme za specifično uporabo, ter jih namestiti v DSP-ju radarskega sistema.



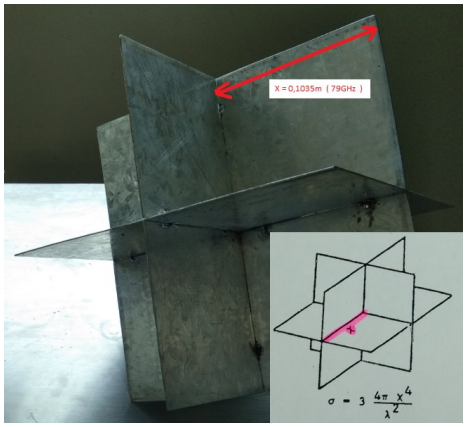
Slika 1. Blokovni načrt radarja AWR 1843

Premišljena arhitektura omogoča uporabo cenovno dostopnega radarja za številne različne aplikacije, kjer pripomore tudi njegova kompaktna zgradba prikazana na fotografiji slike 2.



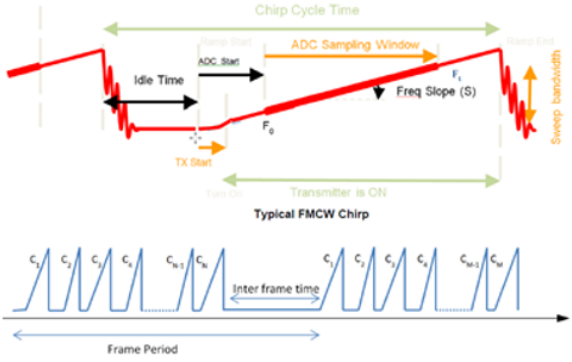
Slika 2. Fotografija radarja AWR 1843

Umerjanje radarja se izvede s postavitvijo radarskega cilja na znano lokacijo, običajno 1 m navpično od radarja. Za umerjanje in različne poskuse je bil izdelan kvadratni radarski reflektor, prikazan na sliki 3, z radarsko odmevno površino 300 m^2 na frekvenci 79 GHz.



Slika 3. Radarski reflektor

Radar oddaja okvirje, ki so sestavljeni iz niza kratkih frekvenčno linearnih moduliranih signalov (chirpov oziroma preletov) v obliki žage, kot je prikazano na sliki 4. Po mešanju radarskega odmeva z oddajnim signalom, se pridobi signal frekvenčne razlike (angl. beat signal), čigar frekvenca je odvisna od oddaljenosti objekta.



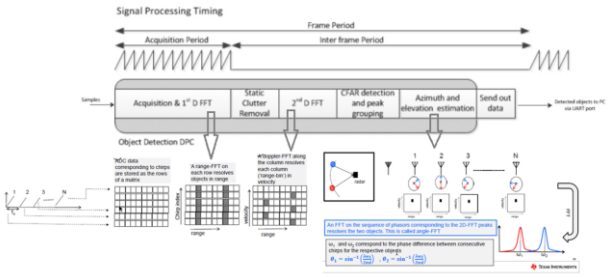
Slika 4. Časovni potek signala, ki ga oddaja radar

Z nastavitvijo parametrov žage, števila in lastnosti preletov v okvirju radar nastavljamo za specifične načine uporabe, ki zahtevajo različni maksimalni domet radarja, maksimalno hitrost objekta, prilagojen kot vidnega polja, ter razločljivost vzdolž in prečno na trenutno os radarskega snopa. Postopek nastavljanja parametrov je vedno kompromis in je opisan v [6].

Procesiranje radarskega signala poteka v več fazah, kot je prikazano na sliki 5. Po zajemu in vzorčenju se zaporedoma v vrstice shranjujejo signali frekvenčnih razlik (angl. beat signal), oziroma mešalni produkt odbojev od tarč in oddajanega signala. S hitro Fourijerjevo transformacijo (angl. fast Fourier transform – FFT) za vsako vrstico dobimo frekvenčni spekter mešalnega signala (angl. beat signal), ki pokaže odmeve objektov na skali, katera predstavlja maksimalni domet radarja. Število vrstic je enako številu preletov v radarskem okvirju. Če so objekti nepremični (statični), se spekter ne spreminja. Statično okolico radarja je mogoče tudi odšteti (angl. static clutter removal) in iskati zgolj spremembe med zaporednimi signali odmevov. S hitro Furijerjevo transformacijo nad vrsticami dobimo hitrost detektiranih objektov na isti skali maksimalnega dosega radarja.

Za tem sledi faza določanja praga detekcije (angl. Constant False Alarm Rate – CFAR), ter grupiranja objektov. Prag CFAR je funkcija tako verjetnosti zaznave kot tudi verjetnosti napačnega alarma.

Po shranjevanju zaporedja okvirja, dobimo radarsko spominsko kocko (angl. radar cube), iz katere se lahko določi smer odboja, ter z bolj zahtevno matematično obdelavo vzorcev v radarski kocki, lahko naredimo specifične aplikacije.



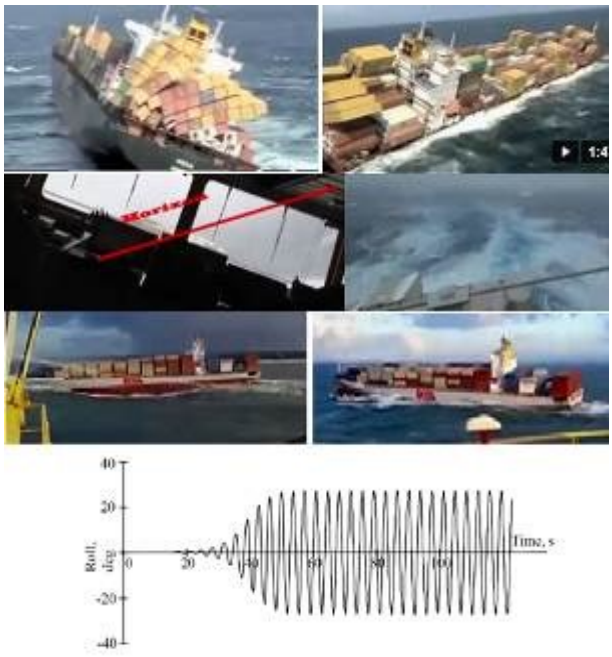
Slika 5. Procesiranje radarskega signala

3 Pojav parametrskega in ekstremnega nihanja ladij

Pojav parametrskega nihanje ladje, ki ga prikazuje slika 6, je več kot pol stoletja znan naravni pojav, ko ladja, ki ima določeno hitrost, zaradi energije valov iz smeri premca (sprednjega dela ladje) ali krme (zadnjega dela ladje), z zadostno višino in valovno dolžino od 0,8 do 1,2 dolžine ladij s periodo primerljivo dvakratni periodi nihanja ladje, začne nihati z zelo velikimi odkloni. Pri ladjah z ravnim trupom, kot so potniške ali kontejnerske ladje, zaradi izgube izpodrivnega volumna morja v zgoraj opisanih pogojih pride do delnega zmanjšanja potiska vode ter se energija vzbujanja naprej/nazaj (angl. pitch) parametrsko prenese oziroma sinhronizira v energijo nihanja levo/desno. Pojav se razvije v manj kot minuti časa, in če se ne spremeni vzbujanje, npr. ladja spremeni hitrost in/ali smer, lahko pride do ekstremnih odklonov, tudi za več kot 45°.

Preseganje maksimalnega nagiba opreme, v primeru srednje napetostnih generatorjev proizvajalca ABB je ta omejitev ±25° lahko povzroči izpad pogona (angl. blackout), izgubo tovora, ki potem predstavlja navigacijsko nevarnost, poškodbe posadke, poškodbe potnikov na potniških ladjah, in v primeru ekstremnega nihanja tudi obračanje ter potopitev ladje.

Ker je v pomorski skupnosti pojav premalo znan, so na žalost tovrstne nesreče pogoste. Leta 2010 M/S Brilliance of the Seas ob obali Aleksandrije doživi ekstremno nihanje. Pojav se razvije v manj kot minuti po spremembi smeri ladje, traja dve in pol minuti z nihanjem od 35° do 40°. Nihanje levo desno je bilo pospremljeno z zelo močnim nihanjem naprej/nazaj, kar kaže na pojav parametrskega vzbujanja. Pred pojavom ekstremnega nihanja, se je ladja nihala do 5°. Poškodovano je bilo 138 potnikov in 7 članov posadke, materialna škoda je znašala 1,7 milijona dolarjev.



Slika 6. Pojav parametrskega nihanja ladje

Sile, ki delujejo na trup ladje, so sestavljene iz sil propulzorjev, krmila in aktivnih plavuti (angl. stabilizers), sil vetra in valov ter reakcijskih sil zaradi ladijskih premikov.

Matematična razlaga dinamičnega obnašanje ladje v težkih okoliških razmerah je podana v [7]. Za majhne kote nihanja ladje velja linearna obravnava, ob večjih kotih nihanja pride do nelinearnih pojavov, kjer lahko reakcijskih sile obrnejo ladjo.

Detekcija nastanka pojava parametrskega nihanja temelji na analizi dinamike gibanja ladje, kjer se uporabljajo vgrajeni dvo ali tri osni merilniki pospeška (angl. accelerometers).

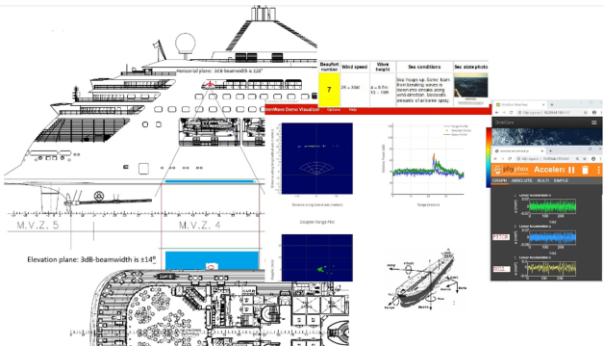
V komercialnem produktu PAROLL [8] se poleg detektorja sinhronizacije frekvence nihanja naprej/nazaj in levo/desno, uporablja detektor sinhronizacije faze obeh nihanj, ker se je izkazalo, da pojav parametrskega nihanja predhodno pospremi niz 5 do 6 močnih nihanj naprej/nazaj.

4 Detekcija parametrskega nihanja z radarjem

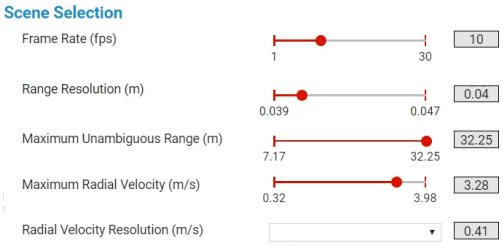
Za oceno stanja morja pred ladjo se lahko uporabljajo odboji od valovanja morja (angl. sea clutter), ki se običajno izvedejo s pomočjo navigacijskega radarja v frekvenčnem področju X. Radar podaja informacije, ki vsebujejo podatek o višini valov, dolžini valov, periodi valovanja in površinskih tokovih.

Izboljšavo detekcije parametrskega nihanja dobimo z opazovanjem, oziroma meritvijo oblike valov vzdolž trupa ladje.

Za namen testiranja uporabe radarja AWR 1843, ki deluje v milimetrskem frekvenčnem področju, je bila opravljena meritev valov na potniški ladji M/S Silverea Muse. Os radarja je bila usmerjena navpično v morsko površino z osmega nadstropja ladje, kot prikazuje slika 7. Za potrebe verifikacije se je dinamiko gibanja ladje spremljalo z uporabo triosnega merilnika pospeška na mobilnem telefonu in programa Phyphox [9] in še snemalo video valovanja. Na sliki 8 so prikazane 2D (2Tx/4Rx) nastavitve radarja za opazovanje valov z osmega nadstropja ladje. Na sliki 9 so prikazani rezultati meritev pet metrskih visokih in dolgih valov iz smeri premca (ang. sea swell). Iz grafa x,y,z pospeška se vidi nihanje naprej/nazaj, graf x (angl. pitch) predaja energijo v nihanje levo/desno, graf z (angl. roll).

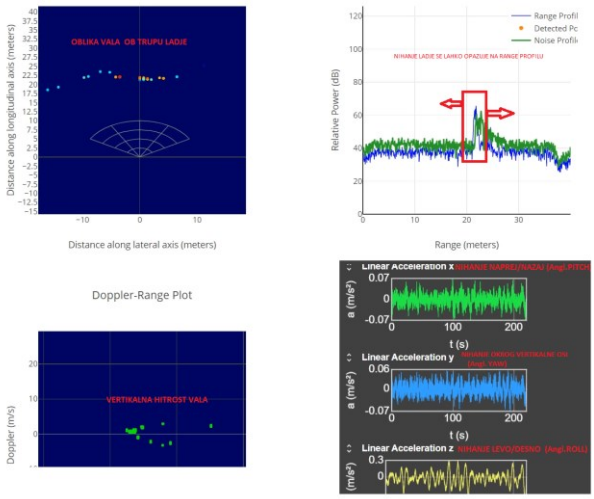


Slika 7. Prikaz namestitve radarja na ladjo



Slika 8. 2D (2Tx/4Rx) nastavitve radarja

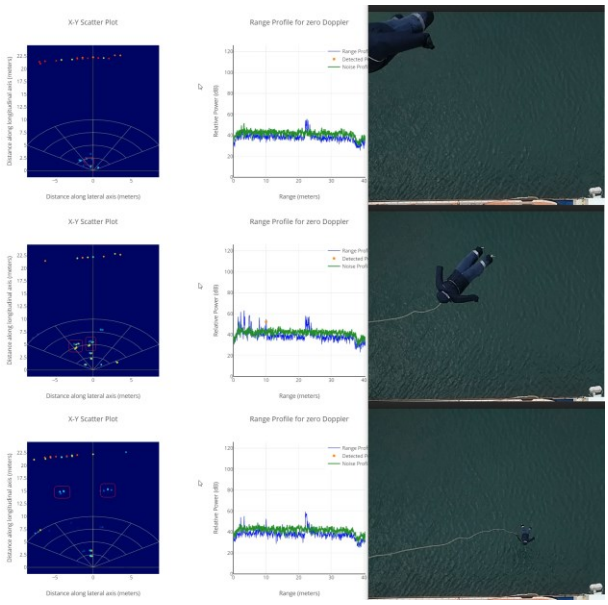
Iz radarskega odmeva od površine morja dobimo obliko, višino in periodo valov (slika 9). Z izvedbo eno-dimenzijskega (1D) FFT se pridobi graf profila področja (angl. range profile graph), kar omogoča meritev velikosti in oblike valov vzdolž trupa ladje in nihanja ladje levo desno. Z izvedbo dvo-dimenzijskega (2D) FFT se pridobi informacija o hitrosti objekta (angl. doppler range profile graph), kar omogoča spremljanje dinamike morskega valovanja.



Slika 9. Meritev 5 m visokih dolgih valov iz smeri premca

Poleg omenjene meritve oblike valov ob ladji lahko radar zazna tudi padec v morje, ki je pogost pojav na potniških ladjah. Na sliki 10 je pokazan padec lutke v morje in detekcija padca z radarjem. Po nekaj metrih postane opazen pojav dvojne tarče, ki nastane zaradi neumerjenega radarja.

Z uporabo drugega radarja postavljenega nižje v horizontalni smeri in s tri-dimenzijonalnim snemanjem, lahko verjetnost detekcije padca v morje občutno povečamo. Dodatni drugi radar lahko uporabljamo tudi kot pomoč pri privezu ladje, kot alternativa ali dopolnitev laserskemu sistemu. [10] S pravilnim načrtovanjem obdelave podatkov v radarski kocki, lahko še dodatno izboljšamo verjetnost detekcije in zmanjšamo verjetnost napačnih detekcij (angl. false true), kar bo predmet nadaljnjega dela.



Slika 10. Meritev padca v morje (angl. man over board)

5 Zaključek

V članku je opisano testiranje radarja na frekvenčnem področju milimetrskih valovih v pomorstvu. Prikazan je primer uporabe detekcije parametrskega in ekstremnega nihanja ladij vključno s detekcijo padca potnika v morje. Uporaba cenovno dostopnega radarja se je izkazala kot zelo primerna in fleksibilna rešitev, ki ponuja možnosti prilagajanja in reševanja nekaterih problemov v pomorstvu.

Literatura

[1] Yury Yu. Yurovsky, Vladimir N. Kudryavtsev, Semyon A. Grodsky, Bertrand Chapron, »Sea Surface Ka-Band Doppler Measurements:Analysis and Model Development«, Remote Sensing 11(7):839, april 2019.

[2] S. Clark and H. Durrant-Whyte, "Autonomous land vehicle navigation using millimeter wave radar," Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146), Leuven, Belgium, 1998, pp. 3697-3702 vol.4, doi: 10.1109/ROBOT.1998.681411.

[3] Shao-Qiu Xiao, Ming-Tuo Zhou, Yan Zhang, »Millimeter wave technology in wireless Pan, Lan, and Man«, CRC Press, 2008.

[4] T. Wen, Z. G. Yao, Z. L. Zhao, L. F. Lin, Z. G. Han and L. D. Guo, "Retrieval of Sea Surface Wind Speed Using Spaceborne Millimeter-Wave Radar Measurements," in IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 12, pp. 1807-1811, Dec. 2018, doi: 10.1109/LGRS.2018.2865196.

[5] Mladen Radovanović, Boštjan Batagelj, »Single-chip mmW Radar for Marine and Industrial Applications«, seminar radijskih komunikacije 2020 (SRK 2020), Ljubljana, februar 2020.

[6] Programming Chirp Parameters in TI Radar Devices, Texas Instruments Application Report, maj 2017, dostopno na: https://www.ti.com/lit/an/swra553a/swra553a.pdf

[7] R. A. Ibrahim, I. M. Grace, »Modeling of Ship Roll Dynamics and Its Coupling with Heave and Pitch«, Mathematical Problems in Engineering, Hindawi Publishing Corporation, Article ID 934714, 2010.

[8] Roberto Galeazzi, »Real-time Detection of Parametric Roll: Full-scale Validation«, CeSOS Highlights and AMOSVisions Conference, Trondheim, 27-29 May 2013.

[9] phyphox.org

[10] M. Perkovič, L. Gucma, M. Bilewski, B. Muczynski, F. Dimc, B. Luin, P. Vidmar, V. Lorenčič, M. Batista, »Laser-Based Aid Systems for Berthing and Docking«. J. Mar. Sci. Eng. 2020, 8, 346.

Spremljanje satelitov GPS v vidnem polju na osnovi modula NEO-6M

Blaž Bertalanič, Aljaž Blatnik, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija

E-pošta: blaz.bertalanic@outlook.com

Real time tracking of GPS satellites in view with module NEO-6M

Abstract. *With the help of module NEO-6M we created an application with which we are able to track GPS satellites in view in real-time. The application is able to interpret the results that it gets from the satellite and show them to a user in a friendly manner. With it we can observe the signal-to-noise ratio of the signals received from the satellites in view and also display its approximate direction of position in the sky through azimuth and elevation. We were able to make a product that can be a sufficient tool for students to work with at laboratory work in the school process.*

1 Uvod

Radijski sistem globalne navigacije (angl. Global positioning system – GPS), temelji na osnovi serije satelitov, ki krožijo okoli Zemlje. Ta sistem nam omogoča določanje naše geolokacije, ki je določitev lokacije podane glede na Zemljo v danem koordinatnem sistemu [1] v primeru, da imamo v vidnem polju štiri ali več satelitov. Večina komercialnih naprav nam podaja le informacije o naši lokaciji v obliki zemljepisne širine in dolžine, ali pa nam na zemljevidu prikažejo točko, kjer se trenutno nahajamo. Drugi podatki so skriti, oziroma nimajo večjega pomena za uporabnika. Seveda je eno od vprašanj, ki se nam porodi, kako močan signal sploh potrebujemo, da lahko uspešno določimo našo lokacijo, oziroma pri kakšni moči signala lahko rečemo, da je nek satelit v našem vidnem polju.

GPS sicer spada v skupino GNSS (angl. Global Navigation Satellite System), ki predstavlja skupino satelitov, ki krožijo okoli Zemlje in so primarno namenjeni določanju geolokacije. GPS je bil narejen s strani ameriškega obrambnega ministrstva za potrebe varnosti . Projekt se je začel že daljnega leta 1973, prvi prototip pa so izstrelili v vesolje 5 let kasneje. Do leta 1993 je ameriška vlada že imela konstalacijo 24 satelitov [2]. Do danes se je ta številka povečala na 31 delujočih satelitov [3]. Glavna značilnost GNSS sistemov je, da imajo vsi sateliti na krovu atomsko uro, ki predstavlja časovno referenco, potrebno za delovanje sistema, hkrati pa je komunikacija do končnega uporabnika enosmerna. Vsak satelit posreduje svojo natančno časovno referenco in navigacijsko sporočilo s svojo efemeridо, s pomočjo katerih nato sprejemnik preračuna svojo geolokacijo, vendar za to potrebuje vsaj 4 satelite v vidnem polju, ki so potrebni za triangulacijo položaja [4]. Sateliti GPS se gibljejo na približni višini 20200 kilometrov od Zemljine površine, ki jo v svoji tirnici preletijo dvakrat na dan [6]. Tirnice satelitov so izbrane tako, da je na večini točk na Zemlji vedno v vidnem polju sprejemnika vsaj 6 satelitov [7].

Končni cilj aplikacije je bilo pripraviti merilno orodje za uporabo na laboratorijskih vajah, kjer bi lahko študenti v realnem času opazovali kaj se dogaja z močjo sprejetih signalov satelitov v vidnem polju, ko na sprejemnik vplivamo z različnimi motilniki, ki bodo v signal vnašali motnjo, oziroma koliko močno motnjo moramo proizvesti, da sprejemnik ni več zmožen zaznave signala iz satelitov. Aplikacija mora temeljiti na odprtokodnih rešitvah in ni odvisna od omrežne povezave, oziroma se mora izvajati lokalno. Pri tem mora biti uporabniški vmesnik čim bolj preprost za uporabo, hkrati pa morajo biti podatki prikazani na razumljiv in nedvoumen način.

Na trgu se nahajajo tudi konkurenčni produkti izdelani aplikaciji, ki pa so večinoma neprimerni za uporabo v didaktične namene. Aplikacija proizvajalca naprave NEO-6M, z imenom u-Center [8], omogoča spremljanje vseh podatkov o satelitih, toda te prikazuje v neprijaznem grafičnem vmesniku, težave pa ima tudi s stabilnostjo saj po nekaj minutah uporabe postane neodzivna. NEO-6M se zaradi cenovne dostopnosti uporablja tudi v številnih domačih projektih, vendar so vsi namenjeni predvsem spremljanju zemljepisne dolžine in širine ter posledično določanju geolokacije, ne pa tudi za prikaz in interpretacijo satelitov v vidnem polju.

Naš merilni sistem je sestavljen iz dveh delov in sicer strojnega dela, ki ga sestavlja modul proizvajalca uBlox NEO-6M in služi kot sprejemnik GPS signala. Modul posreduje želeno informacijo v obliki stavka NMEA (angl. National Marine Electronics Association) na računalnik, kjer ta stavek obdelamo in primerno prikažemo končnemu uporabniku. Koncept sistema je prikazan na sliki 1.

V nadaljevanj je opisana uporabljena strojna oprema za končno aplikacijo, ter programski rešitvi interpretacije pridobljenih podatkov iz strojne opreme in njihovem prikazu.



Slika 1: Koncept našega merilnega sistema

2 Strojna oprema aplikacije

Strojna oprema je sestavljena iz zunanje in notranje enote. Notranjo enoto sestavlja preprost čip FT232R proizvajalca Future Technology Devices International (krajše FTDI), ki deluje kot posrednik za komunikacijo med vodilom USB in serijsko komunikacijo med zunanjo enoto in računalnikom. V zunanji enoti najdemo izdelek proizvajalca uBlox, katerega srce predstavlja modul GPS NEO-6M (Slika 2).

2.1 NEO-6M

To je izjemno zmogljiv sprejemni modul, saj je sposoben na enkrat spremljati kar 22 satelitov na 50 različnih frekvenčnih kanalih nosilca L1 (1575,42 MHz), ki so v vidnem polju. Naprava je zmožna zaznati signal jakosti zgolj -161 dBm in pri tem porabi samo 45 mA toka, v primeru varčevanja z energijo, ko sprejemnik selektivno vklaplja in izklaplja svoje komponente, pa je poraba toka v območju 11 mA in je zato primeren za baterijske naprave. NEO je s predpisano anteno sposoben do petkrat na sekundo posodobiti lokacijo z do 2,5 m horizontalno točnostjo. Zraven je potrebno poudariti tudi, da je pri vročen zagonu čas do prvega izračuna pozicije (angl. Time-To-First-Fix) manjši od 1 sekunde, se pravi potrebuje sprejemnik manj kot 1 sekundo od vklopa do prvih natančno izračunanih in oddanih podatkov. Na modulu sta prisotna tudi električno izbrisljiva programirljiva bralna pomnilnika (angl Electrically Erasable Programmable Read-Only Memory - EEPROM) in baterija, ki skupaj poskrbita, da se ob izklopu na modulu ohranijo nekateri podatki o pozicijah satelitov (GNSS podatki) in druge nastavitve modula, ki so pomembne za hitrejši zagon lokalizacije. V teoriji se hitrost do prve uklenitve s satelitom izračuna po enačbi (**Napaka! Vira sklicevanja ni bilo mogoče najti.**), kjer t_0 predstavlja čas potreben za vsake faze v zaporedju kodnega multipleksa, krajše C/A (angl. »Code division multiplex«) in znaša približno 10 ms. Črka N predstavlja število vseh satelitov, ki v GPS konstalaciji znaša trenutno 31, N' pa predstavlja število satelitov, ki jih lahko trenutno vidimo na nebu, kar v povprečju znaša 8 satelitov. Črka M število poskusov uklenitve na posamezni frekvenci, ki pri naši napravi znaša 10 poskusov, medtem ko oznaka K število frekvenčnih kanalov na katerih lahko naša naprava zazna satelite. Ko vnesemo vse podatke v enačbo lahko izračunamo, da bi se moral modul NEO-6M sinhronizirati v dobrih 8 sekundah. Zaradi že omenjenega shranjevanja določenih podatkov s pomočjo baterije, ki te podatke drži v spominu, se naprava lahko sinhronizira bistveno hitreje.

$$t = t_0 \cdot 1023 \cdot \frac{N}{N'} \cdot M \cdot \frac{1}{K} \quad (1)$$



Slika 2: GPS modul NEO 6M z anteno

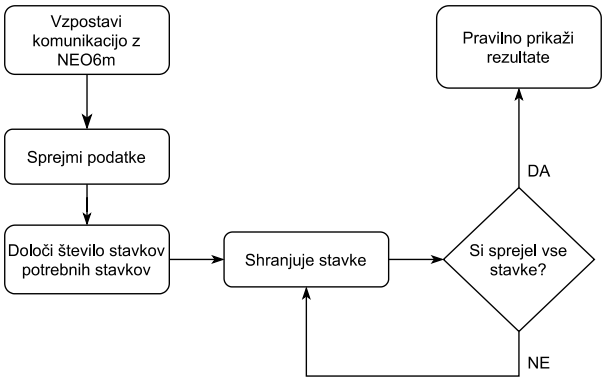
2.2 Antena

Na čip je preko priključka U.FL priklopljena krpičasta antena z oznako 1575R-A. Antena sprejema satelitski signal v območju med 1561 MHz in 1575,42 MHz, ima dobitek 26 dBi in desno sučno polarizacijo. Polarizacija je potrebna predvsem za odpravo neželenih odbojev. Sateliti namreč svoj signal oddajajo z desno sučno polarizacijo, vendar vsi ne pridejo neposredno do sprejemnika ampak se lahko pred sprejemom odbijejo od tal in bi tako lahko vnašali pogrešek pri določanju lokacije. Od tal odbitim signalom se zaradi odbojnosti spremeni polarizacija v levo sučno in na tak način lahko sprejemnik loči neposredni signal od odbitega.

3 Programski del in prikaz rezultatov

Programski del aplikacije je bil izdelan v programskem jeziku Python [9]. Celotna komunikacija s čipom NEO poteka preko serijske povezave s pomočjo knjižnice PySerial [10], ki pošilja stavke NMEA, medtem ko je za grafični vmesnik uporabljena knjižnica Tkinter [11]. Tabela rezultatov je narejena s pomočjo knjižnice Tkinter-table [12], vsi grafi pa so narejeni s pomočjo Matplotlib [13].

Program deluje tako, da najprej vzpostavi komunikacijo s čipom NEO in preveri število stavkov, ki jih lahko pričakuje in si jih shranjuje toliko časa, da si zabeleži vse pričakovane. Iz tako sprejetih podatkov nato izlušči tiste vrednosti, ki nas zanimajo (identifikacijsko številko satelita, razmerje signal šum, azimut in elevacijo). V primeru razmerja signal šum podatke tudi uredi po velikosti od najvišjega proti najnižjemu in nato vse izluščene podatke prikaže v uporabniškem vmesniku. Diagram poteka programa je predstavljen na sliki 3.



Slika 3: Diagram poteka prikaza rezultatov

3.1 Stavki NMEA in obdelava podatkov

Stavki NMEA so bili definirani s strani National Marine Electronics Association (NMEA) in so bili prvotno namenjani za komunikacijo med različno pomorsko elektronsko opremo. Vsak stavek se začne z znakom »$« in konča z »*« in rezultatom zgoščevalne funkcije (angl. hash function). V standardu je definiranih 29 različnih stavkov NMEA za različne uporabe, od tega jih je 10 namenjenih za GPS.

Za našo aplikacijo smo uporabili stavek GSV (»Satellites in view«), ki prikazuje podatke o satelitih v vidnem polju. Tipična oblika sprejetega podatka je »$GPGSV,4,1,16,01,40,083,46,02,17,308,41,12,07,344, 39,14,22,228,45*75«, kjer so z vejico ločeni posamezni podatki. Prvi podatek predstavlja tip stavka, v našem primeru GSV, drugi podatek predstavlja število stavkov, ki jih potrebujemo za sprejem vseh podatkov. Tretji podatek nam pove kateri stavek je po vrsti, od števila ki predstavljajo vse podatke, četrti pa nam pove koliko satelitov se nahaja v vidnem polju. Nato sledijo podatki o samih satelitih in sicer PRN, kar predstavlja psevdo naključno kodo (angl. pseudo-random noise sequence), ki se uporabi v kodiranju signala. Vsak satelit ima svoj PRN ,ki je tudi unikatna identifikacija satelita. [9] Podatki, ki sledijo PRN so azimut glede na pravi sever, tudi geografcki sever [8], elevacija, ki je prav tako normirana na pravi sever in razmerje signal šum (angl. Signal-to-noise ratio – SNR). Vsak stavek lahko vsebuje te podatke za maksimalno 4 satelite, zato za večje število sprejetih satelitov potrebujemo več stavkov NMEA, kar nam tudi pove drugi podatek. Potrebno je omeniti, da sta podatka, ki jih čip poda o azimutu in elevaciji vezana na položaj uporabnika, torej je v središču, na točki (0,0) v polarnih koordinatah, sprejemnik. To pomeni, da če bi se obrnili v smeri geografskega severa, ki se nahaja na azimutu 0°, lahko v grobem ocenili smer v kateri se nahaja posamezni opazovani satelit.

Zaradi načina zapisa podatkov o satelitih, ki se izpisujejo v zaporednih stavkih NMEA, je bil glavni izziv kako zajeti vse in jih potem na koncu uporabiti kot celoto. To nam je uspelo narediti s pomočjo stalne zbirke, v katero smo zapisovali podatke o satelitih. Te podatke smo posodabljali z vsakim naslednjim prispelim stavkom, dokler nismo zapisali vseh prispelih podatkov, oziroma vseh potrebnih stavkov. Število teh podajajo stavki NMEA za zajem vseh satelitov v vidnem polju. Aplikacija po končani zapolnitvi zbirke razvrsti satelite po jakosti signala, od najnižjega do najvišjega in posodobi prikaz na uporabniškem vmesniku. Nato izprazni zbirko in naredi prostor za vnos novih podatkov iz NMEA stavkov, celotni cikel pa se ponavlja.

3.2 Tabela zaznanih satelitov

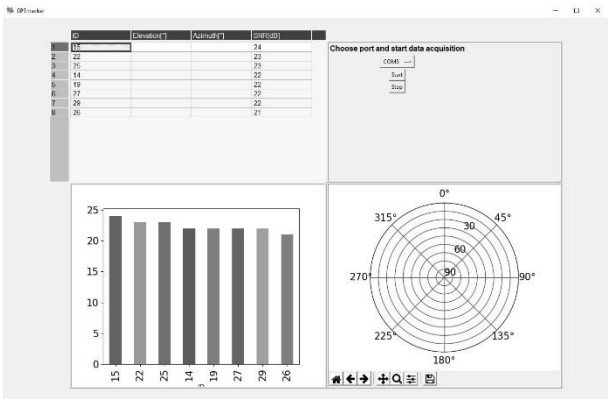
V zgornjem levem kotu uporabniškega vmesnika se nahaja tabela s podatki za posamezne satelite v vidnem polju. V prvem stolpcu je prikazana PRN oznaka satelita, ki je za vsakega unikatna, nato njegova elevacija, ki se lahko giblje med 0° in 90°, sledi azimut, ki se giblje med 0°in 360° glede na pravi sever in na koncu še SNR, ki se lahko giblje med 0 in 99 decibeli. Vsi vnosi so urejeni po velikosti SNR in sicer so na vrhu tisti sateliti z največjim SNR, medtem ko so čisto spodaj tisti z najnižjim razmerjem.

3.3 Graf razmerja moči zaznanih satelitov

V levem spodnjem kotu aplikacije se nahaja vizualni prikaz razmerja moči sprejetih satelitov v obliki stolpčnega grafa. Podobno kot v tabeli rezultatov so tudi tukaj posamezni sateliti predstavljeni s svojim PNR, nato pa višina stolpca predstavlja velikost razmerja med signalom in šumom sprejetega satelita. Podatki so urejeni tako, da se prikazana moč niža iz leve proti desni. Osveževanje prikaza poteka vsakih 100 ms, kar je nekoliko hitreje kot poteka osveževanje podatkov iz modula, ki je v območju 200 ms.

3.4 Prikaz položaja satelitov

Desno spodaj se nahaja grafika, kjer so v polarnih koordinatah prikazani položaji posameznih satelitov glede na njihovo elevacijo in azimut glede na geografski sever (Slika 2). Elevacija je predstavljena z radijem, kjer koordinata (0,0) predstavlja 90° elevacijo, medtem ko azimut poteka v smeri urinega kazalca od 0° do 360°. Prikazi se osvežujejo na 200 ms, ker podatki o položaju satelita zelo malo variirajo in zato je tolikšno osveževanje potrebno predvsem zaradi lepše animacije premikanja.



Slika 4: Uporabniški vmesnik aplikacije

4 Diskusija in možne nadgradnje

Na koncu nam je uspelo izdelati program, ki je preprost za uporabo in z razumljivim prikazom podatkov, ki jih pridobimo v obliki NMEA stavkov iz satelitov. Celotna aplikacija deluje brez tehničnih težav in je sposobna podatke osveževati vsakih 100 ms, medtem ko se podatki o približni smeri položaja satelitov osvežujejo z 200 ms, ker ti bolj kot ne ostajajo konstantni. Zaradi tega je aplikacija primerna za prikaz še tako majhnih sprememb pri meritvah razmerja signal šum, kar je tudi eden od pogojev za uporabo na laboratorijskih vajah. V programski kodi so predvideni nastavki za morebitno nadgradnjo aplikacije, s katerimi lahko bistveno olajšamo delo in pospešimo nadaljnji razvoj.

Aplikacijo bi bilo mogoče tudi nadgraditi, da bi lahko obdelala in prikazovala tudi druge stavke NMEA. V prvi vrsti bi lahko na interaktivnem zemljevidu prikazovala trenutno lokacijo sprejemnika, kot tudi projekcijo položaja satelitov v vidnem polju na azimutnem zemljevidu, vendar bi za to potrebovala internetno povezavo, kar pa ni bilo zaželeno v trenutni izvedbi. Eden od načrtov je tudi, da bi si aplikacija beležila vse posredovane podatke o jakosti signala skozi čas in nato ob koncu meritev prikazala graf celotnega potek upadanja moči v odvisnosti od časa in tako omogočila dodaten podatek za kasnejšo analizo.

5 Zaključek

V članku je predstavljena rešitev za pridobivanje posameznih podatkov o satelitih GPS v vidnem polju in njihov prikaz na grafičnem vmesniku. V realnem času je izvedena meritev razmerja signal-šum satelitov v vidnem polju. Celotna aplikacija temelji na strojni opremi, kjer je uporabljena nizkocenovna naprava NEO-6M, ki omogoča zajem satelitskih signalov s frekvenco 5 Hz. Uporabnik tako lahko pridobi podatke o razmerju signal šum sprejetega signala posameznega satelita in njegovo približno smer lokacije na nebu v obliki azimuta ter elevacije. Podatki, ki se zajemajo s pomočjo navedene naprave se posredujejo v obliki NMEA stavkov v Python skripto, ki te stavke obdela, iz njih izlušči pomembne podatke in jih pretvori v primerno obliko za urejanje ter končni prikaz. Grafični vmesnik, ki je razumljiv in preprost za uporabo, prikazuje podatke na tri različne načine, ki skupaj ponujajo najboljšo reprezentacijo za končnega uporabnika. Uspelo nam je narediti produkt, ki predstavlja primerno orodje za študente pri izvedbi laboratorijskih vaj.

Literatura

[1] [1] TRIGLAV, Joc, 2012, Analiza pomena geolokacije kot funkcije časa : doktorska disertacija [na spletu]. Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo . [Dostopano 20 marec 2020]

[2] https://en.wikipedia.org/wiki/Global_Positioning_System(prvi dostop 12.3.2020)

[3] US Government. "GPS.gov". GPS.gov. US Govt. Pridobljeno 21 marca 2015.

[4] McNeff, J. (2002), “The Global Positioning System”, IEEE Transactions on Microwave theory and techniques, VOL 50, No.3, Marec 2002

[5] https://en.wikipedia.org/wiki/Global_Positioning_System(prvi dostop 12.3.2020)

[6] Agnew, D.C.; Larson, K.M. (2007). "Finding the repeat times of the GPS constellation". GPS Solutions. 11 (1): 71–76.

[7] https://en.wikipedia.org/wiki/List_of_GPS_satellites (prvi dostop 12.3.2020)

[8] https://www.u-blox.com/en/product/u-center (prvi dostop 12.3.2020)

[9] https://www.python.org/ (prvi dostop 10.3.2020)

[10] https://pypi.org/project/pyserial/ (prvi dostop 10.3.2020)

[11] https://docs.python.org/3/library/tkinter.html (prvi dostop 10.3.2020)

[12] https://github.com/dmnfarrell/tkintertable (prvi dostop 10.3.2020)

[13] https://matplotlib.org/ (prvi dostop 10.3.2020)

[14] Massatt, Paul; Wayne Brady (Summer 2002). "Optimizing performance through constellation management". Crosslink: 17–21. Archived from the original on January 25, 2012.

[15] *McClure, Bruce (2013). "Polaris is the North Star". Astronomy Essentials. EarthSky. Archived from the original on 2014-03-09. Retrieved 2014-04-28*

Komunikacijsko satelitsko omrežje Starlink

Žiga Andrejc, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Laboratorij za sevanje in optiko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: ziga.andrejc@gmail.com

Starlink Satellite Communications Network

Abstract. *This paper introduces the Starlink-satellite broadband-communications network, developed by the American company SpaceX. The main technical characteristics of this special segment of broadband-communications networking based on the constellation network of small-satellite technology are reviewed and commented on. The main motivation of the paper is to clearly show the technical characteristics of the Starlink network that SpaceX does not disclose in popular-science presentations.*

Keywords: Starlink, satellite network, low Earth orbit, telecommunications network, broadband access

1 Uvod

Izgradnja in poskusi poslovne uporabe svetovnega širokopasovnega omrežja elektronskih komunikacij preko sistema umetnih satelitov so se začeli že pred več kot desetimi leti [1]. Tovrstne satelitske storitve, ki so alternativa zemeljskim omrežjem iz optičnih vlaken ali brezžičnim povezavam, pa niso tuje niti slovenskemu okolju za zagotavljanje širokopasovnih povezav na podeželju [2], kjer se je v Posočju začela zgodba odvijati že 2009 [3]. Za razliko od običajnih zemeljskih komunikacijskih omrežij, pri katerih internetni promet potuje po optičnih kablih ali krajši razdalji po zraku, satelitsko internetno omrežje prenaša informacije s pomočjo mikrovalovnega valovanja, ki potuje med satelitom v zemeljski tirnici in uporabnikom nekje na Zemlji. Ne glede na trenutno razširjenost zemeljskih fiksnih in mobilnih omrežij ter na nekatere omejitve današnjih satelitskih omrežij pa pri vzpostavitvi resnično globalnega internetnega omrežja, ki bo lahko ponujalo storitve uporabnikom ne glede na lokacijo, brez satelitske tehnologije ne bo šlo. Nov globalen internet, ki bo zagotovljen brez lokalnih telekomunikacijskih operaterjev, bo odporen na lokalno cenzuriranje in omejevanje storitev, bo pa na žalost vodil v globalni monopol in ogrozil nevtralnost interneta [4].

V članku je predstavljeno satelitsko širokopasovno komunikacijsko omrežje Starlink, ki ga razvija ameriško podjetje SpaceX. Pregledane in komentirane so glavne tehnične značilnosti tega zelo posebnega segmenta širokopasovnih komunikacijskih omrežij, osnovanih na konstelacijskem omrežju tehnologije majhnih satelitov. Glavna motivacija prispevka je jasno prikazati tehnične lastnosti omrežja Starlink, ki jih podjetje SpaceX običajno ne razkriva podrobno v javno dostopnih predstavitvah.

2 Satelitska internetna omrežja

Trenutno satelitsko internetno omrežje zagotavljajo veliki in zmogljivi sateliti, ki so umeščeni v geostacionarni tirnici (angl. Geostationary orbit – GEO) na višini 35.786 km nad zemeljskim ekvatorjem in pokrivajo določeno področje na Zemlji [5]. Slabost geostacionarne tirnice je zakasnitev pri oddajanju ali sprejemanju signala zaradi oddaljenosti satelita, kot prikazuje Tabela 1. Ker geostacionarni sateliti nimajo medsatelitskih zvez, mora signal običajno prepotovati štirikratno razdaljo med Zemljo in geostacionarno tirnico [1]. Z namenom zmanjšanja zakasnitev so se že pred več kot 20 leti začele vzpostavljati satelitske konstelacije v nizki zemeljski tirnici (angl. low Earth orbit – LEO) in srednji zemeljski tirnici (angl. medium Earth orbit – MEO) z namenom zagotavljanja mobilnih in fiksnih storitev (Iridium, Globalstar, Skybridge,..) ter sateliti O3b podjetja SES S.A. v MEO.

Tabela 1: Lastnosti tirnice za satelitsko internetno omrežje

vrsta tirnice	LEO	MEO	GEO
oddaljenost	100 km–2.000 km	2.000 km–36.000 km	36.000 km
zakasnitev	10 ms	120 ms	260 ms
hitrost satelita	7,8–6,9 km/s	6,9–3,1 km/s	>3,1 km/s
vidljivost satelita	kratka	srednja	vedno
predaja zveze	pogosta	srednja	ni
območje pokrivanja	majhno	srednje	veliko
življenjska doba satelita	3–7 let	10–15 let	10–15 let

Poleg nizkih zakasnitev imajo sateliti v LEO tudi manjšo pokritost, kar poveča zmogljivost uporabniških povezav v območju pokrivanja, vendar jih za globalno pokritost s signalom rabimo več. Zaradi ekonomije obsega in kratke življenjske dobe se v nizko tirnico umeščajo majhni sateliti, kjer svojo vlogo v zadnjih letih odigrava tudi Slovenija [6].

Osnovni problemi, ki se pojavljajo pri gradnji širokopasovnega satelitskega internetnega omrežja v LEO [1], so:

- Komunikacijski protokoli morajo omogočati mobilnost tako mrežnih vozlišč kot terminalov.
- Vzdrževanje zvez na relacijah satelit-satelit in satelit-Zemlja ter obratno ni trivialno, ker se topologija omrežja spreminja.
- Hitra predaja zvez zaradi velikih hitrosti gibanja satelitov.
- Komunikacijski protokoli morajo imeti toleranco za relativno visoke zakasnitve in slabljenja signalov pri

komunikaciji satelitov z zemeljskimi postajami ali terminali.
- V novi arhitekturi omrežja bodo podatke procesirali in usmerjali tudi sateliti.

Izmed aktualnih projektov majhnih satelitov na LEO je medijsko najbolj odmeven Starlink. Projekt je prvič ugledal luč januarja leta 2015, ko je svetu svojo idejo predstavil direktor in lastnik podjetja Elon Musk. Ob predstavitvi je povedal, da je njihov cilj prenos večine internetnega prometa na dolgih razdaljah in okrog 10 % lokalnega internetnega prometa nekega območja. Prav tako pa je izpostavil, da imajo satelitska internetna omrežja prednost pred zemeljskimi, saj je hitrost elektromagnetnega valovanja v vakumu za približno 30 % višja kot v optičnem vlaknu in ker je potrebnih manj vozlišč, da informacija prispe na oddaljen cilj [7]. Projekt je požel mnogo odobravanja kot tudi nasprotovanja v svetovni in slovenski astronomski skupnosti profesionalnih in amaterskih astronomov [8, 9].

3 Glavne značilnosti

SpaceX bi rad s Starlinkom omogočil hiter in cenovno ugoden prenos velikih količin podatkov na odročnih lokacijah, kamor žične povezave ne sežejo. Pri tem pa pri uporabnikih ne bo potrebno nameščati fiksnih satelitskih krožnikov in podobne opreme.

V splošnem so problemi satelitskih omrežij visoke cene satelitov, izstrelitve le-teh v tirnice in dražje opreme na uporabnikovi strani. Slednjemu so se pri SpaceX-u še posebej posvetili in cena uporabniškega terminala naj ne bi presegala meje 200 ameriških dolarjev.

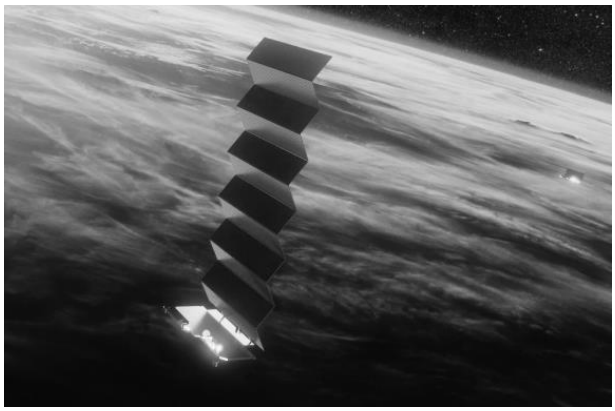
Starlink bo časovne zakasnitve minimiziral s tem, da bodo sateliti krožili okrog Zemlje v LEO. Glavna pomanjkljivost satelitov v LEO je majhno vidno polje. To pomeni, da lahko posamezen satelit opazuje oz. komunicira le z majhnim območjem Zemlje in to za kratek čas (na višini 550 km približno 5 minut). Zato je za stalno pokritost potrebna konstelacija večjega števila manjših satelitov, ki skupaj delujejo kot omrežje in zagotavljajo povezljivost v vsakem trenutku. Poleg majhnega vidnega polja pa prinaša LEO tudi hitro upadanje višine tirnice zaradi zračnega upora v višjih Zemljinih plasteh ozračja, zaradi česar je za ohranjanje tirnice potrebno ponovno pospeševanje ali pa izstrelitev novega nadomestnega satelita. Hitro upadanje tirnice pa ima tudi pozitivno stran, saj bodo nedelujoči sateliti sami padli proti Zemlji in izgoreli v ozračju, kar pomeni tudi manj vesoljskih smeti.

Internetni promet, ki poteka prek današnjih satelitskih omrežij v geostacionarnih tirnicah, ima minimalne teoretične zakasnitve med satelitom in uporabniškim terminalom vsaj 477 ms. V praksi so številke malo višje, in sicer okrog 600 ms ali več [10]. Nižje tirnice Starlinkovih satelitov prinašajo bolj praktične nižje zakasnitve od 25 ms do 35 ms, kar je primerljivo z že obstoječimi kabelskimi optičnimi omrežji. Sistem bo uporabljal protokol 'vsak z vsakim' (angl. peer-to-peer – P2P), ki bo preprostejši od IPv6, vključeval pa bo tudi šifriranje na celotni poti – od konca do konca (angl. end-to-end).

Leta 2017 so dobili od Zvezne komisije za komunikacije (angl. Federal Communications Commission – FCC) dovoljenje za izstrelitev 12.000 satelitov, za katere je načrtovano, da bodo krožili okrog Zemlje na treh različnih višinah. Prvih 1.600 satelitov bo v tirnicah na višini 550 kilometrov, naslednjih 2.800 satelitov bo delovalo v K_u (12–18 GHz) in K_a (26,5–40 GHz) frekvenčnem pasu na višini 1.150 kilometrov, ostalih 7.600 satelitov pa bo uporabljalo V frekvenčni pas (40–70 GHz) in bodo krožili okrog Zemlje na višini 340 kilometrov (angl. Very low earth orbit – VLEO [11]. Oktobra 2019 pa je SpaceX že podal nov zahtevek za 30.000 novih satelitov, ki naj bi krožili okrog Zemlje na višinah od 328 do 580 kilometrov. S prvimi 12.000 sateliti v LEO bodo pokrivali ves svet med geografskimi širinami 70° severno in 70° južno. Konstelacija takšnih razsežnosti pa skrbi astronome, saj verjamejo, da bi tolikšno število svetlih predmetov motilo opazovanje vesolja. SpaceX je v odgovor na Starlink 2 (misija) satelite nanesel eksperimentalni premaz, ki naj bi zmanjšal bleščanje.

4 Vesoljski segment

Vsak Starlink satelit tehta 227 kilogramov in je po oceni velikosti 3,5 m x 1,5 m (slika 1). Prvih 1584 satelitov bo krožilo okrog Zemlje v 72 tirnicah (v vsaki 22 satelitov), razmaknjenih na različnih višinah in položajih, da bi se izognili trkom. Sateliti bodo enakomerno razporejeni po tirnicah, tako da bodo imeli vedno na vidiku sosednje satelite v svoji tirnici in prav tako satelite v sosednji tirnici.

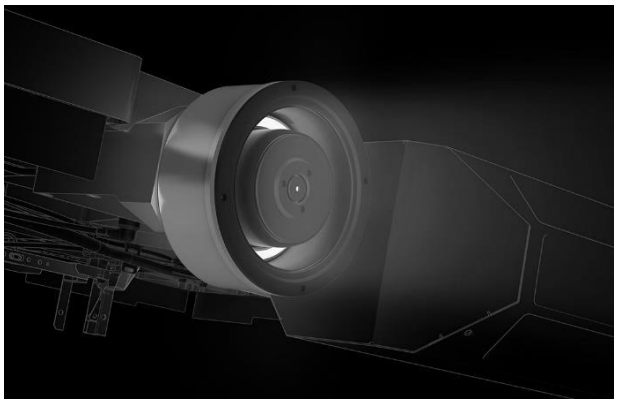


Slika 1: Vsak satelit bo velikosti 3,5 m x 1,5 m in bo tehtal 227 kg, opremljen bo z ionskim potisnikom, enim sončnim panelom in štirimi laserskimi komunikacijskimi terminali. (Vir: spacex.com)

Internetni promet in usmerjanje bo potekalo preko satelitov in ne preko vozlišč na Zemljinem površju. Posamezna zemeljska postaja se bo povezala s

satelitom, ki je v njenem vidnem polju, in ko se bo ta satelit premaknil izven dosega, bo to povezavo predal na naslednji satelit po svoji tirnici. Nato se bo signal preusmeril skozi niz medsatelitskih povezav, dokler ne bo dosegel satelita, ki je v vidnem polju ciljne zemeljske postaje.

Sateliti so rezultat masovne proizvodnje, kar pomeni veliko nižjo ceno na posamezno enoto kot pri ostalih podobnih satelitih. Za ohranjanje tirnic, pridobivanje višine in posledično tudi za izhod iz tirnice se uporabljajo pogonski motorji HET (angl. Hall effect thruster) s plinom kripton, ki so vrsta ionskih potisnikov, v katerih je pogonsko gorivo pospešeno s pomočjo električnega polja (slika 2). HET potisniki s plinom kripton ponavadi kažejo bistveno večjo erozijo pretočnega kanala v primerjavi s podobnim električnim pogonskim sistemom, ki deluje s ksenonom, vendar z nižjimi stroški pogonskega goriva.



Slika 2: Ionski potisniki, ki služijo za dosego končne tirnice in ohranjanje le-te. (Vir: spacex.com)

4.1 Medsatelitske zveze

Sateliti bodo med seboj komunicirali s pomočjo optičnih laserskih zvez (čeprav do sedaj izstreljeni sateliti niso opremljeni z laserskimi zvezami, temveč z radiofrekvenčnimi (RF) zvezami), podatke pa bodo na Zemljo pošiljali v frekvenčnih pasovih Ka, Ku in V, odvisno od višine tirnice, v kateri bodo. Povezava do Zemlje bo radijska, saj je bolj zanesljiva kot laserska. Radijski signali za razliko od infrardečih mnogokrat bolje prodirajo skozi oblačnost in področja padavin. Za podjetja, ki se bodo vseeno odločila za laserske povezave do Zemlje, je organizacija za standarde Consultative Committee for Space Data Systems (CCSDS) priporočila uporabo 1550 nm za promet navzdol (angl. downlink) in 1030–1080 nm za promet navzgor (angl. uplink), tako da so valovne dolžine ločene.

Optična komunikacija med sateliti bo uporabljala laserje in je v večini razvita na podlagi dosedanjih zemeljskih optičnih sistemov. Laserji oddajajo infrardečo svetlobo, ki je prav tako elektromagnetno valovanje kot radijski valovi, le da ima kar nekaj prednosti. Ker je frekvenca infrardeče svetlobe veliko višja (od 300 GHz do 470 THz), je na razpolago več pasovne širine, kar pomeni, da lahko teoretično prenesemo mnogokrat več informacij kot z radijski valovi (dosežena rekordna hitrost z milimetrskimi valovi je 40 Gbit/s, z lasersko povezavo pa 13160 Gbit/s). Prav tako infrardeči signali niso tako regulirani kot radijski. Razlog za neregulacijo svetlobe ni v tem, da je novejša tehnologija, ampak v mehanizmu širjenja radijskega valovanja. Radijski signal, poslan z nizkozemeljske tirnice, lahko pokrije več tisoč kvadratnih kilometrov področja na Zemlji in če več ljudi uporablja isti frekvenčni kanal, pride do prekrivanja oz. motenj. Po drugi strani pa veliko bolj usmerjen laserski žarek, poslan na primer z Lune, pokrije le okrog 20 km^2, zaradi tega in velikega razpona frekvenc zelo težko pride do motenj. Prav tako se za lasersko komunikacijo porabi veliko manj energije (radijski oddajnik na Luni bi porabil več kot 50 x več energije kot infrardeči oddajnik). V vesolju je doseg laserske komunikacije več deset tisoč kilometrov, zato je primerna za medsatelitske povezave. Zelo usmerjen žarek predstavlja inženirske izzive, kot je »ciljanje« z žarkom in sprejemanje le-tega, saj že najmanjše vibracije pomenijo na razdalji več deset tisoč kilometrov odstopanje (par metrov). Pri laserski komunikaciji je bistven kompromis med usmerjenostjo in močjo – manjše je odstopanje snopa, večja bo sprejemna moč, a večje zahteve po usmerjenosti.

Prvo lasersko komunikacijo so uspeli izvesti Japonci leta 1995 [12], od takrat so povezave vedno hitrejše (danes že tudi do 10 Gbit/s), tehnologijo pa razvija in uporablja vedno več podjetij. Laserske komunikacije so uporabili v projektu European Data Relay System leta 2014 med sateliti v LEO in GEO, LeoSat za satelite v LEO, prav tako pa tudi BridgeComm in Cloud Constellation.

SpaceX je za tehnologijo svojih medsatelitskih linkov namenil veliko sredstev in ker se je ostala konkurenca tudi odločila za uporabo laserske komunikacije, nočejo razkriti podrobnosti. So pa v vlogi za FCC navedli, da bodo njihovi laserji oddajali elektromagnetno valovanje s frekvenco 10.000 GHz, kar pomeni infrardečo svetlobo. Vsak satelit bo opremljen s štirimi laserskimi komunikacijskimi terminali: eden spredaj, eden zadaj in po eden na vsakem boku (slika 3). Terminala spredaj in zadaj se povezujeta s sateliti v isti tirnici, zato ostajata ves čas enako usmerjena, terminala na bokih pa se morata premikati in slediti satelitom v drugih tirnicah, saj se relativno drug na drugega ves čas premikajo.



Slika 3: Konstelacija Starlink satelitov s štirimi laserskimi terminali na vsakem satelitu

Paul Cornwell iz podjetja Mynaric pravi, da bi lahko 7–15 kg težak laserski terminal prenašal informacije s hitrostjo 10 Gbit/s na 4500 km dolgi medsatelitski povezavi [13]. Mynaric je nemško podjetje za lasersko komunikacijo, ki razvija in izdeluje laserske izdelke, ki omogočajo hitre brezžične prenose podatkov na dolge razdalje. Njihovi trenutni izdelki podpirajo 10 Gbit/s hitrosti, v prihodnosti pa obljubljajo večterabitne povezave. Na svoji spletni strani imajo zapisano, da bodo od leta 2020 naprej pripravljeni na masovno produkcijo svojih laserskih terminalov, kar lahko med drugim pomeni, da s svojimi produkti ciljajo na podjetja z večjimi konstelacijami, kot je SpaceX.

5 Zemeljski segment

Nedvomna prednost geostacionarne tirnice je v dokaj preprosti izdelavi sprejemnih zemeljskih postaj, saj je satelit vedno nad isto točko Zemlje in je zemeljska antena usmerjena vedno v isto točko na nebu.

Ker se sateliti Starlink gibljejo zelo hitro v svojih tirnicah, je sledilna oprema za antene zelo draga (ne glede na to, ali je sledenje izvedeno pri zemeljski opremi ali pa sateliti usmerjajo svoje sevalne snope v določene točke ob vsakem preletu).

Uporabniški terminali za omrežje Starlink se za razliko od drugih podobnih ponudnikov (Iridium, Globalstar, Thuraya in Inmarsat) ne bodo direktno povezovali na satelitsko omrežje, ampak se bodo povezali z bazno postajo v velikosti satelitskega krožnika s premerom 0,48 m. Bazna postaja Starlink bo opremljena z več antenami, ki bodo s pomočjo faznih sukalnikov sledile satelitom. [14] Bazne postaje so lahko nameščene kjerkoli, le pod pogojem, da imajo odprt pogled na nebo.

Za prenos podatkov sateliti uporabljajo mikrovalovne frekvence, ki jih glede na valovno dolžino razdelimo v več frekvenčnih pasov. Pri satelitskem internetu se uporabljajo pasovi L (1–2 GHz), Ku (12–18 GHz), K (18–27 GHz), Ka (27–40 GHz) in V (40–75 GHz). Načeloma višja frekvenca omogoča možnost uporabe širših frekvenčnih kanalov in s tem prenos več informacij, a prinaša tudi svoje težave, kot je neposredna vidljivost (angl. line of sight – LOS), kar pa pri satelitskih komunikacijah ni problem. Zaradi večjega slabljenja skozi ozračje se dandanes večinoma uporablja pas Ka, v prihodnosti pa lahko pričakujemo uporabo pasu V. Glede na dokumentacijo, ki jo je Starlink oddal na FCC, bodo zemeljski terminali imeli antene na 11,83 GHz z dobitkom 33,2 dBi in 14,25 GHz z dobitkom 36,6 dBi. [14] Zemeljski oddajniki bodo imeli oddajno moč manjšo od 4,06 W in adaptivno modulacijo od BPSK do 64QAM [14], kar bo omogočalo prilagodljivost komunikacijskega kanala glede na vremenske razmere.

6 Zaključek

Vloga Starlinka je zapolnjevati telekomunikacijsko praznino, ki vlada v manj razvitih delih sveta, in vzpostaviti konkurenco v predelih, kjer je trenutno mogoč dostop preko enega samega operaterja. Vzdrževanje zvez na relacijah satelit–Zemlja in satelit–satelit ni trivialno, ker se topologija omrežja spreminja. Zahtevana je hitra predaja zvez zaradi velikih hitrosti gibanja satelitov. Protokoli morajo imeti toleranco za relativno visoke zakasnitve in slabljenja signalov pri komunikaciji satelitov z zemeljskimi postajami ali terminali. Poleg satelitov pa so pomembni še zemeljski sprejemniki in oddajniki ter uporabniški terminali.

7 Literatura

[1] Hubert Fröhlich, Boštjan Batagelj “Širokopasovni mobilni dostop preko satelitskih komunikacij”, *Triindvajseta delavnica o telekomunikacijah VITEL: Širokopasovna mobilna omrežja*, 23 in 24. november 2009, Brdo pri Kranju. Slovenija, str. 40–43, 2009.

[2] “Satelitski internet, kjerkoli in povsod”, Novatel d.o.o., dostopano 20. 7. 2020, dostopno na: https://satelitski-internet.si/

[3] Neva Blazetič “Internet prek satelita”, *Primorske novice*, 22. April 2009, št. 93, str. 6, 2009.

[4] Boštjan Batagelj “5G ni nič revolucionarnega, zarota proti človeštvu pa prav tako ne”, *Monitor*, let. 30, št. 3, str. 28–31, 2020.

[5] Matjaž Vidmar “Tirnice umetnih satelitov”, 2. poglavje v *Uporaba vesoljskih tehnologij*, urednik Drago Matko, Radovljica: Didakta, 1996.

[6] Tomaž Rodič, Drago Matko, Krištof Oštir, Boštjan Batagelj, Marko Peljhan, Tomaž Zwitter, Barbara Malič, Tomaž Šuštar “Small satellites technologies from newcomers perspective - Slovenian case Space.si”, *Proceedings of the 4S Symposium. Small Satellites Systems and Services Symposium*, Madeira, Portugal, 31 May - 4 June 2010, Centre National d'Études Spatiales: = CNES, str. 1–11, 2010.

[7] Elon Musk “SpaceX Seattle 2015”, dostopano 20. 7. 2020, dostopno na: https://youtu.be/AHeZHyOnsm4 (3 min–4 min)

[8] “Sateliti Starlink BODO uničili pogled na zvezdno nebo!“, e-Spika, 3. 12. 2019, dostopano 20. 7. 2020, dostopno na: http://astronomska-revija-spika.si/bodo-sateliti-starlink-unicili-pogled-na-zvezdno-nebo/

[9] Tomaž Zwitter, “Bomo gledali le še satelite in ne zvezd?“, Večer, 1. 2. 2020, dostopano 20. 7. 2020, dostopno na: https://www.vecer.com/10123755

[10] Jon Brodnik, “Satellite Internet faster than advertised, but latency still awful“, Arstechnica, dostopano 20. 7. 2020, dostopno na: https://arstechnica.com/information-technology/2013/02/satellite-internet-faster-than-advertised-but-latency-still-awful

[11] Jeff Foust “SpaceX's space-Internet woes: Despite technical glitches, the company plans to launch the first of nearly 12,000 satellites in 2019”, in IEEE Spectrum, vol. 56, no. 1, pp. 50–51, Jan. 2019, doi: 10.1109/MSPEC.2019.8594798.

[12] Kenichi Araki, Yoshinori Arimoto, Motokazu Shikatani, Masahiro Toyoda, Morio Toyoshima, Tetsuo Takahashi, Seiji Kanda, Koichi Shiratama, “Performance evaluation of laser communication equipment onboard the ETS-VI satellite," Proc. SPIE 2699, Free-Space Laser Communication Technologies VIII, (22 April 1996); https://doi.org/10.1117/12.238434

[13] Jeff Hecht, “Laser Links will link small satellites to Earth and each other”, Laser Focus World, marec 2020, dostopano 20. 7. 2020, dostopno na: https://www.laserfocusworld.com/lasers-sources/article/14104017

[14] “Application for Earth station authorizations”, dostopano 11. 7. 2020, dostopno na: https://fcc.report/IBFS/SES-LIC-20190211-00151/1619048.pdf

Generator THz optičnega signala z DFB lasersko diodo

Blaž Pongrac[1], **Dušan Gleich** [1]

[1] *Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, 2000 Maribor, Slovenija*
E-pošta: blaz.pongrac1@um.si

THz signal generator using a Single DFB Laser Diode

This paper presents a frequency-modulated optical signal generator in the THz band. The proposed method is based on a fast optical frequency sweep of a single narrowband laser diode used together with an optical fiber interferometer. The optical frequency sweep using a single laser diode is achieved by generating short current pulses with a high amplitude, which are driving the laser diode. Theoretical analysis showed that the modulation frequency could be changed by the optical path difference of the interferometer or optical frequency sweep rate of a laser diode. This paper presents a method for constant frequency and chirped signal generation using a programmable current pulse waveform. The experimental results showed that the proposed signals could be generated within a microwave (1 - 30 GHz) and THz band (0.1 - 0.3 THz).

1 Uvod

V zadnjih letih se je povpraševanje po visoki nosilni frekvenci, visoki pasovni širini in visokih hitrostih prenosa močno povečalo. Frekvenčni pas mikrovalov in frekvenčni pas THz valov nam ponujata doslej še neraziskane možnosti hitrega prenosa podatkov, sploh frekvenčni pas THz valov (med 0,1 in 10 THz [1]), ki nam, zaradi svojih lastnosti [2], omogoča npr. detekcijo eksploziva, zajem slik z visoko resolucijo in širokopasovne komunikacije. Trenutne tehnologije generiranja in detektiranja THz valov so drage, saj na trgu primanjkuje cenovno dostopnih oddajnikov, detektorjev, valovodov z nizkimi izgubami in linearnih ter širokopasovnih naprav.

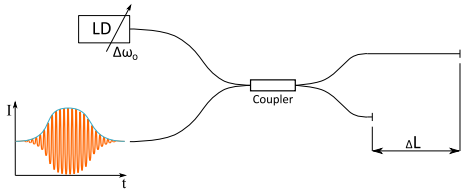
Opto-električne metode generiranja in detektiranja THz valov v večini temeljijo na vzbujanju fototoka v kristalih in polprevodniških napravah [3], ki jih imenujemo foto prevodne antene (ang. Photoconductive Antenna, PCA). PCA uporabljamo tako za generiranje, kot tudi za koherentno detektiranje THz valov. Avtorji v [4] so pokazali, da moramo za koherentno detektiranje THz valov PCA oddajnik in PCA detektor osvetliti s svetlobo iz istega svetlobnega vira. Trenutne metode generiranja optičnih signalov, ki jih uporabljamo pri generiranju/detektiranju THz valov s PCA, vključujejo uporabo fs-laserjev, s pomočjo katerih lahko generiramo kratek optični impulz [3], in množenje dveh periodičnih optičnih signalov z različno valovno dolžino, ki nam omogoča generiranje periodičnea signala s THz frekvenco [1]. Pri generiranju periodičnih signalov s THz frekvenco uporabljamo laserske diode s porazdeljeno povratno vezavo (ang. Distributed Feedback, DFB), ki jim lahko s temperaturo nastavljamo valovno dolžino generiranega laserskega valovanja. Nastavljanje te valovne dolžine je počasno [5] in zato ni primerno za aplikacije, kjer potrebujemo hitre spremembe.

V tem prispevku je prikazan način hitrega premika optične frekvence laserske svetlobe, ki nam omogoča generiranje optičnih signalov s frekvenco v THz frekvenčnem pasu. Predlagana metoda temelji na hitrem frekvenčnem premiku svetlobe, ki jo generiramo z ozkopasovnim laserskim virom, in neuravnoteženim optičnim vlakenskim interferometrom. Izpeljan je bil teoretični model predlagane metode, ki je pokazal, da je frekvenca generiranega signala odvisna od hitrosti frekvenčnega premika in neuravnoteženosti optičnega interferometra oz. razlike optičnih poti (ang. Optical Path Difference, OPD). OPD lahko nastavimo s spreminjanjem dolžine optičnega vlakna, za kar se uporabljajo piezoelektrični (PZT) kristali [6]. Raztegovanje optičnega vlakna s PZT kristalom ni primerno v predstavljeni metodi, saj lahko spremenimo dolžino vlakna za nekaj milimetrov pri dolžini vlakna nekaj metrov [6]. Avtorji v [7] in [8] predlagajo hitro linearno frekvenčno premaknitev laserske svetlobe z uporabo tokovnega sunka z veliko amplitudo in kratko dolžino. S predlaganim hitrim frekvenčnim premikom lahko frekvenco generiranega signala spreminjamo s frekvenca ponavljanja tokovnih sunkov. Glavna prednost predlaganega generatorja optičnih signalov je sposobnost generiranja kompleksnih oblik signalov. Iz eksperimentalnih rezultatov in teoretičnega modela lahko sklepamo, da lahko generiramo optične signale s frekvenco v mikrovalovnem in THz frekvenčnem pasu.

2 Visoko-frekvenčna optična modulacija

Predlagan optični generator je prikazan na Sliki 1. Sestavljen je iz frekvenčno nastavljivega optičnega vira in neuravnovešenega vlakenskega interferometra. Svetlobna jakost, odbita iz vlakenskega interferometra, je v splošnem definirana z:

$$I_r(t) = \frac{I_0}{2}\left(1 + \cos\left(\frac{2n\Delta L\omega(t)}{c}\right)\right) \qquad (1)$$



Slika 1: Predlagan optični generator, ki uporablja optični vir s frekvenčno premaknitvijo in debalansiran interferometer.

kjer je I_0 svetlobna jakost, n efektivni lomni količnik, ΔL predstavlja OPD, $\omega(t)$ optična krožna frekvenca in c hitrost svetlobe. Če uporabimo optični vir s frekvenčni premikom, postane $\omega(t)$ funkcija časa, fazo kosinusne funkcije v enačbi (1) pa lahko opišemo z:

$$\phi(t) = \frac{2n\Delta L}{c}\left(\omega_0 + \frac{\partial\omega}{\partial t}t\right) \tag{2}$$

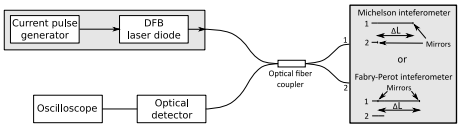
kjer je ω_0 izhodiščna optična krožna frekvenca. Privzamemo, da je hitrost premika optične krožne frekvence konstanta. Optična krožna frekvenca se po času linearno dviguje, odbiti optični signal iz interferometra pa dobi sinusno časovno karakteristiko. Neuravnovešen optični interferometer, ki ga napajamo z optičnim virom z linearno frekvenčno premaknitvijo, bo tako generiral sinusni optični signal, katerega frekvenco lahko določimo z:

$$f = \frac{2n\Delta L}{2\pi c}\frac{\partial\omega}{\partial t} = \frac{2}{c}n\Delta L\frac{\partial\nu}{\partial t} \tag{3}$$

kjer je $\partial\nu/\partial t$ hitrost optičnega frekvenčnega premik. Vidimo, da je frekvenca generiranega sinusnega signala enaka produktu hitrosti optične frekvenčne premaknitve in OPD. Poudarimo, da smo pri izbiri teh dveh parametrov omejeni, saj je optična frekvenčna premaknitev vedno omejena z obsegom in hitrostjo v vseh praktičnih virih. Zato je optična frekvenčna premaknitve v praktičnih optičnih virih vedno omejeno. Iz zapisanega sledi, da je dolžina optičnega signala vedno omejena, kar pa nam omeji največji možen OPD optičnega interferometra, saj moramo zagotoviti, da se dva optična signala prekrivata v prostoru. Poudarimo, da pri uporabi laserskih virov s počasno optično frekvenčno premaknitvijo potrebujemo izjemno velik OPD, če želimo generirati signale s frekvenco v THz frekvenčnem pasu. Kadar želimo generirati signala s frekvenco v mikrovalovnem frekvenčnem pasu ali THz frekvenčnem pasu, moramo uporabiti laserske vire s hitro frekvenčno premaknitvijo, zaželeno v dolžini nekaj 10 ns, kjer je OPD pod 1 km.

Kadar optično krožno frekvenco iz enačbe (2) v času spreminjamo nelinearno, lahko generiramo zahtevnejše oblike signalov. Če privzamemo, da je hitrost premaknitve optične frekvence $\partial\nu/\partial t$ linearna funkcija, lahko generiramo optični frekvenčni žvižg, ki je definiran kot:

$$f(t) = \frac{2}{c}n\Delta L\frac{\partial\nu}{\partial t}(t) \tag{4}$$



Slika 2: Zgradba predlaganega generatorja optičnih signalov z Michelsonovim in Fabry-Perotovim interferometrom.

3 Generator optičnih signalov s hitro premaknitvijo optične frekvence

Predlagan optični generator je prikazan na Sliki 2. Sestavljen je iz DFB laserske diode in generatorja tokovnih sunkov, ki nam služi za generiranje frekvenčno premaknjenega signala, in dveh različnih tipov optičnih vlakenskih interferometrov (Michelsonov in Fabry-Perotov interferometer). Uporaba vlakenskih interferometrov nam omogoča kompaktno in robustno izvedbo signalnega generatorja. Optični signali so bili merjeni z optičnim detektorjem N7004A proizvajalca Keysight in z osciloskopom UXR0334A proizvajalca Keysight.

3.1 Načrtovanje in izdelava neuravnovešenih vlakenskih interferometrov

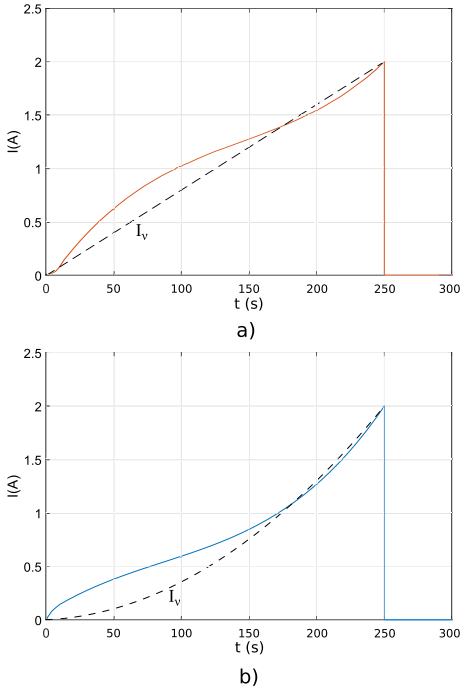
V predlaganem generatorju optičnih signalov sta bila uporabljena dva tipa optičnih vlakenskih interferometrov: (i) Michelsonov in (ii) Fabry-Perotov interferometer. Zrcala v Fabry-Perotovih vlakenskih interferometrih so bila izdelana s tehniko napraševanja visoko-odbojnih materialov (TiO_2), ki so bila nato varjena z ustrezni programom varjenja optičnih vlaken [9]. Odbojnost zrcal v Fabry-Perotovih interferometrih je dosegla 10 %. Zrcala v Michelsonovih vlakenskih interferometrih so bila izdelana z tehnologijo naprešavanja materialov (Al), dosežena pa je bila 30 % odbojnost zrcal. Izdelanih je bilo več interferometrov, ki so bili načrtani s pomočjo enačbe (3) glede na želeno frekvenco generiranega signala pri dolžini tokovnega sunka $T_s = 250$ ns. Dolžine interferometrov so predstavljene v Tabeli 1.

Tabela 1: OPD Michelsonovih in Fabry-Perotovih interfeormetrov ΔL za želeno frekvenco f.

f (GHz)	1	8	15	30	100	200	300
ΔL (m)	0.02	0.16	0.3	0.61	2.03	4.06	6.1

3.2 Generator tokovnih sunkov

Predlagan laserski vir s hitro optično frekvenčno premaknitvijo je zgrajen iz standardne telekomunikacijske DFB laserske diode, ki je tokovno vodena s kratkim tokovnim sunkom visoke amplitude. Tokovno vodenje temelji na metodi, ki so jo avtorji podrobneje opisali v [7]. Metoda uporablja selektivno in hitro gretje aktivnega področja DFB diode, pri čemer z ustrezno valovno obliko tokovnega sunka lahko ustvarimo linearni premik optične frekvence laserskega vira.



Slika 3: Oblika tokovnega sunka: a) linearni premik optične frekvense s $T_s = 250$ ns, b) nelinearni premik optične frekvence s $T_s = 250$ ns.

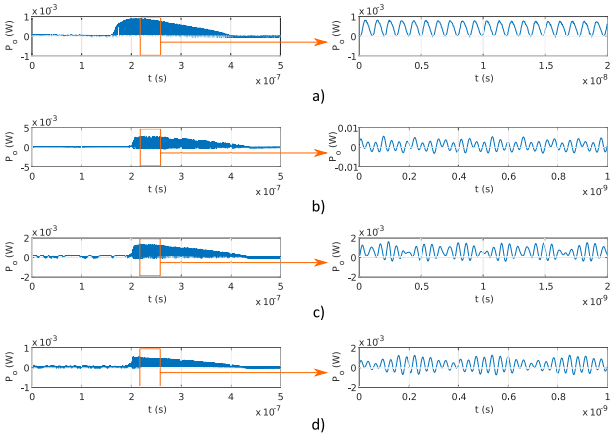
Obliko tokovnega sunka, dolžino sunka in njegovo maksimalno amplitudo so za linearni premik optične frekvence določili avtorji v [8] in je prikazana na Sliki 3 (a). Tokovni sunek lahko določimo kot $I(t) = I_\nu(t) + I_{DFB}(t)$, kjer je I_ν linearna funkcija premika optične frekvence, ki nam pove, da je hitrost spreminjanja optične frekvence konstanta funkcija časa, in $I_{DFB}(t)$ je karakteristična krivulja izbrane DFB laserske diode in je enaka ne glede na obliko signala, ki jo želimo generirati. $I_{DFB}(t)$ je bila eksperimentalno določena v [8]. Če spremenimo I_ν, lahko generiramo različne oblike optičnih signalov. Oblika tokovnega sunka, s pomočjo katere lahko generiramo optični frekvenčni žvižg, je kvadratične oblike, kot je prikazana na Sliki 3 (b).

4 Eksperimentalna evalvacija

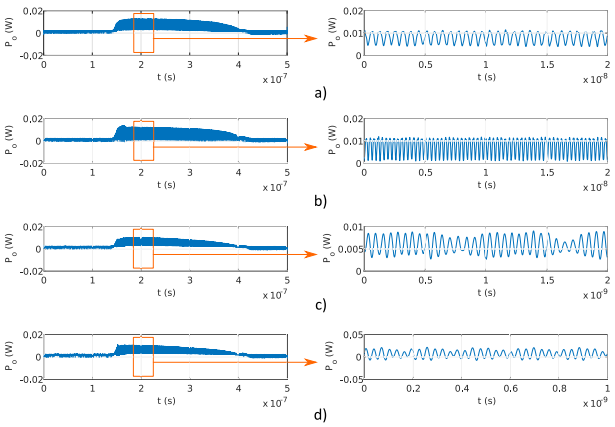
Generator optičnih signalov je bil evalviran v dveh različnih scenarijih. V prvem scenariju je bila uporabljena linearna premaknitev optične frekvence v dolžini tokovnega sunka $T_s = 250$ ns in interferometri obeh tipov z OPD iz Tabele 1. V drugem scenariju sta bila uporabljena Michelsonov interferometer z $\Delta L = 0,033$ m in Fabry-Perotov interferometer z $\Delta L = 0,02$ m ter oblika tokovnega sunka, ki je predstavljena na Sliki 3 (b).

4.1 Generiranje signalov z linearno premaknitvijo optične frekvence

Slika 4 (a)-(d) in Slika 5 (a)-(d) prikazuje generirane optične signale z dolžino tokovnega sunka $T_s = 250$ ns pri uporabi Michelsonovih in Fabry-Perotovih interferometrov z OPD iz Tabele 1, kjer je želena frekvenca manjša ali enaka 30 GHz. Generiranim optičnim pulzom ob začetku



Slika 4: S Fabry-Perotovim interferometrom generirani signal z OPD ΔL iz Tabele 1: (**a**) 1 GHz, (**b**) 8 GHz, (**c**) 15 GHz, (**d**) 30 GHz.

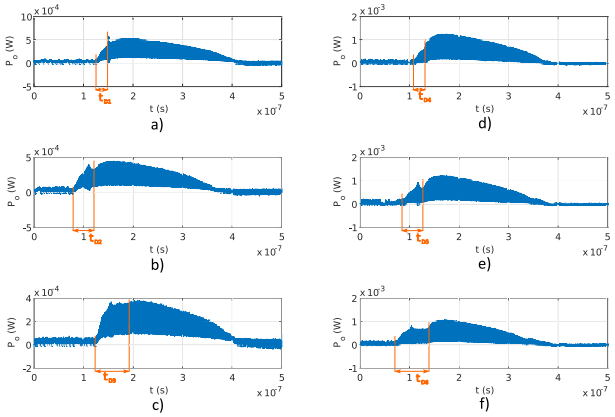


Slika 5: Z Michelsonovim interferometrom generirani signal z OPD ΔL iz Tabele 1: (**a**) 1 GHz, (**b**) 8 GHz, (**c**) 15 GHz, (**d**) 30 GHz.

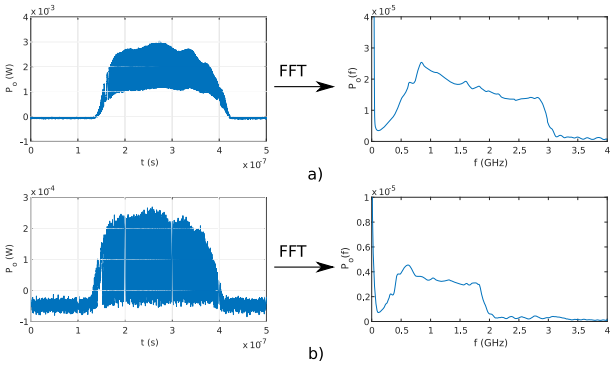
pulza amplituda hitro naraste, nato pa zaradi naraščanje temperature v aktivnem področju DFB laserske diode pride do zmanjšanja emisije fotonov. Moč optičnega signala zato začne padati in pade na vrednost 0 ob koncu tokovnega sunka pri času T_s. Iz signalov, pri katerih je pričakovana frekvenca pod 30 GHz lahko frekvence generiranih signalov razberemo iz Slike 4 (a)-(d) in Slike 5 (a)-(d) ter jih prikažemo v Tabeli 2. Frekvenca generiranega signala je stabilna, odstopa pa od želene vrednosti zaradi razlike med ocenjenim OPD in med praktično implementacijo interferometra. Iz Slike 6 (a)-(f) lahko razberemo zakasnitve t_d, ki so posledica večjega OPD. OPD je pri želenih frekvencah nad 100 GHz v Tabeli 1 še dovolj majhen, da v interferometrih pride do prekrivanja

Tabela 2: Izmerjene frekvence f_M and f_{FP} pri uporabi Michelsonovega in Fabry-Perotoveg interferometra z OPD iz Tabele 1.

f(GHz)	f_M (GHz)	f_{FP} (GHz)
1	1.6	1.02
8	6.25	7.9
15	16.3	15.3
30	29.2	31.1



Slika 6: Generirni optični signali z OPD interferometrov ΔL iz Tabele 1: (**a**) Fabry-Perotov interferometer, $f = 0.1$ THz (**b**) Fabry-Perotov interferometer, $f = 0.2$ THz, (**c**) Fabry-Perotov interferometer, $f = 0.3$ THz (**d**) Michelsonov interferometer, $f = 0.1$ THz (**e**) Michelsonov interferometer, $f = 0.2$ THz, (**f**) Michelsonov interferometer, $f = 0.3$ THz.



Slika 7: Generiran optični frekvenčni žvižg in pripadajoči spekter: (**a**) Michelsonov interferometer z $\Delta L = 0,033$ m, (**b**) Fabry-Perotov interferometer z $\Delta L = 0,02$ m.

odbitih signalov. Zaradi omejitve merilne opreme lahko z opremo, ki nam je na voljo, detektiramo samo ovojnico generiranega signala.

4.2 Generiranje signalov z nelinearno premaknitvijo optične frekvence

Slika 7 (a)-(b) prikazuje generiran optični frekvenčni žvižg. Generiran je bil s pomočjo tokovnega sunka, ki je prikazanega na Sliki 3 (b). Uporabljena sta bila Michelsonov interferometer z OPD $\Delta L = 0,033$ m in Fabry-Perotov interferometer z OPD $\Delta L = 0,02$ m. Iz spektra generiranih signalov lahko določimo generirane frekvence, in sicer 1 GHz do 3 GHz za Michelsonov interferometer in 0,5 GHz do 2 GHz za Fabry-Perotov interferometer. Ker s segrevanjem aktivnega področja DFB diode optična moč pada, padajo tudi amplitude generiranih frekvenc, saj so višje frekvence generirane proti koncu dolžine tokovnega sunka T_s.

5 Sklep

V članku je predstavljena metoda generiranja frekvenčno moduliranega optičnega signala, s pomočjo katerega lahko generiramo optične signale s frekvenco v THz frekvenčnem pasu. Generator optičnih signalov je sestavljen iz DFB laserske diode in generatorja tokovnih sunkov, s pomočjo katerega lahko izvedemo linearno in nelinearno premaknitev optične frekvence laserskega sevanja. Predlagani generator optičnih signalov je bil eksperimentalno evalviran. V prvem eksperimentu je bil uporabljen tokovni sunek, s pomočjo katerega smo izvedli linearno premaknitev optične frekvence. Kot rezultat so bili generirani optični signali s sinusno karakteristiko. Merjeni signali so bili v skladu s teoretičnim modelom. V drugem eksperimentu je bil generiran optični frekvenčni žvižg s pomočjo tokovnega sunka, ki omogoča linearno spreminjanje hitrosti premika optične frekvence. S pomočjo obeh eksperimentov je bilo pokazano, da lahko s predlaganim generatorjem optičnih signalov generiramo tako signale s konstantno frekvenco, kot tudi bolj kompleksne oblike signalov (frekvenčni žvižg), generirani optični signali pa so primerni za generiranje in detektiranje THz valov z uporabo PCA. V nadaljnjem delu je smiselno raziskati možnosti nastavljanja generirane frekvence s spremembo oblike toka, saj je spreminjanje OPD nepraktično.

Raziskave in razvoj so bile sofinancirane s strani ARRS, št. projekta J7-9408.

Literatura

[1] Safian, R.; Ghazi, G.; Mohammadian, N. Review of photomixing continuous-wave terahertz systems and current application trends in terahertz domain. Opt. Eng. 2019, 58, 1. doi:10.1117/1.oe.58.11.110901.

[2] Naftaly, M.; Vieweg, N.; Deninger, A. Industrial applications of terahertz sensing: State of play. Sensors (Switzerland) 2019, 19.

[3] Yun-Shik Lee. Principles of Terahertz Science and Technology; Springer US, 2009. doi:10.1007/978-0-387-09540-0.

[4] Saeedkia, D. Terahertz photoconductive antennas: Principles and applications. Proc. 5th Eur. Conf. Antennas Propagation, EUCAP 2011 2011, pp. 3326–3328.

[5] Guo, R.; Lu, J.; Liu, S.; Shi, Y.; Zhou, Y.; Chen, Y.; Luan, J.; Chen, X. Multisection DFB Tunable Laser Based on REC Technique and Tuning by Injection Current. IEEE Photonics J. 2016, 8, 1–7. doi:10.1109/JPHOT.2016.2585923.

[6] Henderson, D.A.; Hoffman, C.; Culhane, R.; Viggiano III, D. Kilohertz scanning all-fiber optical delay line using piezoelectric actuation. Fiber Opt. Sens. Technol. Appl. III 2004, 5589, 99.

[7] Njegovec, M.; Donlagic, D. Rapid and broad wavelength sweeping of standard telecommunication distributed feedback laser diode. Opt. Lett. 2013, 38, 1999. doi:10.1364/ol.38.001999.

[8] Njegovec, M.; Donlagic, D. Interrogation of FBGs and FBGs Arrays Using Standard Telecom DFB Diode. J.Light. Technol. 2016, 34, 5340–5348.

[9] Bendoula, R.;Wacogne, B.; Pieralli, C.; Gharbi, T. Fabrication of an intrinsic fiber Fabry-Perot interferometer involving simple in situ control of multilayer TiO2-SiO2 mirrors: Material properties studies and cavity interrogation procedure. Optical Engineering - OPT ENG 2005, 44. doi:10.1117/1.2049273.

Izboljšanje dobitka BoR-monopolnih anten z algoritmom Diferencialne Evolucije

Marko Radović, Peter Kitak, Gorazd Lešnjak, Peter Planinšič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, Maribor, Slovenija
E-pošta: marko.radovic@student.um.si

The Gain Enhancement of BoR-Monopolde Antennas using Diferential Evolution Algorithm

Abstract. *Two electrically thick body BoR monopole antennas are proposed with enhanced antennas gain performances. Both antennas are designed above infinite ground plane, fed with coaxial cable and to operate in frequency range from 0.3 GHz to 1.2 GHz. Voltage standing wave ratio and realized gain characteristics are observed and studied. The enhancement was obtained using optimization with Differential Evolution (DE) algorithm. Proposed antennas have better gain performances in comparison to reference conical cylindrical antenna. Results obtained for second antenna show improved impedance bandwidth performances in the upper part of frequency band in comparison to reference antenna.*

1 Uvod

Zanimanje za razvoj in izboljšanje rotacijsko simetričnih kompaktnih anten traja že več kot četrt stoletja [1], [2]. Današnje širokopasovne in ultra - široko-pasovne antene UWB (*Ultra-Wide Band*), z visokimi in vedno višjimi nosilnimi frekvencami, pogosto uporabljajo kompaktne in majhne antene na tiskanem vezju ali celo antene, integrirane v visokofrekvenčnem čipu. Vendar so kompaktne antene, z majhnimi dimenzijami in fizično robustne rotacijsko simetrične antene, še vedno zelo atraktivne, zaradi njihovih neusmerjenih sevalnih karakteristik, dobrega dobitka in drugih lastnosti. Primerne so za zahtevna okolja, kot so moderne brezžične komunikacije pri avtomobilih in vojaških komunikacijah. Primer takšnih robustnih in simetričnih anten so antene oblikovane z vrtenjem profila- BoR-antene (*Body of Revolution*).

Analizo in sintezo električno debelih BoR-anten so podali Djordjević, Dragović in Popović v prispevku [2]. Predstavili so metodo analize in snovanja teh anten s pomočjo računalnika. Svojo metodo so potrdili z eksperimentalnimi rezultati. Med primeri njihovih anten je BoR- antena, imenovana cilindrično-konična monopolna antena, ki smo jo v našem delu vzeli kot referenčno in poimenovali Referenčna antena.

Drugi pristop k oblikovanju BoR-monopolnih anten je bila metoda za oblikovanje in izboljšanje lastnosti UWB-anten s profilom "naključne hoje" (*random walk profile*), ki jo je predstavil Zhao v prispevku [3]. Nizko-profilno UWB monopolna antena z obliko obrnjenega klobuka – IHA (*Inverted Hat*), pa je Zhao predstavil v [4]. Oblikovana je bila za delo v frekvenčnem področju od 50 MHz do 2 GHz, ki bi bila primerna za majhna brezpilotna zračna plovila. V prispevku [5] je isti avtor uporabil genetske algoritme – GA (Genetic Algorithms) za optimizacijo nizkoprofilne UWB-IH- monopolne antene. Miniaturne kompaktne osno-simetrične resonančne lečne antene za Ka-frekvenčni pas so opisane v literaturi [6]. Celovit pregled širokopasovnih in UWB anten za brezžične aplikacije je podan v literaturi [7]. Metode analize in aplikacije nizko-profilnih naravnih in metamaterialnih anten je podan v [8].

Moderne stohastične, bio-zasnovane optimizacijke metode, omogočajo učinkovito optimizacijo kompleksnih nelinearnih problemov in so bile pogosto uporabljene za optimizacijo različnih vrst anten. Pregled uporabe evolucijskih algoritmov (EA) za antene in razširjanje valovanj je bil narejen v [9].

Eden od popularnih bio-zasnovanih algoritmov iz družine Evolucijskih Algoritmov (EA) je algoritem Diferencialne Evolucije (DE). V naši študiji oblikovanja BoR-monopolnih anten smo uporabili klasični algoritem DE, [10], kombiniran z uporabo modeliranja in analiziranja anten s programskim orodjem za reševanje 3D-problemov elektromagnetike, imenovan WIPL-D Pro [11]. WIPL-D Pro omogoča hitro in natančno analizo prevodnih in/ali dielektričnih/prevodnih struktur. Klasični DE je znan kot zelo enostaven in učinkovit. Prav tako so BoR-antene relativno enostavne za analizo in oblikovanje. Kljub temu je bila motivacija naše študije raziskati, ali je mogoče doseči nadaljnja izboljšanja oblikovanja BoR-anten z uporabo (klasične) DE.

Predlagali smo dva nova tipa električno debelih BoR-monopolnih anten, BoR-model$_1$ and BoR-model$_2$. Oblikovali smo ju s primerjavo njihovih električnih karakteristik (dobitek) s kakteristiko Referenčnega modela (antene). Geometrijo trojno eliptične IH-antene, ki jo je predstavil Zhao v [3], smo uporabili za oblikovanje geometrije antene BoR-model$_1$. Drugo anteno BoR-model$_2$ smo zasnovali na osnovi geometrije optimirane UWB BoR monopolne antene #2, ki jo je prav tako predstavil Zhao, in sicer v prispevku [4]. Vse antene smo modelirali v orodju WIPL-D Pro v11, s predpostavko, da so narejene iz električno idealnega prevodnika in impendančno prilagojene na koaksialni kabel 50 Ω.

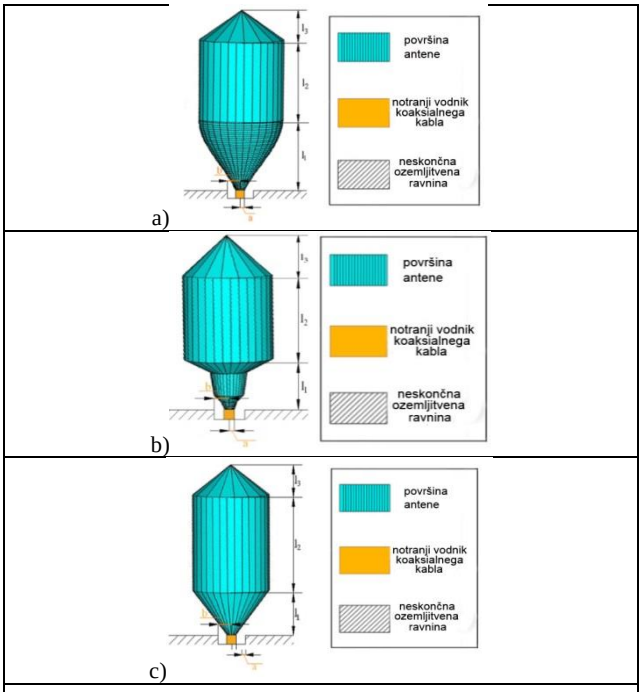
Analizo smo izvajali v frekvenčnem območju od 0,3 GHz do 1,2 GHz. Simulacije dobitka obeh novih predlaganih monopolnih anten kažejo izboljšanje dobitka v frekvenčnem pasu od 0,45 GHz do 1,1 GHz, Rezultati za napetostno razmerje stojnega vala – VSWR (Voltage Standing Wave Ratio) izkazujejo izboljšanje. Tudi za impendačno pasovno ujemanje pri anteni BoR-model2 vidimo izboljšanje v zgornjem delu frekvenčnega pasa in sicer od 0,9 GHz do 1,15 GHz.

2 Metodologija in rezultati

2.1 Geometrijske oblike modelov anten

Slika 1 prikazuje, da vse tri antene (BoR-model$_1$, BoR-model$_2$, in Referenčni model) sestavljajo trije geometrični deli dolžin l_1, l_2 in l_3. Polmer srednjega cilindričnega dela pri vseh antenah označimo z R. Referenčna antena (model) je sestavljena iz dveh področij in cilindra.

Modifikacijo geometrijske oblike obeh predlaganih anten v primerjavi z Referenčno anteno smo naredili le v spodnjem delu, ki je povezan na koaksialno linijo. Vse tri antene smo modelirali kot nameščene pravokotno nad neskončno ozemljitveno ravnino in napajane preko koaksialnega kabla s karakterističnо impendanco 50 Ω. Geometrijski modeli vseh treh anten zmodelirani s programskim orodjem WIPL-D Pro so bili izdealni iz idealnega prevodnega materiala. Polmer a na sliki 1 je notranji polmer koaksialnega kabla in sicer znaša 3.1 mm, zunanji polmer b pa znaša 7.59 mm, za vse tri antene.



Slika 1. Geometrijski modeli električno debelih monopolnih anten: a) BoR-model$_1$, b) BoR-model$_2$, c) Referenčni model Geometrijski obliki obeh predlaganih anten smo dobili z geometrijsko modifikacijo spodnjega dela Referenčne antene.

2.2 Kratek opis algoritma DE

Algoritem Diferencialne Evolucije sta prva predlagala Storn in Price v članku [10], ki je pozneje postal zelo opazno, popularno in močno optimizacijsko orodje v inžinerstvu iz družini evolucijskih algoritmov [12]. Do danes je bilo predlagano mnogo izboljšav klasičnega algoritma DE [13].

Osnovni princip algoritma DE je sledeč. Je stohastični, na populaciji zasnovan algoritem za globalno optimizacijo nelinearnih, ne-odvodljivih, šumnih in ravnih funkcij z več parametri (spremenljivkami), z omejitvami in mnogimi lokalnimi minimumi. Prične z uniformnim nakjučnim naborom kandidacijskih rešitev iz izvedljivega iskalnega volumna. V vsaki iteraciji, to je generaciji algoritma, izvaja enake računske korake kot standardni Evolucijski Algoritem (EA). Vendar se DE znatno razlikuje od GA v tem, da mutira osnovni vektor, to je sekundarne starše s skaliranimi razlikami pomembnih članov trenutne generacije. To lastnost imenujemo samo-referencirana mutacija. Medtem ko EA in nekateri Genetični Algoritmi (GA) zahtevajo adaptacijo vsake spremenljivke pri iteracijah, kanonični DE zahteva le adaptacijo enojnega skalirnega faktorja za vse spremenljivke. V nasprotju od mnogih EA tehnik, je osnovni DE zelo enostaven algoritem, katerega implementacija zahteva le nekaj vrstic kode v katerem koli standardnem programskem jeziku, kot je na primer Matlab, ki smo ga uporabili v našem primeru (slika 2). Dodatno, osnovni DE zahteva le tri kontrolne parametre, skalirni faktor, prečno hitrost in velikost populacije (slika 3). Toda presenetljivo, DE se zelo dobro obnaša pri optimiranju objektivne funkcije z več spremenljivkami v smislu končne natančnosti, računske hitrosti in robustnosti.

Kot smo omenili, smo pri konstrukciji prvega geometrijskega dela dveh novih anten, izhajali iz idej, ki jih je predstavil Zhao za konstrukcijo profila IH-antene in BoR-monopolne antene. Profili prvega dela predlaganih in referenčne antene so prikazani na sliki 4.

Za oblikovanje profila prvega geometrijske oblike prvega dela antene BoR-model$_1$, smo uporabili profil, ki je podoben profilu trojne-eliptične IH-antene, ki ga lahko matematično opišemo kot:

$$y = \frac{R}{2}(1 + cos\ \frac{\pi n t}{l_1}) \qquad (1)$$

Kjer spremenljivka t pripada intervalu [0, R], faktor n pa izberemo tako, da dobimo $y = a$ pri $t = R$. Izraz za n je:

$$n = \frac{l_1}{\pi R}\left(arccos\frac{2a}{R} - 1\right) \qquad (2)$$

2.3 Oblikovanje in optimizacija anten z algoritmom DE

Proces optimizacije oblike obeh novih anten smo izvajali z iskanjem štirih optimalnih geometrijskih

parametrov oziroma dimenzij R, l_1, l_2 in l_3, da bi dosegli izboljšano obnašanje novih anten. Dobljeni rezultati dimenzij za obe novi anteni in referenčno anteno so podani v tabeli 1.

Dobitek antene je definiran kot razmerje med referenčno gostoto moči P isotropičnega vira in gostote moči P_{di} v določeni smeri opazovanja. Običajno se podaja v dBi [14]:

$$G_{dBi} = 10 \cdot log_{10}(\frac{P_{di}}{P}) \quad (3)$$

V reševanju našega optimizacijskega problema z algoritmom DE smo spreminjali štiri geometrijske parametere antene l_1, l_2, l_3, in R, da bi maksimirali dobitek antene G_{dBi} v opazovanem frekvenčnem področju v primerjavi z referečno anteno (modelom). V algoritmu DE smo minimizirali kriterijsko funkcijo, ki je bila obratno sorazmerna vsoti absolutne vrednosti razlik ojačenja obravnavane novo oblikovane in Referenčne antene, v 31 točkah.

Za vsako dobljeno novo anteno smo opazovali tudi napetostno razmerje stojnega vala VSWR, ki je merilo odbite moči, ki potuje nazaj k oddajniku po priključenem kablu zaradi neprimerne prilagoditve karakteristične impedance kabla na vhodno impendanco antene. VSWR zapišemo kot [14]:

$$VSWR = \left(\frac{Vmax}{Vmin}\right) \quad (4)$$

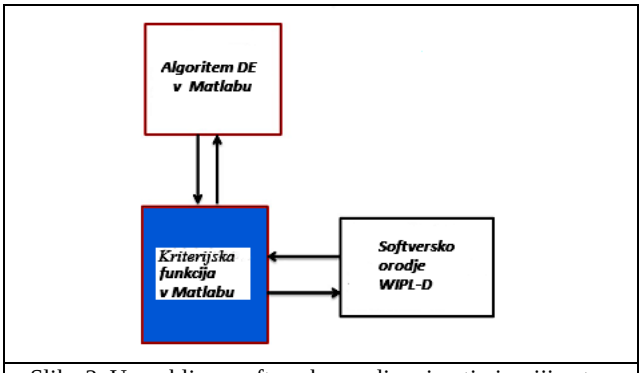
kjer sta V_{max} in V_{min} maksimalna in minimalna amplituda napetosti stojnega valovanja na napajalnem kablu. Dobra prilagoditev (ujemanje) je pri vrednost približno 1.2. V primeru popolne prilagoditve je ta vrednost 1. Povezavo VSWR z impedancami lahko zapišemo z izrazom [15]:

$$VSWR = \frac{1+|\Gamma|}{1-|\Gamma|} \quad (5)$$

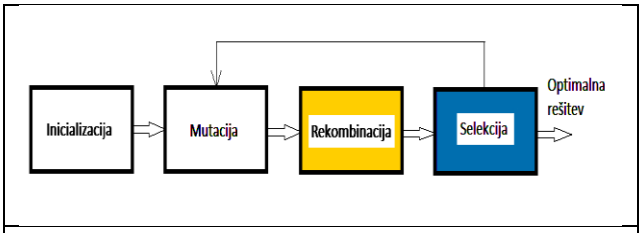
kjer je Γ napetostni reflekcijski koeficient na vhodnem priključku antene. Definiran je kot:

$$\Gamma = \frac{Z_{in}-Z_0}{Z_{in}+Z_0} \quad (6)$$

kjer je Z_{in} vhodna impedanca antene in Z_0 karakteristična impendanca napajalnega kabla (prenosne linije). V primeru popolnega ujemanja je Z_{in} enaka Z_0 (v našem primeru 50 Ω). Potem je vrednost Γ enaka 0 in VSWR enaka 1. VSWR je povezana s kompleksnim frekvenčno odvisnim S-parametrom S_{11}, ki je enak refleksijskemu koeficientu Γ in je običajno narisan v Smithovem diagramu [15].



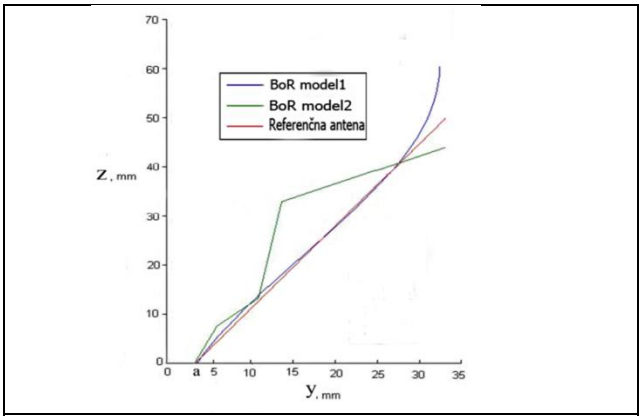
Slika 2. Uporabljena softverska orodja pri optimizaciji anten



Slika 3. Principielna blokovna shema algoritma DE

Tabela 1. Optimirani parametri anten z DE

Parameter	l_1 (mm)	l_2 (mm)	l_3(mm)	R (mm)
BoR-model$_1$	60,62	72,61	23,55	32,41
BoR-model$_2$	44,12	80,67	34,24	33,00
Referenčni model	50,00	87,00	3,00	33,00



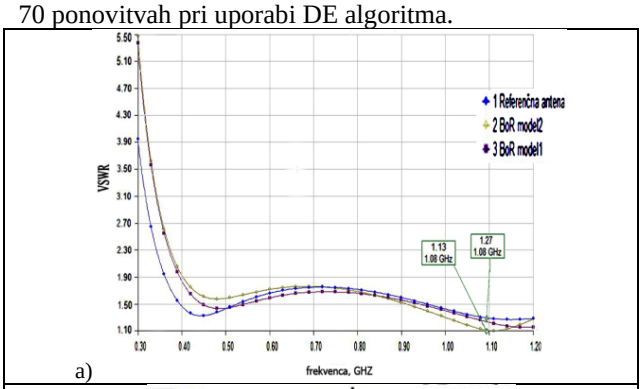
Slika 4. Spodnji del 2D-profila novih in Referenčne antene

2-D profil prvega geometričnega dela antene BoR-model$_2$ je sestavljen iz štirih segmentov, podobno kot optimirana antena #2, ki jo je predlagal Zhao. V optimizacijskem procesu UWB IH-monopolne antene predlagatelja Zhao, so linijski segmenti 2D-profila odvisni od njihove dolžine in kota. V našem primeru optimizacije 2D-profila antene BoR-model$_2$ se je ta izvajale le vzdolž z-osi, prikazano na sliki 4.

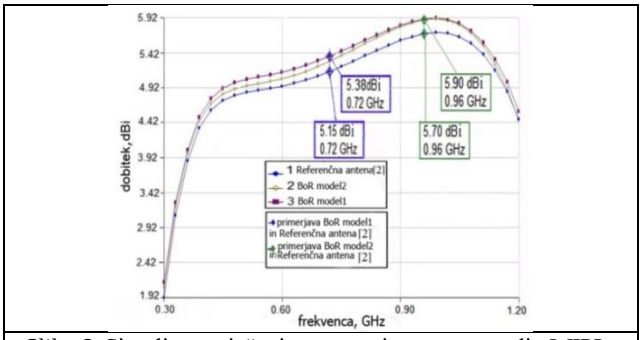
Dve predlagani novi anteni in Referenčno anteno smo simulirali v orodju 3D EM Solver WIPL-D Pro v11. Rezultati VSWR v frekvenčnem področju od 0,3 GHz do 1,2 GHz so za vse tri antene prikazani na sliki 5, VSWR v kartezičnih koordinatah, kompleksni parameter S_{11} pa v Smithovem grafikonu. Opazimo lahko, da se graf antene BoR-model$_1$ skoraj popolnoma prekriva z grafom Referenčne antene. Izboljšanje pasovne impendančne karakteristike antene BoR-model$_2$ je razvidno v zgornjem delu frekvenčnega območja, 0,9 GHz to 1,15 GHz. Maksimalno doseženo izboljšanje impendančne karakteristike za BoR-model$_2$ je 12,39% pri frekvenci 1,08 GHz in je označeno na sliki 5 z zelenim markerjem.

Slika 6 prikazuje simulirane rezultate ojačenja v istem frekvenčnem področju, za vse tri antene. Opazimo lahko, da imata obe predlagani anteni izboljšan dobitek v primerjavi z Referenčno anteno, kar je bil osnovni namen naše optimizacije z algoritmom DE.

To je evidentno v frekvenčnem pasu 0,45 GHz do 1,1 GHz, manj pa v preostalem frekvenčnem pasu. Optimirani parametri za obe novi BoR anteni so pridobljeni na PC AMD FX(tm)-8300 8-jedrni procesor (64 bitni proceser) 3,3 GHz 8, GB RAM v 14 urah po 70 ponovitvah pri uporabi DE algoritma.



Slika 5. a) Simulirane VSWR za vse tri anterne, b) parameter S_{11} v Smithovem grafikonu za vse tri antene



Slika 6. Simulirana ojačenja za vse tri antene v orodju WIPL-D Pro v11

3 Zaključek

Z algoritmom DE smo oblikovali dve novi električno debeli BoR-monopolni anteni, BoR-model$_1$ and BoR-model$_2$. Elektromagnetne simulacije smo naredili za obe novi anteni in Referenčno anteno, v frekvenčnem območju 0,3 GHz to 1,2 GHz. Rezultati simulacije v predlaganem frekvenčnem pasu in konfiguraciji so pokazali izboljšanje dobitka obeh novih anten v primerjavi z referenčno. To je jasno razvidno v frekvenčnem območju 0,45 GHz to 1,1 GHz. Dobili smo tudi izboljšano impendančno prilagoditev antene BoR-model$_2$ v frekvenčnem področju 0,9 GHz to 1,15 GHz, v primerjavi s kanonično-cilindrično referenčno anteno.

ZAHVALA*:* Avtorji se zahvaljujejo prof. dr. Igorju Tičarju in prof. dr. Dušanu Gleichu za podporo.

Literatura

[1] M. G. Andreasen, "Scattering from bodies of revolution", *IEEE Trans. Antennas Propag*., 1965, 13, (2), pp. 303–310.

[2] A. R. Djordjević, M. B. Dragović, B. D. Popović, "Analysis of Electrically thick antennas of revolution", Proc of 3rd ICAP, Norwich , April 1983, pp. 31–35

[3] J. Zhao, T. Peng,T., C.-C. Chen, J. L. Volakis, "Low-profile ultra-wideband inverted-hat monopole antenna for 50 MHz–2 GHz operation", *Electronics Letters*, 2009, vol. 45, no 3, pp. 142-144, July 2011, pp. 3295–3296

[4] J. Zhao, D. Psychoudakis, C.-C. Chen, J. L. Volakis, "Ultra-wideband performance optimization of a body-of-revolution monopole antenna", IEEE (APSURSI) Int. Sym. Antennas Propag., 2008, Spokane, WA

[5] J. Zhao, D. Psychoudakis, C.-C. Chen, and J. L. Volakis, Design Optimization of a Low-Profile UWB Body-of-Revolution Monopole Antenna, *IEEE Transaction on Antennas and Propagation*, vol. 60, no. 12, December, 2012, pp. 5578-5586.

[6] A. Rolland, M. Ettorre, A. V. Boriskin, L. L. Coq, and R. Sauleau, Axisymmetric Resonant Lens Antenna With Improved Directivity in Ka-Band, *IEEE Antennas and Wireless Propagation Letters*, vol. 10, 2011, pp. 37-40.

[7] R. Cicchetti, E. Miozzi, and O. Testa, Wideband and UWB Antennas for Wireless Applications:A Comprehensive Review, *International Journal of Antennas and Propagation*, Volume 2017, Article ID 2390808, 45 pages, Hindawi Publishing Corporation, https://doi.org/10.1155/2017/2390808

[8] H. Nakano, *Low-Profile Natural and Metamaterial Antennas: Analysis Methods and Applications*, Willey, 2016.

[9] S. K. Goudos, C. Kalialakis, and R. Mittra, Evolutionary Algorithms Applied to Antennas and Propagation: A Review of State of the Art, *International Journal of Antennas and Propagation*, Volume 2016, Article ID 1010459, 12 pages, Hindawi Publishing Corporation, http://dx.doi.org/10.1155/2016/1010459

[10] R. Storn, K. Price, "Differential Evolution: A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces", *J. Global Optimization*, 1997, vol. 11, no 44, pp. 341-359.

[11] Wipl-d Pro v11 simulator "Wipl-dpro 3D electromagnetic solver for fast and accurate analysis of arbitrary metallic and dielectric/ma-gneticstructures", http://www.wipl-d.com/products.php?cont=wipl-d-pro, accessed July 2015.

[12] G. C. Onwubolu, B. V. Babu, *New Optimization techniques in Engineering*, Springer, 2004.

[13] D. Swagatam S. S. Mullick, and P. N. Suganthan," Recent Advances in Differential Evolution – An Updated Survey", *Swarm and Evolutionary Computation*, February 2016

[14] T. Manning, *Microwave Radio Transmission Design Guide*, Artech House, 2006.

[15] C. A. Balanis, *Antenna Theory. Analysis and Design*. Wiley-Interscience, 2005.

Satelitska podpora vplutju v pristanišče in privezovanju

Franc Dimc, Blaž Luin, Marko Perkovič

Fakulteta za pomorstvo in promet, Univerza v Ljubljani, Pot pomorščakov 4, Portorož
E-pošta: franc.dimc@fpp.uni-lj.si

Satellite-based harbour entrance and berthing

Abstract. *Navigational measures at the port entrance are well regulated by IMO performance specifications, but with the demand for more transported goods, safety requirements are evolving. Modern maritime electronic positioning systems on their way to autonomous ships develop integrity measures and exchange information with other ships and authorities. An adaptive and satellite-based solutions lead to radio navigation aids similar to those used in aviation domain.*

1 Uvod

Zgoščevanje ladijskega prometa narekuje v priobalni plovbi še posebej natančno spremljanje in pogosto izmenjavo zanesljivih podatkov [1]. Nerazumevanje ali slabo tolmačenje podatkov je eden od glavnih vzrokov trčenj in nasedanj ladij. Zato morajo biti ladijskim častnikom v straži, pilotom na poveljniškem mostu in zaradi avtomatskega pošiljanja podatkov tudi operaterjem obalne nadzorne službe (*Vessel Traffic System*, VTS) na voljo zanesljivi podatki zadosti časa, da lahko ustrezno odločajo in ukazujejo izvajanje potrebnih manevrov. Če odmislimo klasični pogled skozi okno, se pretežno odločajo glede na podatke informacijskega sistema za spremljanje in prikaz podatkov na elektronski karti (ECDIS) o položaju plovila, navigaciji in času (PNT), ki temelji na klasičnih instrumentih (kompas, RADAR, brzinomér, globinomer) in na svetovnih satelitsko podprtih sistemih (GNSS). Navigacijske naprave morajo zaradi doslej povedanega izpolnjevati posebne zahteve po točnosti, integriteti, razpoložljivosti in neprekinjenosti dostopa do podatkov PNT [2].

Ladja s sporočili avtomatskega identifikacijskega sistema (AIS) seznanja posadke bližnjih ladij in operaterja VTS s trenutnim položajem in integriteto svojih naprav. Napake modeliranja v podatkih naprav so posledica predvsem omejitev modeliranja pogojev sprejema satelitskih signalov in le delnega upoštevanja ostalih omejitev ladijskih elektronskih naprav.

Integriteta naprav, ki so izvori podatkov PNT, je razvrščena v tri razrede: visok, srednji in nizek. Sam podatek o integriteti uporabniku sporoča, ali je sistem trenutno sposoben omogočati varno plovbo oz. je zanjo sploh uporaben [3]. Integriteto GNSS opisujejo: mejna vrednost oz. meja opozarjanja (*alert limit, AL*), čas do alarma in ocena tveganja integritete. Glede na zahteve Mednarodne pomorske organizacije (*International Maritime Organization* , IMO) za plovbo v pristanišču je za AL določenih 25 m, napaka določanja položaja do 10 m, čas do alarma zaradi nedelovanja sistema pa 10 s, tveganje integritete pa 10^{-5} v 3 urah [1], [2]. Med plovbo brez znanih pravih vrednosti sprejemnik GNSS sam ocenjuje napake določanja položaja plovila. Po algoritmu za spremljanje stanja integritete naprave (*Receiver Autonomous Integrity Monitoring*, RAIM) sodobni sprejemnik stalno računa na svojo statistiko podatkov psevdorazdalj vezano trenutno oceno napake izračuna položaja, imenovano stopnja zaščite (*Protection Level, PL*). Tveganje izpada integritete pa pomeni verjetnost, da bo sprejemnik kadarkoli na katerem koli mestu na območju pokritosti s signali GNSS uporabniku določil napako položaja, povzročeno ali z napakami instrumentov na ladji (*Vessel Technical Error*, VTE) ali s sistematskimi navigacijskimi napakami navigacijskih sistemov (*Navigational Systematic Error*, NSE), ki bo takrat od pravega položaja oddaljen več kot za AL, ne da bi se v predpisanem času od pojava napake sprožil alarm, ki se lahko prenese kot posebna zastavica (*RAIM flag*) v sporočilih AIS. V praksi zaenkrat nastopajo omejitve, ker nekateri sprejemniki sploh ne izdajajo zastavice RAIM v sporočilih in ker morda ladijski AIS zastavice ne prenaša.

Če torej sprejemnik GNSS v priobalni plovbi pravočasno opozori odgovornega, da je PL presegla AL ali da podatki niso na voljo ter se tako razglasi za nezanesljivega, ostane kot naprava vreden zaupanja, saj je pravočasno opozoril, naj se odgovorni ne zanaša na njegove trenutne podatke. S katerimi podatki pa si bo pri odločanju pomagal v tem primeru, je odvisno od senzorjev in naprav, ki so še vključene v ECDIS.



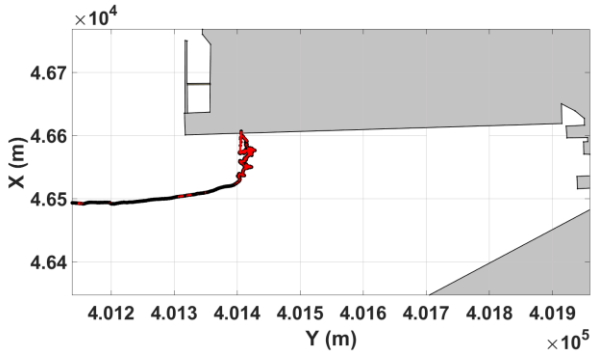
Slika 1. Skica pojava večpotja (večžarkovni prenos), Koper.

2 Območja napak podatkov GNSS

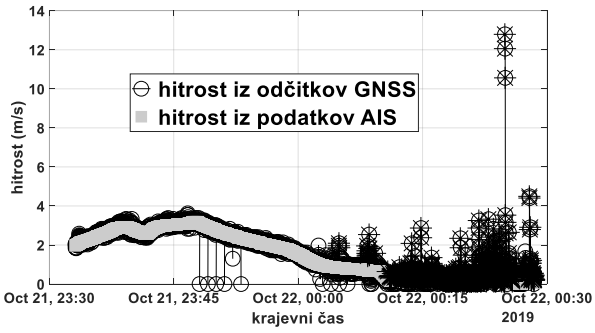
Ladijskim pilotom napake v podatkih sprejemnikov GNSS določanja položaja dejansko otežujejo pregled nad stanjem plovbe in usklajevanje s poveljniki

vlačilcev, privezovalci ladje ter službo VTS. V pristanišču se ladjam z mlajšimi generacijami sprejemnikov GNSS dogaja redkeje, ladjam s starejšimi, večinoma eno-frekvenčnimi, pa se sorazmerno pogosto zgodi, da se jim pri pristaniških napravah zaradi odbojev satelitskih signalov od kovinskih konstrukcij pojavljajo težave zaradi nenadnih izrazitih skokov.

Pojav splošno ponazarja Slika 1, rezultate tega pojava pa štiri slike, ki ji sledijo. GNSS sledi ladij sta prikazani na Sliki 2 in Sliki 4, obrežje je prikazano s sivo barvo. Opazen je vpliv večpotja oz. večžarkovnega prenosa (*multipath*), ki se zaradi napak pri določanju psevdorazdalj, odraža v določanju položaja in iz njih izračunanih trenutnih vrednostih hitrosti (na Sliki 3 in Sliki 5) med manevrom pristajanja 1. in 2. bazenu pristanišča Koper.



Slika 2. GNSS sled (D96, UTM) pristajanja kontejnerske ladje Concord: 6500 TEU, LOA 304,16 m, Koper, 22.10.2019



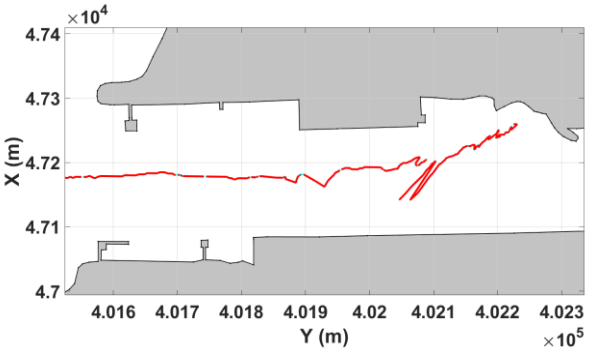
Slika 3. Med plovbo v 1. bazenu zabeležena hitrost ladje Concord, Koper, 22.10.2019.

Krogi na Sliki 3 in Sliki 5 označujejo hitrosti, izračunane iz odčitkov položajev GNSS, sivi kvadrati pa odčitke hitrosti, ki jih je za AIS priredil GNSS. S črnimi zvezdicami poudarjeni izračuni, ko hitrost iz odčitkov položajev GNSS premikajoče ladje več kot dvakrat presega hitrost z AIS, so bili zabeleženi, ko se je ladja z manjšo hitrostjo bližala pomolu.

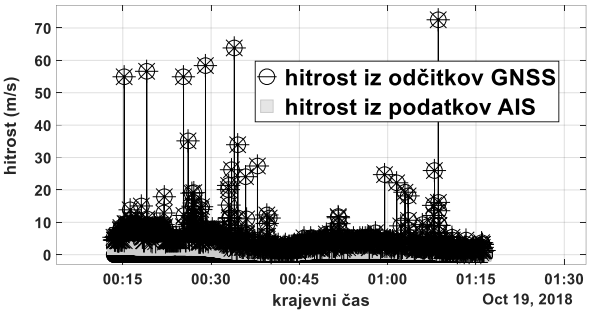
V težavah s podatki položaja je potrebno z RAIM zaznati porast NSE, kar je pojem, uvožen iz letalskega prometa. Največjo NSE je mogoče opisati z vsoto v danem trenutku pričakovanih sistematskih napak, kar ponazarja Enačba (1):

$$NSE_{\max} = nap.karte + nap.GNSS + druge\,napake \quad (1)$$

Iz primerjave Slike 4 in Slike 5 sklepamo, da je hitrost iz samih odčitkov GPS vso plovbo skozi 2. bazen zelo nerealna in vsaj dvakrat presega SOG z merilnika hitrosti.



Slika 4. GNSS sled ladje Couga (D96, UTM) za prevoz razsutega tovora: DWT 50806, LOA 182,98 m, Koper, 18.10.2018.



Slika 5. Med plovbo v 2. bazenu zabeležena hitrost ladje Couga, Koper, 18.10. 2018.

Območja, kjer se pogosto pojavlja večpotje, je možno ocenjevati iz sporočil AIS oz. t.i. NMEA stavkov. Zaradi negotovosti ocenjevanja dolžine psevdorazdalj do satelitov, določenih iz kod, je mogoče statistično določati z zbiranjem podatkov oz. beležiti mesta pogostih nenadnih preskokov podatkov AIS o položaju ladij.

Kartiranje območja večpotja pa je bolj analitično določljivo iz *surovih* podatkov. Oblika takšnih podatkov je za rezultate položaja z največjo točnostjo, neodvisno od uporabljenega sprejemnika, predpisana v obliki, imenovani RINEX (*Receiver Independent Exchange Format*), ki pa jih v sporočilih AIS zaenkrat ni. Kartiranje lahko temelji na analizi močnostnega spektra razmerja signal šum vsakega posameznega zaznavnega satelita [4].

Veliko naporov se vlaga v modeliranje napak večpotja in s tem v njihovo zmanjševanje na različnih področjih uporabe GNSS. Modeliranje napak vključuje poznavanje: oddaljenosti objektov od antene sprejemnika GNSS, njihove relativne usmerjenosti

odsevnih površin na zveznico do antene, sevalnega diagrama antene in iz almanahov znanih položajev satelitov nad obzorjem.

Kartiranje območij, bogatih z napakami, je pomembno v času priprave postopkov certificiranja sprejemnikov GNSS, ki bi ustrezali zahtevam določenih procesov (npr. plutja v območju dvigal), saj služba VTS lahko opozarja poveljnika ladje, ki vpluje brez pilota, naj se na teh območjih v določenih konstelacijah satelitov manj zanaša na podatke, zbrane z GNSS.

Elegantno rešitev v pogojih večpotja predstavlja GNSS kompas. To je sprejemnik GNSS z več antenami, ki je sposoben z obdelavo signalov zmanjšati vplive sofaznih oz. sorazmerno razdalji med antenami zakasnjenih motenj, upoštevaje smer proti posameznemu satelitu.

3 Nadzor pristajanja in privezovanja

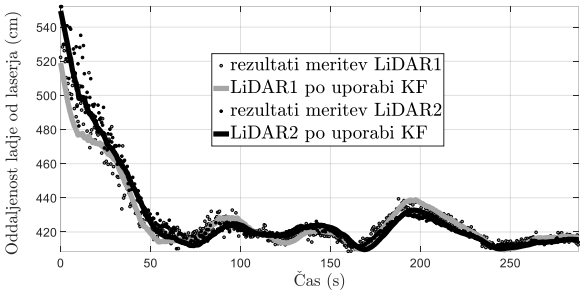
Privezovanje izjemno velikih ladij je vedno težaven proces, ki postane še bolj zapleten, kadar je treba s plovili ravnati 1) na omejenih območjih omejene globine, se hkrati 2) zoperstaviti še silovitemu bočnemu vetru ali 3) manevrirati v močnem toku ali celo v vseh treh omenjenih pogojih hkrati. Končni manever približevanja pomolu in natančno pozicioniranje na privezna mesta sta še posebej zahtevna na kontejnerskih terminalih, kjer je vzdolž pomola nameščenih mnogo dvigal ladja-obala, kar zelo omejuje možnost dopustne napake pri privezovanju, še posebej, če so dvigala nameščena v bližini krila poveljniškega mosta, premec ali krma pa segata čez rob obale (Slika 6).



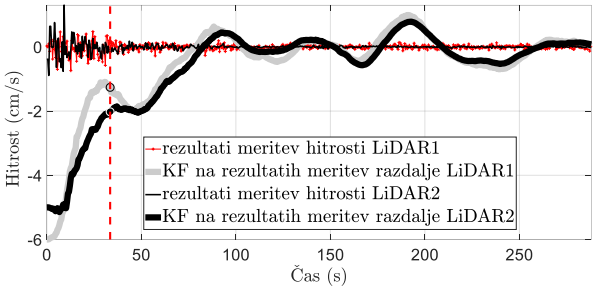
Slika 6. Ladja Concord, privezana v pristanišču Koper.

Da bi se izognili trkom, je potrebno bočno gibanje popolnoma nadzorovati; ladja mora biti vzporedna s pomolom, pri čemer se mora ohranjati še dopustna bočna hitrost približevanja in obenem preprečevati, da bi se plovilo bistveno pomaknilo vzdolžno. Podatki o položaju, pridobljeni z GNSS, tudi tisti, ki vključujejo diferencialne popravke, in sporočeni po AIS, vsebujejo bistveno prevelike napake za tako natančne postopke.

Nadzor omogoča LiDARski privezni in pristajalni sistem, ki je bil po preizkusih in uporabi ustrezne obdelave podatkov laserskih merilnikov z metodo Kalmanovega filtriranja (KF) (Slika 7), razvit za pristanišči Koper in Świnoujście (Poljska) [5].



Slika 7. Obdelava podatkov LiDARjev, zajetih med manevrom vzporednega približevanja pomolu na mirni gladini.



Slika 8. Obdelava generičnih podatkov o hitrosti in podatkov o oddaljenosti, obdelanih s KF s Slike 7.

Vpeljani LiDARski sistem v pristanišču Koper omogoča zanesljivo odčitavanje usmerjenosti ladje glede na os pomola (*angle-of-approach*) z ločljivostjo do 0,5°, posebno pomembna pa je zanesljiva meja določljivosti hitrosti 3,0 cm/s, saj se sme najtežja kontejnerska ladja, ki lahko vpluje v koprsko pristanišče, s tolikšno hitrostjo nasloniti na odbojnike pomola. Iz rezultatov na Slika 8 sklepamo, da sta se ob naslonitvi na odbojnike (trenutek označen s črtkano črto) obe *laserski točki* na boku ladje proti pomolu premikali pod mejno vrednostjo 3,0 cm/s.



Slika 9. Prikaz podatkov na prenosni pilotovi napravi (PPU).

Pilot na ladjo prinese prenosno enoto (PPU) in jo priključi na AIS vtičnico ladje. S PPU pilot zbira podatke LiDARskega sistema preko povezave Zig Bee in 4G, zraven pa tudi podatke senzorjev plime in dveh ultrazvočnih merilnikov vetra (Slika 9), preko priključka na ladijski sistem pa prejema še izvleček podatkov z navigacijskih sistemov plovila ter vlačilcev, ki delujejo na osnovi GNSS, da lahko vodi proces pristajanja in izplutja.

4 Prihodnost s prilagodljivimi rešitvami

Ker je manevriranje v pristaniščih omejeno, so raziskovalci skovali MVPA (*Marine Vessel Protection Area*) [7], na dve razsežnosti razširjen pojem AL, namenjen ladjam, ki za plovbo poleg integritete GNSS uporabljajo podatke o integriteti EGNOS (*European Geostationary Navigation Overlay Service*) ter poznajo ocene napak ladijskega kurza, ki je odvisen od lastnosti uporabljanega kompasa. Varno območje se na ECDIS predstavlja kot posebno obarvana površina.

Projektna skupina, v kateri sodeluje tudi Fakulteta za pomorstvo in promet Univerze v Ljubljani, si je v zaključenem projektu MAGS (*Maritime Adaptive GNSS Safety Concept*), ki ga je financirala Evropska vesoljska agencija, zastavila cilj ugotoviti kolikšen varen manevrski prostor potrebuje posamezna ladja glede na vremensko-prometne okoliščine ter na svojo opremljenost s senzorji in instrumenti ter svoj HPL (*Horizontal Protection Level*), za varno vplutje v območje pristanišča.

Ladja bo po načrtih skupine že v bližnji prihodnosti izmenjevala podatke o plovbi preko satelita (*VHF Data Exchange System*, VDES) z okoliškimi ladjami (podobno kot sedaj z AIS) in jih bo v predpisani obliki posredovala tudi obalnim postajam VTS oz. ustrezni pristaniški službi. Služba bo v skladu s parametri vplutja in trenutno prometno situacijo vsakemu plovilu posebej določila MVPA in odločila ali odobri njegovo približevanje pristanišču po zastavljeni poti in z načrtovanimi manevri ali ne, upoštevajoč MVPA vseh v nadzorovanem območju prisotnih plovil.

Po odobritvi vplutja se bo začela faza dejanskega vplutja v pristanišče, v kateri bo služba VTS spremljala premike vseh plovil, razdalja robov MVPA do robov območja varne plovbe se bo sproti prilagajala vremenskim podatkom (METOCEAN), vsem lastnim spremembam kurza in hitrosti pa se bo prilagajala tudi površina MVPA. Odgovorni na poveljniškem mostu bo lahko spremljal na ECDIS trenutno vrednost MVPA svojega plovila in vsa morebitna opozorila, saj jih bo obalna služba osveževala v realnem času. Za povečanje zanesljivosti je poleg rabe storitev EGNOS predvidena tudi vpeljava storitve evropskega satelitskega navigacijskega sistema Galileo, imenovana HAS (*High Accuracy Service*), v kateri je v sodobnih pogojih telekomunikacij zelo zaželena tudi storitev NMA (*Navigation Message Authentication*). NMA s preverjanjem pristnosti zagotavlja zaščito prenesenih podatkov in signalov, njena uporaba s sistemom javnih in zasebnih ključev praktično onemogoča potvarjanje (*spoofing*).

5 Sklepi

Ocena, da uporabljamo toliko satelitsko podprtih navigacijskih naprav, kolikor nas trenutno živi na svetu, odseva ne le izredno razširjenost ampak tudi izredno zanašanje uporabnikov na takšen način navigacije. Zato je mera upravičenega zaupanja v navigacijski podatek za varno navigacijo zelo pomembna. Odgovorni na poveljniškem mostu je ob pojavih različnih šumov in celo namernih motenj [6] mirnejši, če prejme vsaj verodostojno opozorilo naj bo previden, kaj šele, da v jasni grafični podobi sproti dobiva obvestila, kako varna je njegova plovba.

Sprejemnik GNSS, sposoben obdelave signalov več sistemov, združen z GNSS neodvisnimi senzorji, postaja običajna oprema v pomorskem prometu. Satelitsko podprta nadgradnja s prilagodljivimi rešitvami spremljanja in opozarjanja plovil sledi rešitvam v letalskem prometu [8], v katerem je varnost pri pristajanju letal in ostajanje v predvidenem krajevno-časovnem oknu na pristajalni stezi enakovredna vplutju ladje v območje pristanišča in ostajanje v dodeljeni MVPA ob vstopu v pristanišče v določenem, malo daljšem, časovnem obdobju.

Zahvala

Avtorji se zahvaljujemo kap. Matjažu Felicjanu za strokovni pregled in izboljšave besedila.

Literatura

[1] IMO, „Resolution A.915(22) Revised Maritime Policy and Requirements for a Future Global Navigation Satellite System (GNSS),“ 2001.

[2] IMO, „Resolution A.1046(27), Worldwide Radionavigation System.,“ 2011.

[3] IMO, „MSC.1/Circ.1575, Guidelines for Shipborne Position, Navigation and Timing (PNT) Data Processing,“ 2017.

[4] A. Bilich in K. M. Larson, „Mapping the GPS multipath environment using the signal-to-noise ratio (SNR),“ *Radio Science,* p. 16, 2007.

[5] M. Perkovič, L. Gucma, M. Bilewski, B. Muczynski, F. Dimc, B. Luin, P. Vidmar, V. Lorenčič in M. Batista, „Laser-Based Aid Systems for Berthing and Docking,“ *J. Mar. Sci. Eng.,* p. 20, 2020.

[6] Uredništvo, „Real-time interference detection by GIDAS makes satnav safer,“ *GPS World,* p. 1, 2020.

[7] P. Zalewski, L. Gucma, S. Gewies, K. Urbanska, S. Schlueter in M. Porretta, „Concept of EGNOS Implementation in the Maritime Domain,“ v *8th ESA Workshop on Satellite Navigation Technologies and European Workshop on GNSS Signals and Signal Processing, NAVITEC,* Noordwijk, Nizozemska, 2016.

[8] J. Burns, B. Clark, R. Cassell, C. Shively, T. Murphy in M. Harris, „Conceptual Framework for the Proposal for GBAS to Support CAT III Operations,“ ICAO NSP WGW WP11, Montréal, Kanada, 2009.

Multimodal Electronic Access Control System Using Digital Identity for Enabling Authorized Physical Access to Premises

Matej Rabzelj, Tjaša Jereb, Aljaž Martinčič, Andrej Kos

University of Ljubljana
Faculty of Electrical Engineering
Tržaška cesta 25, 1000 Ljubljana
E-mail: matej.rabzelj@ltfe.org, tjasa.jereb@ltfe.org, aljaz.martincic@ltfe.org, andrej.kos@fe.uni-lj.si

Abstract

With the advancement of consumer electronics and expansion of connectivity technologies, more and more devices are connected to the cloud. These include smart home appliances such as smart lights, home video systems and most importantly – smart locks. However, installation of an EAC (electronic access control) system should consider a number of security and usability aspects. This paper presents the developed EAC system which enables multimodal context-based authentication and provides access to the Student council's office at the Faculty of electrical engineering. It describes the integration of hardware mechanism and it's connection to the cloud, as well as the authorization process to ensure controlled access. Finally, it explains the importance of interaction design and overviews the security aspects of different modalities.

1 Introduction

Premises locked by means of electronic access control (EAC) offer enhanced physical security due to centralized monitoring and access logging systems. The most common type of physical EAC is role-based, where access to parts of premises is determined by the user's role in an organization [1]. Replacement of standard keys with EAC systems minimizes difficulties in case of their loss, as there is no need to replace the entire lock and keys of other users. Instead, restricting the access goes only as far as denying permissions for an individual electronic key [1]. Their usage is not exclusive to office buildings, as nowadays, a wide array of smart access solutions are presented through a trend of smart houses, introducing keyless entry to simple households and regular customers. Changes towards easier access are also visible in the car industry. While most car manufacturers already introduced key-less entry as a part of their basic equipment, some went even a step further, developing the new Digital key standard which will enable any smartphone or smartwatch to act as a digital car key [2].

The access control to the majority of cabinets, laboratories and offices at the University of Ljubljana, Faculty of electrical engineering is realised with an electronic system unlocked by appropriate RFID (radio-frequency identification) cards or keys. To open the door with this solution, one must approach the key next to the keypad near the door. Amongst premises with this system is the Student council's office on the ground floor of the faculty. Each year a newly elected set of students gains access to the office, leading to issuance of a new set of keys at the beginning of each mandate. These keys – while functional – bring some inconveniences. Starting with key programming which requires a time-consuming manual entry of each name using a dedicated control program. Then, when entering and exiting the office frequently, keys might be forgotten inside, leaving the user locked out. Furthermore, in case a key is lost, the access has to be manually revoked. These are only some aspects of a much wider challenge of designing a secure access control system while offering a seamless user experience.

To achieve their functionality, EAC systems make use of various technologies ranging from simple RFID tags to complex biometric scanning solutions [3]. However, the exact choice of authentication method typically represents a tradeoff between a system's security and users' convenience. Nevertheless, multiple modalities of interaction may be implemented to increase the solution's versatility and satisfy different use cases.

Therefore, our solution proposes the use of the already established digital university identity database to provide premises access control with the following modalities of interaction:

- voice-invoked access control,
- geolocation-based access control,
- proximity-based access control;

2 System Architecture

The developed EAC system consists of three main functional building blocks:

- a user-facing (mobile) client component for enabling various modalities of invocation,
- an authentication proxy for authentication against University of Ljubljana's remote identity directory,
- an electromechanical part comprised of physically secured components for latching action control;

An overview diagram of these modules and their interconnected interfaces is depicted in Figure 1. The configuration of components on the diagram follows the logical flow of an unlocking action - starting with the user intent at the top and resulting in mechanical action at the bottom of the figure.

The user who intends to access locked premises can interface with one of our four developed client applications. These provide the ability to control the door lock with both iOS and Android-based mobile terminals. Besides the touch input, these mobile applications can also be invoked using a corresponding Siri voice command or an on-location NFC (near field communication) tag. Aside from providing entry for themselves, users inside the building may also respond to incoming visitors without having to reach for their mobile phones or a remote doorbell switch. Instead, the unlocking mechanism may be activated using voice control provided by the Amazon Alexa device with a custom invocation command or using web and desktop applications.

Mobile clients and web application require users to log in prior to their first use, while Alexa requires the user to be present within its audible range. Login credentials are accepted in the form of username and password, which are securely conveyed to the authentication proxy server in exchange for a session token. This eliminates the need to store sensitive data on the device or in the cloud and only stores the time-constrained token instead. After initial authentication the application allows for fingerprint-assisted and geolocation-based unlocking. These methods provide some elementary restrictions, such as geofencing, to limit the application's functionality outside of the building's vicinity.

Further upstream, a LEMP stack (Linux, Nginx, MariaDB, PHP-FPM) middleware with secure API (application programming interface) endpoints serves as an authentication and session management gateway. It features multiple virtual network interfaces providing firewalled internet connectivity and communication with the University of Ljubljana's AD (Active Directory) server and the on-premise microcomputer module. At the forefront, Nginx performs TLS (transport layer security) termination of incoming requests, while the PHP OpenLDAP implementation escapes and authenticates the provided client credentials against the central AD server. Assuming success, the retrieved user data is further compared against local filters enforcing conditional access. Matching records from the unified authentication procedure then generate a pseudo-random session key, which is written to the MariaDB storage and returned to the client. All subsequent incoming requests with included session token are then compared against existing records and either forwarded to the server's command and control component or dropped if invalid. Lastly, the command and control module uses a virtual private network to communicate with the on-site microcomputer.

Contrary to the extensive list of software components, the on-premise electromechanical installation is rather uncomplicated. It consists of a preinstalled electric door latch (lock), mechanical relay, and an SBC (single board computer). The latter severely aids the prototyping process, as it offers on-board Ethernet connectivity and direct programmatic control of its GPIO (general purpose input and output) ports. It accepts network requests from the control module and issues executable scripts that activate the mechanical unlocking action. The whole installation is enclosed in a locked network cabinet and remains compatible with the preinstalled RFID EAC system.

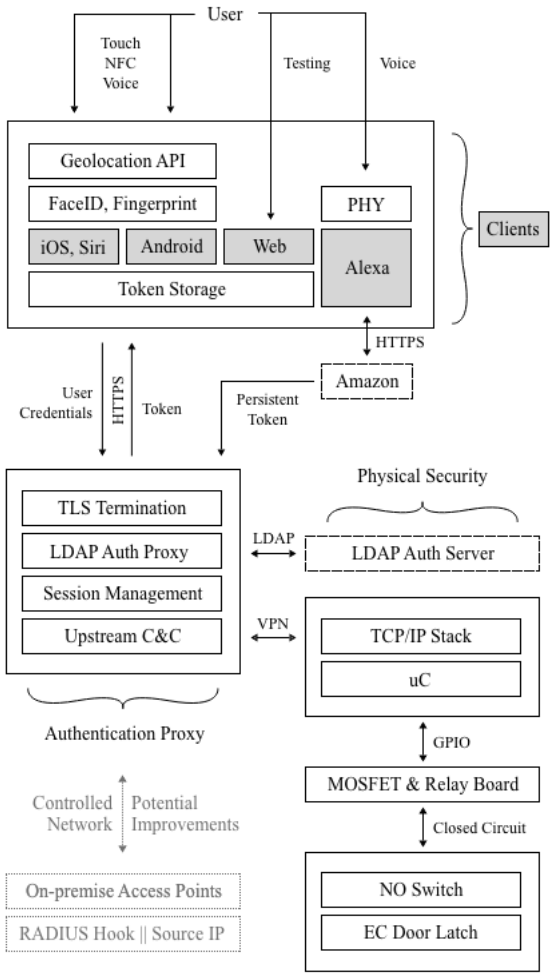


Figure 1: System architecture diagram depicting a typical request flow through different clients and various software and hardware components.

3 Interaction Design

During the mobile development, our efforts were aimed towards creating a very streamlined operation experience. Therefore, the application interface was designed to serve one core purpose – granting physical access in a rapid and secure manner. This approach severely aids the overall user experience and aims to position the touch-based smartphone unlocking mechanism near the existing RFID keycards in terms of convenience. Accordingly, two different interface design proposals were developed – one

employing a tap-based unlocking system and another relying on a swiping unlock gesture. Each proposal was implemented on one of the mobile platforms as a rudimentary approach towards A/B testing, neglecting the platform dissimilarities and relying on user familiarity instead. Selected views of mobile interfaces are depicted in Figure 2 and Figure 3.

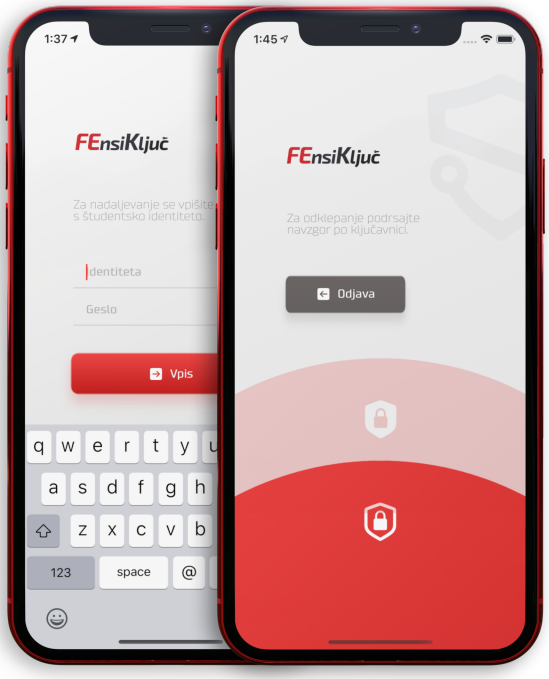


Figure 2: Mobile client interface on iOS operating system displaying a large surface area for a "swipe to unlock" gesture. The unlocking action also features audible and haptic feedback to aid user experience.

Despite different interface compositions, both applications follow the same crucial design patterns. They use the red color from the University of Ljubljana's emblem to establish a visually relating appearance and color-code their primary functions. The interfaces adapt to different screen sizes and resolutions to always position the primary action widgets within a natural resting position of a user's thumb. The geolocation-conditioned availability of login and unlock actions are color-coded along with matching symbols or text.

The iOS application, however, provides a much larger surface area for the touch-start event, making it easier to operate while walking as it does not require the user to keep eyes on the screen. It also provides audible and haptic feedback upon the successful unlock action, aiding the user in case of a silent latching mechanism. Contrary, its Android counterpart only offers feedback in form of a toast message but provides additional invocation mechanisms. Namely, it allows for direct invocation of the unlock intent from the system launcher menu and the immediate application exit upon the command execution. The latter nearly transforms the application into a home screen widget, allowing for a very transparent operation.

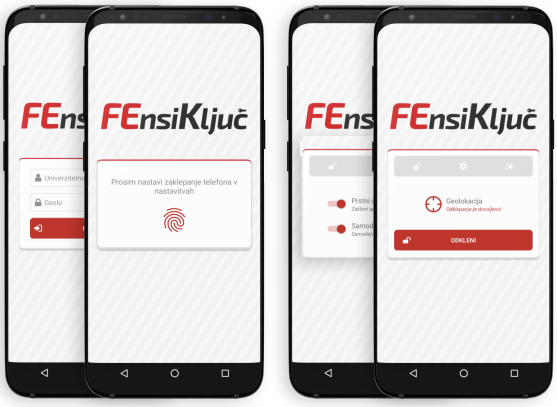


Figure 3: Android interface features a smaller primary action surface positioned within a natural resting position of a user's thumb. However, the application also offers several configuration options, including direct unlocking from home screen and fingerprint-based unlock protection.

Besides manually launching smartphone applications, their invocation using Siri voice commands on iOS and NFC detection on Android was also implemented. To enable the latter, a NFC tag with a password protected memory was placed outside of the office. Its scanning with an Android smartphone allows for activation of background unlock actions implemented in the mobile application.

Additionally, voice invocation of the unlock controller was realized using Amazon Alexa. This required a different approach, as Alexa only provides users with a voice-operated interface. Its invocation is therefore only possible from inside the office where the Echo Dot resides. In order to enable its smart home functionality an intermediary device emulation server was installed in local network, bridging the communication between Alexa and SBC. This allowed the use of natural language to control access to premises.

4 System Security

Our development approach considered the separation of concerns (SoC) principle in terms of software as well as hardware components. The modular design provided flexibility in terms of system installation and ensured the physical security of critical parts. Further, the authentication proxy server and the SBC both store private key files for OpenVPN authentication and TLS termination, as well as session and database information on the fully encrypted storage. This configuration prevents possible data exfiltration attacks on the physical media. Besides storage, all participating networks and interfaces are packet-filtered, rate-limited, and directly addressed by IP, evading potential DNS poisoning attacks. All system connections, login attempts, and executed mechanical actions are logged to provide an authorization history catalog.

Despite our efforts, several security weaknesses were identified and changes and updates for their mitigation were proposed. While DMA side channel and cold boot encryption attacks [4] on the authentication infrastructure are highly unlikely due to tight security measures, client terminal devices remain largely unprotected. Most notably, the system's geolocation-conditioned access is based on the information directly provided by the client devices. The latter should not be trusted as a reputable source of information, as they remain outside of our direct control. While such attack does not permit unauthorized access, mobile application users may modify their mobile systems to provide mock location data, falsely indicating their geographical position and allowing the premises to be unlocked remotely. Such settings are rather easily accessible in form of a software toggle on Android platforms, but may also be enabled on rooted (jailbroken) iOS devices or by using a compromised external GPS module or a GPS signal spoofer (e.g. using a software-defined radio - SDR) [5]. Moreover, mobile client security features such as fingerprint authentication and token storage all rely on their vendor implementations. In case the terminal device gets stolen, and an exploitable vulnerability exists, an attacker might be able to hijack active, time-constrained session keys.

Therefore, our security propositions include a migration to a network-based terminal localization method or source IP-based local access restrictions. On IP networks, these could range from simple firewall policies to RADIUS authentication hooks on eduroam networks, utilizing the same digital university identity and confirming the user presence on-site. However, all of the above options require direct control over the on-premise network, potentially sacrifice user convenience and possibly even introduce new attack vectors (e.g. a RADIUS evil twin attack).

Lastly, possible attacks on voice recognition hardware - Amazon's Echo Dot - were evaluated. While the device is securely installed inside the gated area, an attacker within a direct line of sight (e.g. from an elevated nearby building) could use a novel laser-based audio injection attacks on voice-controllable systems utilizing integrated MEMS (micro-electro-mechanical systems) microphones [6]. Although a cheaper (albeit more conspicuous) version of voice-based attack could also include a loud voice recording projected through door slit using a directional loudspeaker to activate the Echo Dot.

5 Conclusion

Electronic access control is undeniably a fundamental part of most commercial premises and smart homes. Its use has many advantages, such as revokable control and authorization log compared to the classic approach of sharing a common key. With the use of a wide range of standard options ranging from RFID installation to biometric scanners, additional methods, such as voice assistants and mobile and desktop applications enable users with new methods of authentication. However, each such novelty potentially instigates new security risks that have to be taken into consideration during the development and implementation these solutions.

Our solution addresses the problem of RFID key possession and enables context-based authentication using smartphone or voice-activation instead. Further, it simplifies privilege management operations by utilizing a pre-existing user database particularly suited for academic environment. Based on limited user feedback, these additions represent a significant improvement in terms of access convenience, as well as user experience by tackling repetitive daily tasks with gamification. However, the long-term reliability and security of our solution remain to be continuously evaluated. The first system assessment primarily indicated the requirement for transition to a network-based geolocalization, as well as inclusion of support for smart wearable devices to further aid user convenience.

References

[1] A. C. Caputo, "Digital video surveillance and security (second edition)," [Accessed 03.09.2020]. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124200425000113

[2] "Car connectivity consortium – digital key task group 2019," [Accessed 03.09.2020]. [Online]. Available: www.trustonic.com/events/car-connectivity-consortium-digital-key-task-group-2019

[3] W. Deutsch, "An introduction to electronic access control systems," [Online; accessed 6.7.2020]. [Online]. Available: https://www.thebalancesmb.com/introduction-to-electronic-access-control-394578

[4] J. Longo, E. De Mulder, D. Page, and M. Tunstall, "SoC it to EM: ElectroMagnetic side-channel attacks on a complex system-on-chip," pp. 620–640.

[5] N. O. Tippenhauer, C. Pöpper, K. B. Rasmussen, and S. Capkun, "On the requirements for successful GPS spoofing attacks," p. 75. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2046707.2046719

[6] T. Sugawara, D. Genkin, B. Cyr, S. Rampazzi, and K. Fu, "Light commands: Laser-based audio injection attacks on voice-controllable systems."

Zagotavljanje varnosti v celičnih omrežjih

Jernej Mušič, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: music.jernej@gmail.com

Ensuring security in cellular networks

Abstract. *This paper will discuss the properties of 2G and 4G cellular networks including the main components and authentication process during the primary tracking area update process. The authentication process walkthrough will show us the main security risks of previous generations of cellular networks and their negative influence on newer generations which inherit them. Inheriting these security risks show the negative side of providing backward compatibility to users which is one of the goals the provider wants to achieve during implementation. Security risks which are discussed in this paper will be listed and complimented with security solutions which have emerged during the years. Some of these are provided as integrated solutions where the whole cellular infrastructure and authentication is changed and is therefore in theory the responsibility of each provider to implement them.*

1 Uvod

S postopnim uvajanjem mobilnega omrežja pete generacije (5G) se soočamo s problematiko varnosti omrežja, predvsem z vidikov kakšne spremembe na tem področju prinaša nova generacija celičnega omrežja in ali bo ta podedoval ranljivosti prejšnjih generacij. V prispevku se osredotočimo na četrto generacijo (4G/LTE) celičnih omrežij ter na ranljivosti v protokolu, ki omogočajo potencialnemu napadalcu zlorabo celovitosti, integritete in dostopnosti omrežja.

Potrebno se je zavedati, da je LTE še vedno izpostavljen ranljivostim omrežjem predhodnih generacij. V nadaljevanju bomo za razumevanje omenjenih slabosti opisali protokol navezovanja in vanj umestili možnosti napadov. Od možnih napadov se bomo osredotočili tudi na uporabo naprav za prestrezanje celične komunikacije . Omenjene naprave namreč izkoriščajo ranljivosti protokolov omrežij prejšnje generacije, zato bomo v prispevku opisali tudi oblike napada le-teh.

2 Protokoli in gradniki v celičnem omrežju

Temeljni cilj ponudnikov ob vzpostavljanju novih generacij celičnega omrežja je skrb za nemoteno delovanje storitev, z namenom zagotavljanja kakovostne uporabniške izkušnje. Ob vzpostavljanju nove generacije celičnega omrežja se le to zaradi več razlogov vzpostavlja postopoma. V prvi vrsti se ob vzpostavitvi novih tehnologij ustvari cela veriga naprav, ki so zastarela oz. ne ustrezajo novemu standardu. Kot se je pokazalo že v preteklosti, je ob izgradnji nove generacije omrežja nujno potrebno stare tehnologije integrirati v novo generacijo. Z zagotavljanjem združljivosti s starejšo tehnologijo nova generacija podeduje tudi prednosti in slabosti prejšnje generacije, vključno s varnostnimi težavami. V tem prispevku bomo zato tudi obravnavali komponente in postopek preverjanja istovetnosti identitete v omrežju 2G in 4G, tipe najbolj pogostih napadov na omrežje 4G in varnostne izzive, ki jih je ta podedoval od svojih predhodnikov.

2.1 Gradniki omrežja 2G

Omrežje GSM je sestavljeno iz naslednjih osnovnih komponent, in sicer:

Mobilne postaje (ang. mobile station – MS) so kakršnekoli naprave, ki vsebujejo odstranljive Subscriber Identification Module (SIM) kartice [1]. V splošnem to velja za mobilne telefone in ostale naprave, ki omogočajo povezovanje z mobilnim omrežjem. MS-ju je dodeljen tudi unikatni identifikator (ang. International Mobile Equipment Identification – IMEI), ki v primeru odtujitve omogoča omrežju identifikacijo in zavračanje povezovanja naprave v omrežje [1].

Bazne postaje (ang. base stations – BS), ki direktno komunicirajo z MS in upravljajo celico v celičnem omrežju. Velikost celice je odvisna od geografskih značilnosti terena in gostote uporabnikov. Celice so manjše v gosto poseljenih območjih, kjer je večje število porabnikov. BS zagotavljajo tudi šifriranje in dešifriranje podatkovne komunikacije [1].

Krmilnik baznih postaj (ang. Base Station Controller – BSC) upravlja množico BS. Opravlja postopek predaje MS iz ene BS na drugo v primeru prehoda iz ene celice v drugo [1]. Njihova naloga je torej zagotavljati kakovosten prehod MS brez prekinitev v komunikaciji.

Preklopni center za mobilne usluge (ang. Mobile Services switching Center – MSC) upravlja mobilnost storitev in je zadolžen za usmerjanje telekomunikacijskih storitev ter komuniciranje z zunanjimi omrežji [1]. V njegovi sestavi so tudi naslednje podatkovne baze:

Register lokacij MS (ang. Home Location Register – HLR), vsebuje vse podatke o uporabnikih in njihovi dodeljeni mednarodni identiteti (ang. International Mobile Subscriber Identity – IMSI). Vsako omrežje ima samo en register lokacij MS [1].

Register lokacij gostujočih uporabnikov (ang. Visitor Location Register – VLR), je dodeljen posameznemu preklopnemu centru mobilnih uslug in

vsebuje podatke o uporabnikih, ki spadajo v območje odgovornosti posameznega MSC [1].

Register naprav (ang. Equipment Identity Register – EIR), vsebuje podatke o MS, ki nimajo dovoljenja za dostop do omrežja. V tem primeru gre za IMEI naprav, ki so bile prijavljene kot odtujene ali prepovedane [1].

Avtentikacijski center (ang. Authentication Center – AuC), vsebuje uporabniške podatke za dostop do omrežja, ki so tudi hranjeni v SIM, in sicer skriti ključ K_i[1, 2].

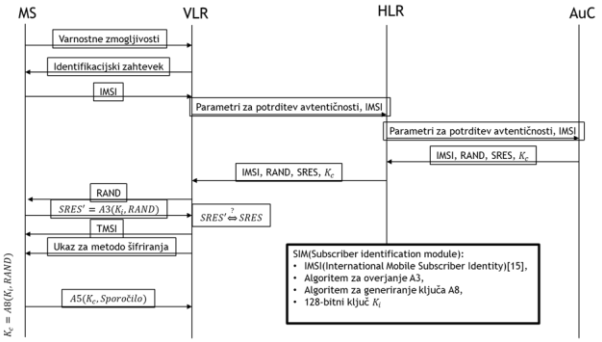
2.2 Postopek avtentikacije v omrežju 2G

MS se ob vklopu želi povezati z omrežjem, kar izvede z zahtevkom za povezavo na omrežje, del katerega je tudi nabor varnostnih zmogljivosti naprave. [1] V tem koraku je potrebno poudariti, da varnostne zmogljivosti MS zaostajajo za potrebami, še posebej na hitro rastočem in konkurenčnem trgu. Te razlike se sicer zmanjšujejo, vendar pa omogočanje uporabe predhodnih algoritmov za šifriranje kljub odkritim ranljivostim v omrežju predstavlja resen varnostni izziv. Vsaka MS ima vgrajen protokol A5, ki ga uporablja za šifriranje komunikacijskih podatkov. MS posreduje zahtevek za povezavo, ki vključuje varnostne zmogljivosti naprave, preko BS v VLR. Ta odgovori MS z zahtevkom po identificiranju s svojo številko IMSI, ki jo pošlje preko omrežja v odprti oz. nešifrirani obliki. VLR nato sestavi zahtevek za avtentikacijo uporabnika na podlagi prejete številke IMSI do registra lokacij mobilnih postaj (HLR), ki preveri uporabnika in od avtentikacijskega centra (AuC) zahteva generiranje naključnega 128-bitnega števila, imenovanega RAND. AuC nato na podlagi števila RAND in ključa K_i v algoritmu A3 izračuna 32-bitni odgovor imenovan SRES (ang. Signed RESponse). Ključ K_i in algoritma A3 in A8 se tudi nahajata v kartici SIM v MS. V naslednjem koraku AuC s pomočjo algoritma A8 in števila RAND izračuna 64 bitni sejni ključ K_C. SRES, RAND in K_C nato posreduje v VLR, ki do MS posreduje le RAND. S podatki, ki jih ima MS shranjene na SIM, izračuna svoj odziv SRES ter ga posreduje v VLR. Na tem mestu se obe različici SRES primerjata in v primeru ujemanja VLR dodeli MS številko TMSI (ang. Temporary Mobile Subscriber Identification) ter katero različico šifriranja A5 uporabiti. Kot je vidno na Slika 1, MS nato generira lasten sejni ključ K_C, ki ga uporablja v izbrani različici šifriranja A5. [1, 2] IMSI se torej izpostavi omrežju le med začetnim navezovanjem MS in po poteku veljavnosti TMSI, ki je namenjen onemogočanju sledenja uporabnikom storitev.

Simetrični tokovni šifrirni algoritem v rabi za šifriranje uporabniških komunikacijskih podatkov, se imenuje A5 in je vgrajen v vsak MS. Zasnova različic A5 je v osnovi enaka in se razlikuje le v generiranju psevdonaključnega 228-bitnega toka ključev. Za generiranje le-tega, algoritem prejme sejni ključ K_C, ki ga MS generira po uspešni avtentikaciji in števec okvirjev F_n. Ta je nato skupaj z 228-bitnim segmentom odprtih podatkov poslan skozi funkcijo izključujoči ALI (XOR), katere rezultat je 228-bitni tok šifriranih podatkov. Edina razlika med različnimi tipi A5 algoritma je način generiranja psevdonaključnih 228-bitov (GEN), medtem ko je XOR funkcija skupna. [3]

MS imajo na voljo naslednje različice A5 šifriranja:

A5/0, ki splošno predstavlja komunikacijo brez uporabe šifriranja se uporablja v državah, ki so pod sankcijami Združenih narodov in ostalih državah tretjega sveta. Uporaba A5/0 se v tem primeru ne zdi smiselna, saj je to eden od možnih vektorjev napada na omrežje. Njen namen je torej odpravljanje tehničnih težav ali preobremenjenosti v omrežju s strani ponudnika oz. upravitelja omrežja. [2, 3]



Slika 1: Postopek avtentikacije MS v omrežje 2G [1]

A5/1 generira tok ključev s pomočjo treh pomikalnih registrov z linearno povratno vezavo (LFSR), kjer ima vsak LFSR določene prioritetne bite, ki predstavljajo njihov status [3].

A5/2 je manj varna različica šifriranja, ki je bila izdelana zaradi lokalnih omejitev v nekaterih državah in velja za razbito in je zato njena raba tudi odsvetovana. [3]

A5/3 ali KATSUMI je različica s tokovnim šifrirnim algoritmom, ki temelji na blokovnem algoritmu MISTY. Vsebuje 64-bitne bloke, 128-bitni ključ in kompleksne rekurzivne Feistel-ove strukture z osmimi koraki. Znane so nekatere delne analize šifriranja A5/3, medtem ko algoritem do sedaj še ni bil razrešen v celoti ali vsaj ne v realnem času.[4]

2.3 Gradniki omrežja 4G

Struktura celičnega omrežja 4G se po osnovni zasnovi bistveno ne razlikuje od 2G. Pri 4G se nekateri gradniki združujejo, kar samo strukturo naredi manj kompleksno. Struktura omrežja 4G je sledeča:

Uporabniška oprema (ang. User Equipment – UE), predstavlja mobilno napravo, ki jo uporablja uporabnik za povezovanje z mobilnim omrežjem. Ta vsebuje univerzalno kartico z integriranim vezjem (ang. Iniversal Integrated Circuit Card – UICC) s pripadajočo programsko opremo in univerzalnim modulom z identiteto uporabnika (ang. Universal Subscriber Identification Module – USIM). USIM vsebuje enake podatke kot SIM pri omrežju 2G in sodeluje pri avtentikaciji UE ob vstopu v omrežje. Modul posreduje podatke o uporabniku, in sicer njegovo številko IMSI, skriti ključ in identifikacijsko številko mobilnega naročnika (ang. Mobile Subscriber Identification Number – MSIN). Iste podatke vsebuje tudi AuC v domačem omrežju za namene avtentikacije. [5]

Vozlišča (ang. evolved Node B – eNodeB), ki so temeljna komponenta razvitega univerzalno zemeljskega radijskega dostopovnega omrežja (ang.

evolved universal terrestrial radio access network – E-UTRAN), predstavljajo BS, ki omogočajo povezovanje MS z omrežjem. Z namenom zmanjšanja časovnega zamika in izboljšanja delovanja, so omenjena vozlišča med seboj tudi neposredno povezana. [5]

Entiteta za upravljanje mobilnosti (ang. Mobility Management Entity – MME) predstavlja glavno krmilno vozlišče omrežja, katerega naloge so vodenje avtentikacije uporabnika, upravljanje z nosilci, spremljanje lokacije uporabniške opreme in upravljanje prehodov za podatkovne pakete. [5]

Strežnik za upravljanje z uporabniki (ang. Home Subscriber Server – HSS), hrani in upravlja uporabniške podatke in skrite ključe. V primerjavi z gradniki omrežja 2G gre tukaj za združitev AuC s HLR, ki hranita in generirata ves potrebni šifrirni material za izvajanje avtentikacije in jih zagotavlja MME. [5]

Omrežni prehod paketnih podatkov (ang. Packet data network GateWay – P-GW) zagotavlja povezavo osrednjega omrežja z zunanjimi omrežji in zagotavlja uporabniški opremi IP-naslov. Poleg omenjenega je namenjen tudi za vodenje vrednosti porabljenih storitev in zagotavljanje izvajanja storitev skladno s politiko omrežja. [5]

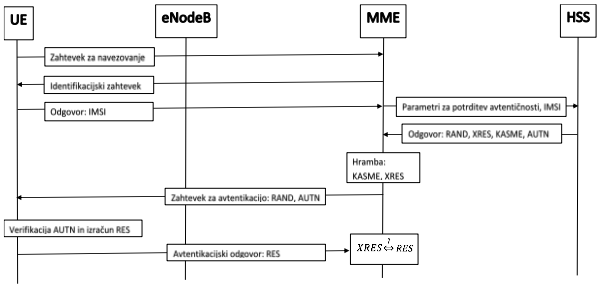
Strežniški prehod (ang. Serving GateWay – S-GW) zajame podatkovni nosilec in usmerja podatkovne paketke uporabniški opremi med proženjem MME. S-GW velja za vmesno opremo za izogibanje konstantnemu preusmerjanju podatkov na nivo P-GW ob vsakem premiku uporabniške opreme iz ene celice v drugo. [5]

2.4 Avtentikacijski postopek v omrežju 4G

Avtentikacijski postopek v omrežju 4G se deli glede na klasifikacijo naprav, ki se želijo povezati v omrežje. Te se delijo na zaupanja vredne in ne zaupanja vredne naprave, ki so ali niso del sporazuma partnerskega projekta tretje generacije (ang. 3rd Generation Partnership Project – 3GPP). V omrežju pa je operater tisti, ki odloča o tem, katera dostopna omrežja, ki niso del sporazuma so zaupanja vredna ali ne. [5]

Upravljanje dostopov s strani naprav, ki niso del sporazuma 3GPP, izvajata strežnik za avtentikacijo, avtorizacijo in obračunavanje (ang. Authentication, Authorization and Accounting – AAA) in prehod paketnih podatkov (ang. Evolved Packet Data Gateway – EPDG) [5].

Proces avtentikacije omrežij 3GPP, kot je vidno na Slika 2, se izvaja s pomočjo protokola Authentication and Key Agreement (EPS-AKA) med UE ter MME in deluje po principu medsebojne avtentikacije. [5] To pomeni, da omrežje avtenticira UE in obratno, kar je rešitev problematike v omrežjih 2G, kjer je ta proces le enosmeren. Rešitev je bila že del omrežja 3G z uvedbo protokola UMTS-AKA. Celoten postopek se začne z zahtevkom za navezovanje (ang. attach request), ki vsebuje IMSI ali GUTI (ang. Global Unique Temporary Identifier). Slednji ima enako funkcijo kot TMSI, in sicer preprečevanje pogostega izpostavljanja številke IMSI in njenega izpostavljanja potencialnemu sledenju uporabnika s strani nepooblaščene osebe. Na Slika 2 je viden postopek, kjer ima uporabniška oprema že dodeljeno GUTI, vendar je ta že potekla, kar pomeni, da je čas za dodelitev nove. MME v takšnem primeru pošlje UE zahtevek za identifikacijo in ta odgovori s svojo številko IMSI, ki je nato posredovana dalje v HSS. Ta v svoji podatkovni bazi poseduje podatke o uporabniku, in sicer njegov IMSI in SNid (ang. Serving Network identifier). S pomočjo slednjega strežnik za upravljanje z uporabniki generira ključ K_{ASME}, ki je specifičen za storitveno omrežje. HSS nato generira naključni izziv RAND ter ga skupaj z uporabniškim skritim ključem *K*, ki ga hrani v podatkovni bazi uporabi za generiranje avtentikacijskih vektorjev s pomočjo šifrirnih funkcij. Vektorji so sestavljeni iz RAND, XRES (odziv generiran s pomočjo skritega ključa *K in RAND*), K_{ASME} in AUTN (avtentikacijski žeton). Avtentikacijski vektor vsebuje tudi števec *SQN*, ki je namenjen preprečevanju ponovne rabe vektorjev, kar se dogaja med napadi s ponovno rabo (ang. replay attacks). Avtentikacijski vektor HSS nato posreduje MME, ki obdrži XRES in K_{ASME} ter posreduje AUTN in RAND uporabniški opremi UE. Ta iz parametra AUTN razbere vrednost SQN s pomočjo skritega ključa *K* in RAND. Nato UE generira vrednost XMAC, ki jo pridobi iz parametra SQN, RAND in AMF (del AUTN) ter jo primerja z vrednostjo MAC, ki je tudi del avtentikacijskega žetona. V naslednjem koraku preveri še vrednost parametra SQN in lastnega parametra SQN, katerih vrednost se ne sme razlikovati preveč. S tem je avtentikacija omrežja zaključena, in UE lahko generira ključ K_{ASME}, s čimer imata sedaj UE in omrežje enak ključ za vzpostavitev varne komunikacije. V odgovor omrežju UE generira tudi lastni odgovor RES, ki se nato na nivoju MME primerja z XRES in v primeru ujemanja zaključi postopek avtentikacije in dogovora o ključu. [5]



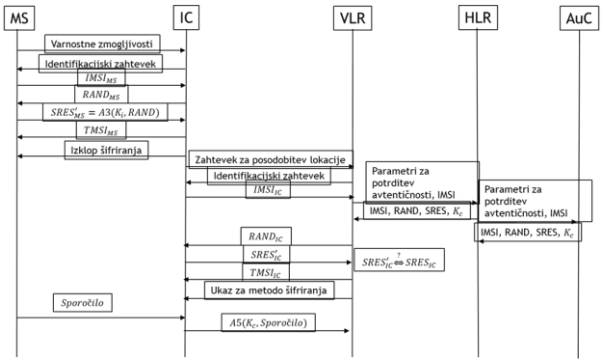
Slika 2: Avtentikacijski postopek v omrežju 4G

3 Ranljivosti v celičnem omrežju

Kot prikazuje **Error! Reference source not found.**, kjer je v postopku navezovanja v omrežju 2G prikazana metoda napada »man in the middle« (MitM), kjer naprava, imenovana lovilec IMSI (ang. IMSI catcher), prestreza paketke med MS in BS.

Lovilec IMSI enostavno spremeni vrednosti paketa, ki vsebuje varnostne zmogljivosti MS na vrednost, da MSni zmožna šifrirati komunikacije ali pa da je zmožna šifrirati samo z algoritmi, ki jih je možno s pomočjo mavričnih tabel dešifrirati v realnem času [1, 2]. Tovrstne naprave imajo za cilj znižanje varnostnega nivoja MS in s prestrezanjem paketkov z informacijami

za avtentikacijo prestrežejo informacije o uporabniku za lažje sledenje, brez da zmotijo postopek avtentikacije. [1, 2] Poznamo torej tri načine spremljanja celične komunikacije, in sicer pasivno, hibridno in aktivno obliko. Kot smo že omenili, je v nekaterih državah zaradi slabe infrastrukture celičnega omrežja in tehničnih težav na voljo le algoritem A5/0, ki pa je enostavno oznaka za to, da šifriranje ni v uporabi in so zato podatki v odprti obliki. Samo v teh primerih je možno povsem pasivno spremljanje mobilnih komunikacij, saj naprava potrebuje le pasivne komponente, in sicer prilagojene sprejemnike.



Slika 3: MITM napad značilen za lovilca IMSI

Na območjih, kjer ni omejitev in je raba šifrirnih algoritmov omogočena, je raba pasivne metode onemogočena. Tukaj pride v poštev hibridna in aktivna metoda, kjer obe vsebujeta tudi aktivno komponento, ki je namenjena posnemanju obnašanja MS ter BS. Temeljna razlika med hibridno in aktivno metodo spremljanja mobilnih komunikacij je v tem, da hibridna uporablja aktivno komponento le v procesu avtentikacije, kjer zagotovi znižanje šifriranja na nivo, katerega lahko v realnem času tudi dešifrira in s tem zagotovi pogoje pasivnega spremljanja. Aktivna metoda, ki je poimenovana tudi lažna BS (ang. fake BS) izvaja enake naloge, kot BS vendar v zmanjšanem obsegu, saj so takšne naprave mobilne in imajo zaradi tega tudi določene omejitve. [1, 2]

V celičnih omrežjih 3G in 4G so takšni napadi omejeni, saj kot smo videli uporabljajo vzajemno metodo avtentikacije, kjer se uporabniku identificira tudi samo omrežje in s tem v določeni meri prepreči vzpostavitev lažnih BS. Na podlagi virov [5, 6, 7] obstajajo v omrežju 4G določene oblike groženj, ki med drugim razkrivajo podatke o opremi v omrežju, pri čemer ni nobenega razloga, da bi bili znani uporabnikom. S pomočjo GPRS tunelskega protokola (ang. tunneling protocol – GTP) lahko uporabnik skenira omrežje in pridobi podatke o fizičnih napravah, ki so v rabi v omrežju, kar je enakovredno izvidovanju pred napadom. [7] Med možne napade spadajo tudi napadi za preprečevanje storitev (ang. denial of service – DoS), kjer z velikim številom sporočil create session v GTP protokolu lahko izkoristimo vsa razpoložljive vire in s tem onemogočimo rabo omrežja ostalim uporabnikom. [7, 8]

Pri možnost napada, ki je opisan v [8] lahko napadalec pridobi tudi telefonsko številko in IMSI uporabnika na omrežju 4G s tem, da mu moti možnost rabe omrežja, kar ima za posledico uporabo omrežja 2G. S tem lahko napadalec pridobi dovolj podatkov, da s svojo napravo oponaša napadeno in z iniciacijo klica na drugo napravo na tej pridobi tudi uporabnikovo telefonsko številko. Napad, opisan v [8], je tipični primer, kako novejše generacije celičnega omrežja zaradi povratne združljivosti lahko podedujejo varnostne luknje svojih predhodnikov, kljub temu da imajo bolj učinkovito varnostno politiko. Kot najbolj učinkovito obrambo pred temi napadi v [7, 8] navajajo opustitev zagotavljanja storitev omrežja 2G pod pogojem, da sta omrežji 3G in 4G v celoti zgrajeni in brez tehničnih pomanjkljivosti. Kot rešitev na varnostne grožnje, ki jih navajajo v viru [6] je predlagana raba asimetričnega šifrirnega algoritma in digitalnega podpisovanja, z namenom preprečevanja prestrezanja oz. zlorabe nekaterih podatkov, ki se med postopkom preverjanja istovetnosti prenašajo v odprti obliki. Vendar pa asimetrični šifrirni algoritmi zahtevajo več virov kot simetrični, ki so v trenutni rabi v omrežju 3G in 4G, kar pomeni dodatno obremenitev za fizične naprave v omrežju.

4 Zaključek

Celično omrežje 3G in 4G na nivoju varnosti zagotavlja velik napredek v primerjavi s svojim predhodnikom (2G) vendar pa ni brez pomanjkljivosti. Kot najbolj očitno grožnjo predstavlja prenos nekaterih občutljivih informacij v postopku navezovanja z omrežjem v odprti obliki. Velika slabost je tudi dedovanje varnostnih groženj s strani omrežja 2G, ki ga ponudniki na območjih kljub izgradnji omrežja 3G in 4G še vedno vzdržujejo. Ker smo trenutno v fazi izgradnje omrežja 5G, se postavlja vprašanje, ali bo omrežje 5G kljub vsem novostim tudi na področju varnosti ogroženo zaradi težav svojih predhodnikov. Treba se je zavedati kritičnosti obdobja med začetkom in koncem izgradnje 5G in pomena ukinitve prejšnjih generacij. Ranljivosti omrežja prejšnjih generacij bodo po izgradnji 5G in ukinitvi prejšnjih generacij postale brezpredmetne. To je le eden od predlogov, ki pa je ekstremno optimističen, saj bodo najverjetneje ponudniki storitev še nekaj časa po izgradnji 5G zagotavljali storitve omrežja 3G in 4G. Kot druga rešitev se ponuja vsaj ukinitev omrežja 2G in izvedba rešitev, ki temeljijo na asimetričnem šifriranju in uporaba digitalnega podpisovanja v postopkih avtentikacije.

Literatura

[1] J. Ooi, IMSI Catchers and Mobile Security, University of Pennsylvania, April 29 2015.

[2] S. Meyer, Breaking GSM with rainbow Tables, Marec 2010, arXiv:1107.1086

[3] O. D. Jensen, K. A. Andersen, A5 Encryption in GSM, Junij 2017

[4] O. Dunkelman, et al., A Practical-Time Attack on the A5/3 Cryptosystem Used in Third Generation GSM Telephony, 2010

[5] S. Behrad, E. Bertin, N. Crespi, Securing Authentication for Mobile Networks, A Survey on 4G issues and 5G answers, ICIN, Pariz, 19-22 Feb. 2018

[6] L. Qiang, et al., Security Analysis of TAU Procedure in LTE Network, Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2014

[7] S. Park, et al., Threats and countermeasures on a 4G Mobile Network, 8th Int. Conf. on Innovative Mobile and Internet Services in Ubiquitous Computing, 2014

[8] C. Yu , et al., LTE Phone Number Catcher: A Practical Attack against Mobile Privacy, College of Computer, Security and Communication Networks, 2019

Senzorski sistem za ocenjevanje agilnosti

Matevž Hribernik, Erik Keš, Anton Umek, Anton Kos

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, Ljubljana, Slovenija
E-pošta: matevz.hribernik@fe.uni-lj.si

Sensor system for assesment of agility

Abstract. *Measurement of athletes' physical abilities in different sport disciplines is one of the most important elements in training. The ability to change direction quickly called agility is just one of these abilities. Various agility tests are regularly included in sport test batteries for objective and wide-ranging athletes assessment. This paper presents a new approach to measuring agility using infrared optical gates and kinematic sensors. The combination of timing, position detection and motion recording enables the acquisition of previously inaccessible kinematic and temporal variables. The evaluation of the athletes' agility can be more objective and acquired with much greater certainty than it would be possible with traditional equipment and methods. The described sensor system can be used in various sport settings and tests, including the frequently used T-test presented in this paper.*

1 Uvod

Težnja po uporabi tehnologije v športu za izboljšanje športnikovih sposobnosti je prisotna skozi celotno zgodovino tekmovalnega športa. Nova znanstvena dognanja in tehnologije se usmerjajo v povečanje ciljnih športnikovih sposobnosti, te pa so večinoma vezane na posamezne športe. Nova dognanja in tehnologije pa niso vezane zgolj na profesionalne športnike in tekmovalni šport; enake ali podobne rešitve se uporabljajo pri rekreativnih športnih, spremljanju telesnega razvoja otrok in mladostnikov ter pri fizični rehabilitaciji. Uporaba različnih senzorjev in senzorskih sistemov profesionalnim športnikom omogoča pridobivanje tekmovalne prednosti. Športnikom, trenerjem in znanstvenikom s področja športa nove tehnologije ponujajo informacije, ki jih s tradicionalni orodji ni mogoče pridobiti ali pa jih lahko pridobimo z mnogo manjšo natančnostjo [1].

Tekmovalna uspešnost športnika je v veliki meri odvisna od učinkov dolgotrajnih treningov in telesne priprave, ki jih spremljamo z različnimi testi in meritvami. Agilnost je sposobnost hitrega spreminjanja smeri in je zgolj ena izmed pomembnih veščin, ki je pomembna v številnih skupinskih športih, kot sta košarka in nogomet, ali pri individualnih športih, kot so borilne veščine. Za oceno agilnosti uporabljajo različne teste, ki so usmerjeni na posamezne dele telesa in specifične komponente agilnosti. Mera učinkovitosti je običajno čas, ki je potreben za dokončanje zadane gibalne naloge. Čeprav je takšna informacija za ugotavljanje razlik med športniki dovolj, ne poda informacij o gibanja športnikov med testom.

Agilnost se navadno meri s pomočjo štoparice in izurjenega očesa trenerja ali pa s pomočjo modernih senzorskih sistemov za merjenje časa. Več raziskovalnih skupin je ugotovilo prednosti uporabe optičnih vrat v primerjavi s štoparico [2] oz. Global Navigation Satellite System (GNSS) [3]. Uporaba optičnih vrat za merjenje časa se šteje med najboljše načine merjenja agilnosti pri različnih skupinskih športnih in posamičnih [4], [5]. Senzorsko merjenje časa smo nadgradili z dodatnim merjenjem kinematike gibanja telesa ter tako razvili in izdelali vsestranski več senzorski sistem za merjenje agilnosti. Sistem je v osnovi zamišljen kot univerzalen, pri razvoju in preizkušanju pa smo kot šablonski primer uporabili T-test. Glavni elementi sistema so senzorji, senzorske naprave, sinhronizacijski in komunikacijski protokoli. Osnovne gradnike in programsko logiko je mogoče prilagoditi za uporabo pri različnih testih v športu. S predlaganim sistemom obravnavamo vprašanja natančnega merjenja časa v različnih fazah gibanja, sinhronizacijo številnih senzorskih naprav in prenos signalov in podatkov senzorjev. Poglobljena analiza uspešnosti športnika med testom je mogoča zgolj z združenimi in usklajenimi časovnimi, prostorskimi in kinematičnimi podatki; kar omogoča znatno izboljšanje kakovosti in uporabnosti informacij za trenerje in športnike.

Glavni prispevek našega članka je nov tehnološki pristop k merjenju agilnosti, ki združi kinematične, časovne in prostorske spremenljivke v časovno vrsto.

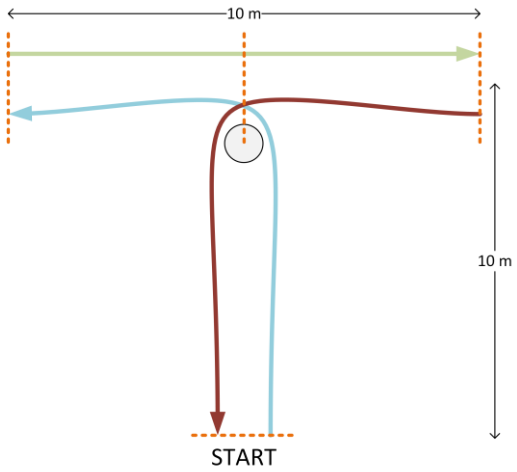
2 Ocenjevanje agilnosti

Agilnost je sposobnost hitrega in učinkovitega spreminjanja položaja telesa (ang. change of direction, COD). Povezana je s fizičnimi lastnostmi, kot sta moč in tehnika, ki jih je mogoče s treningom izboljšati, povezana pa je tudi s kognitivnimi komponentami, kot so tehnika in hitrost vizualnega skeniranja ter predvidevanje [6].

Agilnosti ni mogoče izmeriti z eno samo nalogo, ampak s kombinacijo zaporednih hitrih gibov. Agilnost se oceni z merjenjem časa potrebnega za dokončanje vrste vnaprej določenih nalog. Čeprav je naloga vnaprej določena, lahko vsak trener podaja različne ukaze in čas meri subjektivno. Zato je primerljivost meritev različnih trenerjev omejena.

Obstajajo različni standardni preskusi ocenjevanja agilnosti [7]. Preučili smo najpogosteje uporabljane teste: T-test, X-test, Illinois test in hiter tek na kratke proge. Vsak test je namenjen merjenju in poudarjanju določenih značilnosti športnikovega gibanja. Lahko rečemo, da vsak test ocenjuje različne vidike agilnosti.

Široka uporaba T-testa v različnih panogah je eden od razlogov za njegovo izbiro. T-test je v bistvu preprost, vendar dovolj zapleten za športnika. To nam omogoča načrtovanje in izdelavo gradnikov senzorskega sistema, ki bodo uporabni tudi pri drugih testih agilnosti. T-test, kot ga prikazuje slika 1, se izvede s tekom od začetne črte do sredinskega stožca. Po prečkanju osrednjega stožca, spremenimo smer gibanja in tečemo ali poskakujemo bočno v levo in nato v desno. Ko dosežemo desno stran, se vrnemo bočno do osrednjega stožca in nato vzvratno do začetne linije [8], [9].



Slika 1 Shema izvajanja T-testa. Puščice predstavljajo športnikovo gibanje, siv krog osrednji stožec, oranžne črtkane črte pa pozicijo optičnih vrat in talnih oznak.

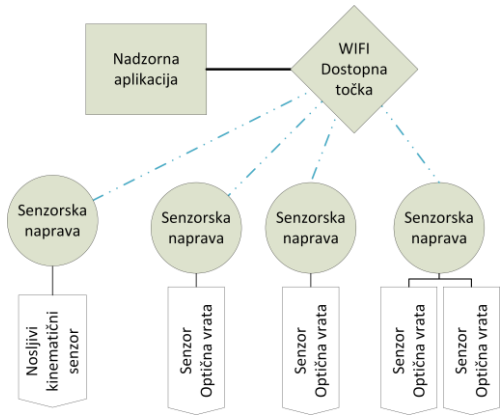
3 Senzorski sistem

3.1 Arhitektura

Arhitektura predlaganega merilnega sistema je prikazana na sliki 2. Vsebuje optična vrata za natančno merjenje časa športnika na določeni lokaciji in kinematični senzor gibanja za merjenje kinematičnih parametrov športnika. Za uspešno delovanje je potreben sinhronizacijski protokol, ki skrbi za časovno usklajenost notranjih ur vseh naprav. Potreben pa je tudi protokol za prenos podatkov preko brezžičnega omrežja in namenska kontrolna aplikacija.

3.2 Delovanje

Elementi sistema morajo biti povezani v isto lokalno omrežje. Brezžične senzorske naprave so lahko povezane na isto ali na različne dostopne točke. Optična vrata so fizično razporejena glede na potrebe testa. Nosljiva senzorska naprava s kinematičnim senzorjem je s pomočjo posebnega pasu pritrjena na ledveni del hrbta športnika, ki predstavlja približno pozicijo težišča človeka. Ko so vse komponente povezane v omrežje, operater sproži postopek sinhronizacije z uporabo nadzorne aplikacije na računalniku. Ko je sinhronizacija končana, se merjenje prične. Prejeti paketi se obdelujejo, zapisujejo v datoteko ter se uporabljajo za vizualizacijo.



Slika 2 Arhitektura senzorskega sistema. Naprave in senzorji so povezani z aplikacijo s pomočjo brezžične dostopne točke. Športnik ima nosljivi kinematični senzor nameščen na hrbtu, optična vrata pa so postavljena glede na potrebe testa. Vsaka naprava, lahko vsebuje ena ali več optičnih vrat.

3.3 Optična vrata

Pri izvajanju testov agilnosti nas zanimajo ekstremne točke gibanja, tam postavimo optična infrardeča vrata. Pri T-testu so to položaji: začetek, središče, levi in desni obrat. To nam omogoča, da določimo lokacijo, smer pristopa in potreben čas za izvedbo celotnega testa ali njegovih odsekov.

Ugotovili smo, da je za našo meritev najbolj primerna rešitev z uporabo oddajno-sprejemnega IR senzorja z odsevnikom, ki je sestavljen iz trirobnikov in deluje na potrebni razdalji 2 m. Izbrano integrirano vezje *IS471F* in primerna infrardeča LED (svetleča dioda) ustrezata našim potrebam. Komponente na vezju smo postavili tako, da sta LED in foto-senzor skupaj. Sprejemnik smo zaščitili pred neposrednim signalom iz oddajnika, ki predstavlja motnjo. Ker je oddajani signal moduliran, drugi IR viri kot so sijalke, predstavljajo manjši problem v smislu motenj, direktna sončna svetloba pa vseeno onemogoči delovanje senzorjev.

Senzor je povezan na *Adafruit Feather M0 WiFi z ATWINC1500*. Ker je izhod senzorja logična enica ali ničla, smo uporabili enega ali več digitalnih vhodov (odvisno od števila senzorjev na napravi). Mikrokrmilnik uporablja prekinitve za spremljanje spremembe stanja na vsakem vhodu. Ob sprožitvi prekinitve, program deluje v skladu s prenosnim protokolom.

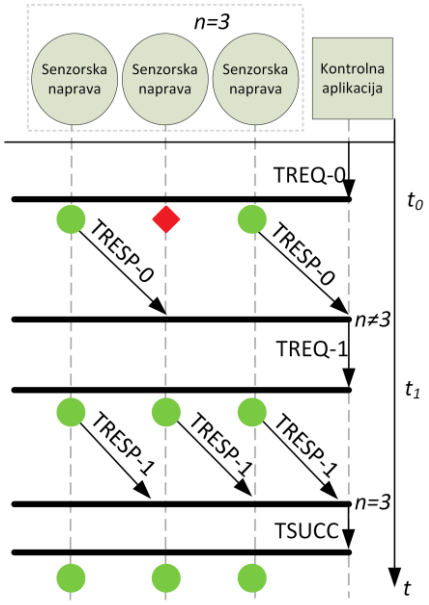
3.4 Nosljivi kinematični senzor

Nosljive naprave se pogosto uporabljajo v športu za merjenje gibanja. Pravilno lociran in pritrjeni senzor gibanja lahko uporabimo za merjenje kinematičnih spremenljivk [10]–[13]. V našem primeru parametre gibanja merimo z eno samo senzorsko napravo pritrjeno na ledveni del hrbta [12], [13] z vzorčevalno frekvenco 200 Hz. Uporabljamo kinematični senzor z vgrajenim pospeškometrom in žiroskopom (*LSM6DS33*). Ta senzor je prav tako pritrjen na *Adafruit Feather M0 WiFi z ATWINC1500*. Skupaj tvorita nosljivo senzorsko

napravo, ki jo napaja Li-ionska baterija. Senzor z mikrokrmilnikom komunicira preko protokola I^2C. Senzorski podatki se obdelujejo na mikrokrmilniku in preko brezžičnega omrežja posredujejo nadzorni aplikaciji na računalniku.

3.5 Sinhronizacijski in prenosni protokol

Časovno usklajeno delovanje sistema je osnovna zahteva za več senzorske sisteme, za kar je potrebna aktivna časovna sinhronizacija vseh senzorskih naprav.



Slika 3 Diagram sinhronizacijskega protokola. Aplikacija pošlje na vse naprave zahtevo po sinhronizaciji (TREQ-m) na naslov za razpršeno oddajanje. Naprave odgovorijo s sporočilom TRESP-m. Če strežnik ne sprejme vseh odgovorov s strani naprav, zahteva ponovno sinhronizacijo m+1. Ko vse naprave odgovorijo pritrdilno, aplikacija obvesti vse naprave o začetku merjenja s TSUCC sporočilom.

Zamislili smo si rešitev, ki deluje prek brezžičnih omrežij IEEE 802.11. Ne gre za optimalno rešitev, vendar smo tako združili že obstoječe merilne sisteme [14], [15] z novimi senzorji in optičnimi vrati. Časovni diagram sinhronizacije je prikazan na sliki 3. Sinhronizacijski protokol je relativno preprost in uporablja naslov razpršenega oddajanja (ang. broadcast) internetnega protokola (IP). Delovanje sinhronizacijskega protokola:

1. Kontrolna aplikacija ob času t_0 zahteva izvedbo sinhronizacije z oddajo paketa TREQ-m na naslov razpršenega oddajanja.
2. Ko naprave prejmejo zahtevo, nastavijo spremenljivko ure sinhronizacije na nič, prilagodijo lokalni števec sinhronizacij in strežniku vrnejo potrditveno sporočilo TRESP-m.
3. Strežnik zbira odgovore in če je število odgovorov enako številu naprav in so vrednosti števcev zaporednih sinhronizacij v vseh potrditvenih sporočilih enaki, je bila sinhronizacija uspešna. V nasprotnem primeru strežnik nadaljuje z zahtevami po sinhronizaciji, dokler vse naprave ne odgovorijo na isto zahtevo v določenem časovnem oknu.

Kot protokol transportnega sloja uporabljamo UDP (User Datagram Protocol). Nosljivi kinematični senzor pošlje 200 paketov na sekundo, v vsakem se nahaja en vzorec meritve. Optična vrata zaznavajo in pošiljajo samo spremembe stanja. Vsak paket nosi informacije o notranji in sinhronizirani uri, zaporedni številki vzorca ter ustrezne podatke. Vsi prejeti podatki so urejeni na skupni časovni osi s pomočjo nadzorne aplikacije glede na sinhronizirani sistemski čas.

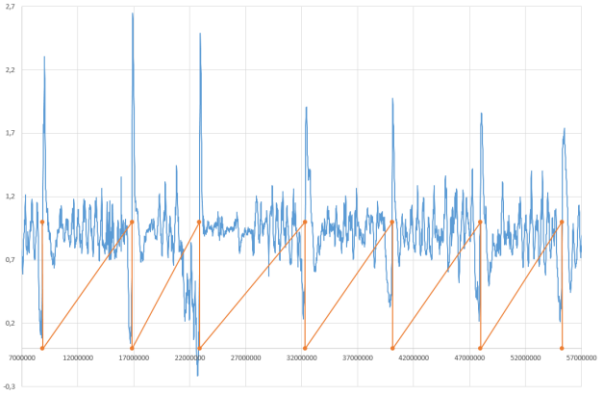
3.6 Aplikacija za nadzor sistema

Nadzor sistema se vrši v okolju LabVIEW v namenski aplikaciji, ki izvaja naslednje naloge: snemanje in analiza senzorskih signalov/podatkov, nadzor in spremljanje stanja sistema, časovna sinhronizacija naprav in inicializacija protokola za prenos podatkov. Operater sistema izvrši sinhronizacijo senzorskih naprav in začne meritev. Ko so podatki s senzorjev prejeti, se obdelajo in shranijo glede na sinhronizacijski čas.

4 Diskusija

S tem sistemom nestandardno opremo in protokole uporabljamo za meritve v realnem času. To pomeni, da je bilo potrebno sistem potrditi in testirati za delovanje v realnih pogojih. S testiranjem smo potrdili, da sistem deluje v okviru zahtev, ima pa tudi pomanjkljivosti.

Prvi problem je bila validacija senzorja optičnih vrat. Z laboratorijskim testom in ob upoštevanju specifikacije proizvajalca smo potrdili, da uspešno zaznamo prekinitve, ki trajajo več kot 1 ms. S tem smo potrdili, da je ločljivost senzorja zadostna za naša merjenja.



Slika 4 Rezultati testa sinhronizacije. Senzor gibanja (modro) smo zamahnili skozi optična vrata (oranžno). Opazimo, da se dogodki zgodijo hkrati. Vrata si sledijo kot v realnem testu: začetna, sredinska, leva, sredinska, desna, sredinska, začetna.

Večji izziv je predstavljal sinhronizacijski protokol. Pri testu sinhronizacije smo zato testirali zakasnitve oz. neskladnost sinhroniziranih ur med napravami. Protokol sam po sebi ima zakasnitve, ki jih zaradi uporabe brezžičnega omrežja IEEE 802.11 ne moremo natančno določiti. Zavedamo se, da vse naprave ne prejmejo zahteve po sinhronizaciji hkrati, in bi bila rešitev z uporabo drugačne brezžične tehnologije vsekakor koristna za sinhronizacijo. Ne moremo izmeriti časa potrebnega za sinhronizacijo, lahko pa izmerimo čas potreben za prejetje vseh odgovorov iz naprav in določili smo mejo, da zaupamo sinhronizacijam z

obhodnim časom manj kot 50 ms. Problem v našem primeru predstavlja neskladnosti notranjih ur, ki jih programsko ne popravljamo, zato sinhronizacijo izvedemo pred vsako meritvijo, a vseeno so rezultati meritev pokazali, da napaka za naše meritve ni usodna. Rezultati takšnega testa so vidni na sliki 4. Ob zamahu kinematičnega senzorja pred optičnimi vrati vrednost težnostnega pospeška pade in hkrati se sprožijo optična vrata. Natančnost je za naše potrebe zadostna. Uporaba brezžične rešitve, ki ne potrebuje sinhronizacije reši vse naše težave, vendar to zahteva spreminjanje celotnega merilnega sistema in osnovnih gradnikov, ki jih uporabljamo v drugih eksperimentih, čemur smo se želeli izognit.

5 Sklep

Predstavljeni senzorski sistem je univerzalen, saj ga je mogoče uporabiti pri različnih preizkusih agilnosti, prav tako pa je uporaben pri drugih testih v športu, kjer je potrebna lokalizacija in časovna sinhronizacija.

Predlagani sistem omogoča opravljanje mnogo bolj natančnih in objektivnih meritev agilnosti, saj vsebujejo spremenljivke in signale, ki tradicionalno niso del teh testov. Informacija o gibanju, vmesni rezultati ter povezavo med njimi je nemogoče pridobiti zgolj s štoparico in ekspertnim znanjem. Naša rešitev integrira obstoječe tehnologije in nudi nove informacije o športnikovem gibanju. Za optični senzor smo uporabili preprost, cenen integriran senzor, s čimer smo zmanjšali stroške in zapletenost sistema v primerjavi z drugimi rešitvami na trgu.

S sistemom smo opravili nekaj preliminarnih testov in potrdili, da izpolnjuje osnovne zahteve. V tem sistemu smo kombinirali nosljiv kinematični senzor z brezžičnimi optičnimi vrati in s tem pridobili informacije, ki jih je nemogoče zajeti s posameznim senzorjem.

Literatura

[1] K. Lightman, ‘Silicon gets sporty’, *IEEE Spectr.*, vol. 53, pp. 48–53, Mar. 2016, doi: 10.1109/MSPEC.2016.7420400.

[2] R. K. Hetzler, C. D. Stickley, K. M. Lundquist, and I. F. Kimura, ‘Reliability and Accuracy of Handheld Stopwatches Compared With Electronic Timing in Measuring Sprint Performance’, *J. Strength Cond. Res.*, vol. 22, no. 6, pp. 1969–1976, Nov. 2008, doi: 10.1519/JSC.0b013e318185f36c.

[3] M. Waldron, P. Worsfold, C. Twist, and K. Lamb, ‘Concurrent validity and test–retest reliability of a global positioning system (GPS) and timing gates to assess sprint performance variables’, *J. Sports Sci.*, vol. 29, no. 15, pp. 1613–1619, Dec. 2011, doi: 10.1080/02640414.2011.608703.

[4] R. Healy, M. Norris, I. C. Kenny, and A. J. Harrison, ‘A Novel Protocol to Measure Short Sprint Performance’, *Procedia Eng.*, vol. 147, pp. 706–711, Jan. 2016, doi: 10.1016/j.proeng.2016.06.252.

[5] N. Sánchez Aldana, J. Velásquez Gómez, J. Villa Bedoya, and J. M. Marín Correa, ‘SpeedMed: device for measuring velocity in track sports’, *Rev. Ing. Bioméd.*, vol. 1, no. 1, pp. 33–37, Jun. 2007.

[6] J. M. Sheppard and W. B. Young, ‘Agility literature review: classifications, training and testing’, *J. Sports Sci.*, vol. 24, no. 9, pp. 919–932, Sep. 2006, doi: 10.1080/02640410500457109.

[7] G. Sporis, I. Jukic, L. Milanovic, and V. Vucetic, ‘Reliability and Factorial Validity of Agility Tests for Soccer Players’, *J. Strength Cond. Res.*, vol. 24, no. 3, p. 679, Mar. 2010, doi: 10.1519/JSC.0b013e3181c4d324.

[8] K. Pauole, K. Madole, J. Garhammer, M. Lacourse, and R. Rozenek, ‘Reliability and Validity of the T-Test as a Measure of Agility, Leg Power, and Leg Speed in College-Aged Men and Women’, *J. Strength Cond. Res.*, vol. 14, no. 4, pp. 443–450, Nov. 2000.

[9] P. F. Stewart, A. N. Turner, and S. C. Miller, ‘Reliability, factorial validity, and interrelationships of five commonly used change of direction speed tests’, *Scand. J. Med. Sci. Sports*, vol. 24, no. 3, pp. 500–506, Jun. 2014, doi: 10.1111/sms.12019.

[10] H. Zeng and Y. Zhao, ‘Sensing Movement: Microsensors for Body Motion Measurement’, *Sensors*, vol. 11, no. 1, pp. 638–660, Jan. 2011, doi: 10.3390/s110100638.

[11] C.-C. Yang and Y.-L. Hsu, ‘A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring’, *Sensors*, vol. 10, no. 8, Art. no. 8, Aug. 2010, doi: 10.3390/s100807772.

[12] A. Kos and A. Umek, ‘Wearable Sensor Devices for Prevention and Rehabilitation in Healthcare: Swimming Exercise With Real-Time Therapist Feedback’, *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1331–1341, Apr. 2019, doi: 10.1109/JIOT.2018.2850664.

[13] A. Kos, S. Tomažič, and A. Umek, ‘Suitability of Smartphone Inertial Sensors for Real-Time Biofeedback Applications’, *Sensors*, vol. 16, no. 3, p. 301, Mar. 2016, doi: 10.3390/s16030301.

[14] A. Kos, V. Milutinović, and A. Umek, ‘Challenges in wireless communication for connected sensors and wearable devices used in sport biofeedback applications’, *Future Gener. Comput. Syst.*, vol. 92, pp. 582–592, Mar. 2019, doi: 10.1016/j.future.2018.03.032.

[15] A. Umek and A. Kos, ‘Smart equipment design challenges for real-time feedback support in sport’, *Facta Univ. Ser. Mech. Eng.*, vol. 16, no. 3, pp. 389–403, Dec. 2018, doi: 10.22190/FUME171121020U.

5G – ali res predstavlja tveganje za zdravje?

Tomi Mlinar[1]

[1]*Univerza v Ljubljani, Fakulteta za elektrotehniko, Katedra za informacijske in komunikacijske tehnologije, Laboratorij za sevanje in optiko, Tržaška cesta 25, 1000 Ljubljana*
tomi.mlinar@fe.uni-lj.si

5G – Is there a health risk?

Abstract. *This article discusses 5G technology and its impact on the environment and human health. The most important changes in the 5G radio interface that might have different impact on the environment as older technologies are the following: antenna beamforming and user tracking, massive MIMO, advanced channel coding, scalable OFDM-based 5G NR radio interface, and extended use of frequency spectrum. The second chapter discusses influence of electromagnetic fields on living organisms. The final conclusion is that 5G has no significant technical innovations compared to previous generations. The recommended limit values, which are in force from before, also apply to 5G technology. Procedures for the safe placement of 5G technology in the environment and methods for assessing the impact of electromagnetic fields are described in details in many documents and partially covered in this paper.*

1 Uvod

Tehnologija 5G je samo še korak od rabe v vsakdanjem življenju. V prvi fazi pričakujemo večjo uporabnost v industrijskih okoljih, kjer se trenutno tudi največ preizkuša, čez leto ali dve pa jo lahko pričakujemo tudi v naših domovih. Prve aplikacije bodo povezane z visokimi podatkovnimi hitrostmi, ki so nujne za storitve navidezne ali obogatene resničnosti in za grafično zahtevne igre, pa tudi za podatkovno nezahtevne storitve s področja interneta stvari, kot na primer krmiljenje in nadzor pametnega doma.

Približno leto po vpeljavi bo omrežje 5G delovalo v kombinaciji z obstoječim omrežjem 4G, šele pozneje pa tudi samostojno, dokler ne bo povsem nadomestilo prejšnje generacije omrežij. Standardizacija 5G poteka v vseh pomembnejših organizacijah (ITU-R, ECC, ETSI, 3GPP) in je že v sklepni fazi.

Omrežje 5G bo lahko delovalo na vseh obstoječih frekvenčnih območjih, ki jih že zdaj uporabljajo operaterji (0,8 GHz, 0,9 GHz, 1,8 GHz, 2,1 GHz in 2,6 GHz), nova frekvenčna območja, namenjena predvsem 5G, pa so 3,4–3,8 GHz in 24,25–27,5 GHz [1]. Širina frekvenčnega spektra na operaterja se povečuje - z zdajšnjih nekaj 10 MHz na prihodnjih nekaj 100 MHz. Pomembnejši parametri so zajeti v preglednici 1 [2]. Omrežja 5G uporabljajo napreden radijski vmesnik in drugačno arhitekturo jedrnega omrežja. Radijski vmesnik pri 5G je zelo učinkovit in deluje tako, da so oddajne moči minimalne, s tem pa tudi poraba energije.

Smernice mednarodnih in nacionalnih institucij za ravnanje z vplivi elektromagnetnih polj sistemov 5G na okolje in človeka se že uporabljajo. Standardizirane so tudi metode preskušanja in ocenjevanja vpliva neionizirnih sevanj na okolje in človeka.

Preglednica 1. Povečevanje zmogljivosti skozi generacije

	3G	**4G**	**5G**
Pasovna širina	2 MHz	200 MHz	> 1 GHz
Zakasnitve	100–500 ms	20–30 ms	<10 ms
Povprečna hitrost	0,144 Mb/s	25 Mb/s	200–400 Mb/s

2 Tehnološke novosti 5G

Pri tehnologiji 5G najdemo novosti na naslednjih področjih [3]:

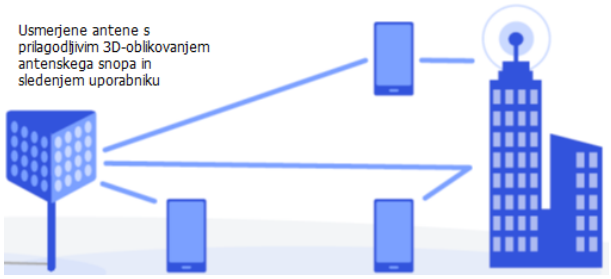
- nova arhitektura omrežja, ki so namenjena podpiranju raznovrstnih storitev,
- programsko nastavljivo omrežje z virtualiziranimi omrežnimi funkcijami,
- omrežje je razrezano na rezine, ki so prilagojene potrebam uporabnikov po hitrosti, pokrivanju in varnosti,
- ločevanje med uporabniško in kontrolno ravnino,
- nov radijski vmesnik (5G NR) in vpeljava majhnih celic.

Pri ocenjevanju vpliva tehnologije 5G na okolje se moramo najprej vprašati, kateri novi tehnološki dejavniki ali spremembe imajo vpliv na okolje. Izmed prej naštetih je to delno prenovljen radijski vmesnik, zato se mu v nadaljevanju posvečamo bolj podrobno. Pomembnejše spremembe v radijskem vmesniku 5G so naslednje:

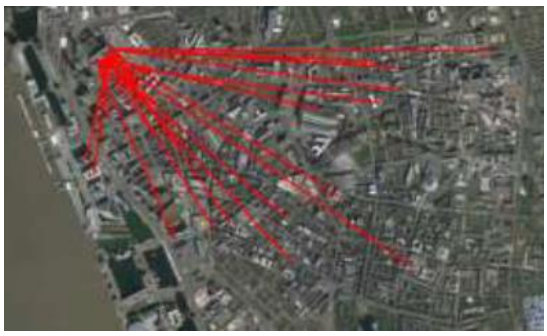
- oblikovanje antenskega snopa in sledenje uporabniku,
- uporaba masovnega MIMO,
- napredno kanalsko kodiranje,
- skalabilnost radijskega vmesnika 5G NR,
- razširjen frekvenčni spekter.

2.1 Oblikovanje antenskega snopa in sledenje uporabniku

Oblikovanje antenskega snopa (slika 1) je primerno za zgoščena omrežja, kjer se ista frekvenca ponavlja na razdaljah od 150 do 200 m [3]. Za to tehniko je značilno hitro usmerjanje antenskega snopa v smeri proti uporabniku (slika 2) in preklapljanje med snopi znotraj ene dostopovne točke. Posamezni antenski snopi so zelo ozki (2° in manj), kar pomeni, da je dobitek lahko tudi 50 dB (preglednica 2) [4]. Ta tehnika je predvsem namenjena za uporabo na frekvenčnem spektru pod 6 GHz, lahko pa se uporablja tudi na višjih frekvencah. Poraba energije je ciljna, s tem pa tudi manjša kot pri klasičnih antenah s širokim antenskim diagramom.



Slika 1. Antene s prilagodljivim 3D-oblikovanjem antenskega snopa



Slika 2. Uporaba zelo ozkih antenskih snopov, t. i. "pencil beams"

Preglednica 2. Širina posameznega snopa antene z oblikovanjem diagrama in njegov dobitek

Širina posameznega snopa	Dobitek
2°	38 dB
1°	43 dB
0,5°	50 dB

2.2 Masovni MIMO

Radijski vmesnik 5G (5G NR) je oblikovan za optimalno delovanje sistemov s časovnim dupleksom (TDD, Time Division Duplex). Uporablja referenčni signala (UL SRS, Uplink Sounding Reference Signal), da daje bazni postaji povratno informacijo o stanju zveze. 5G NR ima izboljšan mehanizem za poročanje o stanju v kanalu (CSI-RS, Channel State Information Reference Signal). Kodna knjiga z veliko prostorsko razločljivostjo podpira sestavo masovnega MIMO z do 256 anten. Poraba energije je ciljna, izgube v kanalu so manjše.

2.3 Napredno kanalsko kodiranje

5G uporablja kanalsko kodiranje ME-LDPC (Multi-Edge Low Density Parity Check). ME-LDPC ima do 5-krat večji dobitek v primerjavi s Turbo-kodami, ki se uporabljajo v prejšnjih generacijah omrežij. To velja predvsem pri obdelavi večjih blokov podatkov, zato je nova oblika kodiranja primerna predvsem za skupino storitev mobilnega širokopasovnega dostopa. Na sprejemni strani je omogočena vzporedna uporaba dekoderjev. Učinkovito kodiranje skrajša prenosni čas pri večji prepustnosti, kar se odraža v manjših zakasnitvah. Na porabo energije ima ta parameter manjši vpliv.

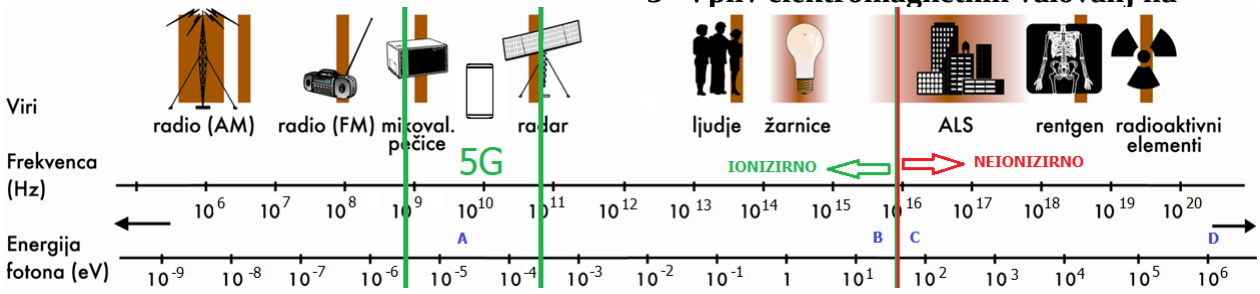
2.4 Skalabilnost radijskega vmesnika

Radijski vmesnik 5G NR je zasnovan tako, da ima prilagodljivo številčenje [3, 5], kar omogoča optimalno dodeljevanje radijskih virov in s tem optimalno porabo energije. Najmanjši razmik med podnosilniki (Δf) je 15 kHz. Razmik se prilagaja razpoložljivemu spektru, namenu uporabe (zunanje krovno pokrivanje, majhne celice, notranje širokopasovno pokrivanje) in vrsti storitve. Spreminja se po izrazu $\Delta f = 2^n \cdot 15$ kHz. Tako npr. uporabimo razmik 15 kHz na frekvenčnem območju 700 MHz, kjer so nosilniki široki 1, 5, 10 in 20 MHz, na frekvenčnem območju 28 GHz, kjer imamo na razpolago npr. 400 MHz spektra, pa izberemo razmik 120 kHz. Poznavanje načina delovanja OFDM je pomembno pri ocenjevanju obremenitve okolja.

2.5 Razširjen frekvenčni spekter

Po splošno sprejetem dogovoru se frekvenčni spekter za uporabo v 5G deli na dve skupini: FR1 in FR2 [1]. V skupino FR1 spadajo deli spektra pod 6 GHz (410–7125 MHz) in v skupino FR2 deli spektra med 24 in 60 GHz (24,25–52,60 GHz). Na frekvencah FR1 že delujejo obstoječe generacije mobilne telefonije (npr. LTE), frekvence iz skupine FR2 pa se uporabljajo npr. v radioastronomiji za daljinsko zaznavanje, v vojaške namene, kot orožje za onesposabljanje nasprotnikov, za skeniranje blaga in ljudi na letališčih, v policijskih radarjih za merjenje hitrosti, v industriji za natančno merjenje debeline materialov, v medicini za zdravljenja bolezni.

3 Vpliv elektromagnetnih valovanj na



Slika 3. Elektromagnetni spekter in energija [6]

3.1 Splošno o vplivih elektromagnetnih valovanj

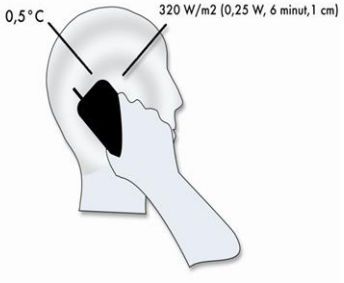
Pri vsaki radijski tehnologiji, pa tudi 5G, imamo dve vrsti vpliva elektromagnetnih valovanj (EMV) na okolico:

- vpliv na druge naprave in
- vpliv na žive organizme.

Poznamo termične in netermične učinke EMV na žive organizme. Večina raziskav in podlag za izdajo priporočil se ukvarja s termičnimi učinki električnih polj, magnetnih polj in elektromagnetnega valovanja na žive organizme [6].

Pri uporabi 5G raziskujemo vpliv EMV v frekvenčnem območju med približno 0,7 in 100 GHz. Na sliki 4 [6] vidimo, da ima EMV v tem delu spektra premalo energije, da bi povzročalo ionizacijo (izbijanje elektrona iz atoma), zato ga imenujemo neionizirno. Vezana energija elektrona v vodikovem atomu je 13,6 eV (slika 3, točka B), kar je meja za ionizacijo. Za primerjavo zapišimo, da je energija valovanja v spektru, ki ga uporablja 5G, okrog 0,000004 eV (slika 3, točka A), energija ultravijoličnega valovanja okrog 41 eV (slika 3, točka C) in energija valovanja pri jedrski eksploziji med 0,3 in 3 MeV (slika 3, točka D). Pri 5G torej nimamo enakih učinkov na organizme, kot jih povzročajo npr. radioaktivni elementi, računati pa moramo na t. i. termične učinke, ki pomenijo segrevanje organizma.

Vsak organizem je sestavljen iz pomembno velikih količin vode. Molekula vode (H_2O) se pod vplivom EMV začne vrteti in zato segrevati.



Slika 4. Segrevanje tkiva pri uporabi mobilnega terminala [6]

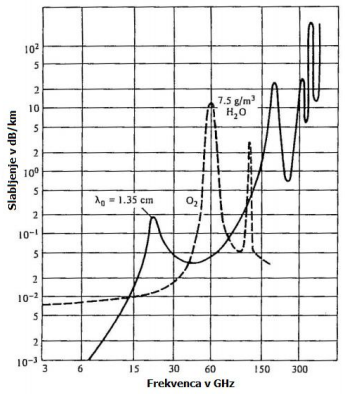
Na sliki 4 vidimo primer uporabe mobilnega terminala, ki ga držimo prislonjenega na uho, ki je približno 1 cm oddaljen od glave. Če se pogovarjamo 6 minut in terminal deluje z močjo četrt vata, se nam tkivo v okolici ušesa segreje za pol stopinje.

Preglednica 3. Vdorna globina EMV v tkivo

Frekvenca (GHz)	Relativna dielektričnost	Prevodnost (S/m)	Vdorna globina (mm)
6	36	4,0	8,10
10	33	7,9	3,90
30	18	27	0,92
60	10	40	0,49
100	7,3	46	0,35
300	5,0	55	0,23

Pri višjih frekvencah, ki jih uporablja 5G, se pojavi t. i. kožni pojav (ang. skin effect), kar pomeni, da se energija valovanja absorbira le na površini kože, v tkivo pa jo prodre zelo malo. V preglednici 3 so zbrane vdorne globine v kožo pri različnih frekvencah EMV. Višja kot je frekvenca, manjša je vdorna globina.

Pri višjih frekvencah mm-valov moramo računati še z zelo velikim slabljenjem zaradi padavin in vodne pare v ozračju [7]. Še posebno veliko je slabljenje na frekvenci 60 GHz (slika 5), kjer se signal oslabi tudi do 15 dB na kilometer, kar pomeni, da so te frekvence primerne le za komunikacijo na zelo kratke razdalje.



Slika 5. Slabljenje elektromagnetnega valovanja na višjih frekvencah zaradi padavin in vodne pare

3.2 Mejne vrednosti

Mednarodne organizacije, pa tudi Slovenija, posvečajo vprašanju vpliva EMV na okolje posebno pozornost že več desetletij. Javnost je postala bolj občutljiva za ta vprašanja od devetdesetih let prejšnjega stoletja, ko se je začela razvijati mobilna telefonija, z njo pa tudi množiti bazne postaje.

Po številnih raziskavah so organizacije, ki se ukvarjajo z vplivom EMV na okolje, že pred časom sprejele priporočene mejne vrednosti, ki se v skoraj treh desetletjih veljave niso spremenile. Nekaj od teh je zajetih v preglednici 4. Mednarodna komisija za zaščito pred neionizirnimi sevanji ICNIRP je lansko leto ponovno pregledala ta priporočila in objavila ugotovitve, iz katerih izhaja, da so mejne vrednosti, ki so bile postavljene leta 1998, ustrezne [8]. Tem priporočilom sledi tudi slovenska Uredba [9]. So pa v zadnjih letih nastali dokumenti, ki bolj natančno opisujejo postopke ugotavljanja obremenitve okolja z EMV in preventivne ukrepe, ki se jih morajo držati mobilni operaterji oziroma drugi postavljavci virov EMV. Nekaj od tega je zbrano v priporočilih ITU-T [10].

Preglednica 4. Mejne vrednosti EMV

	400 < f < 2000 MHz	f > 2000 MHz
ICNIRP 2020 "delavci"	S = f/40 W/m^2 E = = 3f$^{1/2}$ V/m (f v MHz)	S = 50 W/m^2 E = 137 V/m
ICNIRP 2020 "splošna populacija"	S = f/200 W/m^2 E = 1,37 f$^{1/2}$ V/m (f v MHz)	S=f/200 W/m^2 E=61 V/m
SLO UL 70/96 "I. območje"	S=f/2000 W/m^2 E=0,43f$^{1/2}$ V/m (f v MHz)	S=1 W/m^2 E=19 V/m
SLO UL 70/96 "II. območje"	S=f/200 W/m^2 E=1,37f$^{1/2}$ V/m (f v MHz)	S=10 W/m^2 E=61,4 V/m

3.3 Vpliv 5G na okolje

Za izračun vpliva tipične bazne postaje 5G na okolje vzamemo vhodne podatke, ki izhajajo iz prejšnjih poglavij. V priporočilih ITU [11, 12, 13] so zbrane enačbe in metodologija za ocenjevanje obremenitve okolice virov EMV.

Gostoto pretoka moči S izračunamo po enačbi (1).

$$S = \frac{P \cdot G_{\theta,\varphi}}{4\pi \cdot d^2} \quad \left[\frac{W}{m^2}\right] \tag{1}$$

kjer je *P* moč oddajnika, $G_{\Theta,\varphi}$ dobitek oddajne antene v smeri (Θ,φ) in d razdalja od oddajnika.

Če ocenjujemo gostoto pretoka moči v smeri največjega dobitka (kar vedno velja za antene z usmerjanjem snopa), potem velja enačba (2).

$$S = \frac{EIRP}{4\pi \cdot d^2} \quad \left[\frac{W}{m^2}\right] \tag{2}$$

kjer je *EIRP* efektivna izsevana moč oddajnika in so vključene tudi izgube na priključnih kablih in konektorjih.

Oceno, kakšna je gostota pretoka moči na razdalji d od antene, naredimo tako, da primerjamo izračunano ali izmerjeno vrednost z referenčno po enačbi (3).

$$\frac{S}{S_{ref}} < 1 \tag{3}$$

kjer je *S* izračunana/izmerjena vrednost in S_{ref} mejna vrednost iz priporočil.

Za primer 1 vzemimo *malo celico 5G*, ki deluje na frekvenčnem območju 28 GHz. Moč oddajnika P_{Tx} je 6 W (37,78 dBm), dobitek antene G_{Tx} je 7 dBi. Upoštevamo še slabljenje na priključnih kablih, ki je na teh frekvencah približno 1,5 dB/m.

Za primer 2 vzemimo *makro celico 5G*, ki deluje na frekvenčnem območju 3,5 GHz. Moč oddajnika P_{Tx} je 200 W (53 dBm), dobitek antene G_{Tx} je 25 dBi. Upoštevamo še slabljenje na priključnih kablih, ki je na teh frekvencah približno 0,15 dB/m.

Izračunane vrednosti primerjamo z mejnimi vrednostmi po priporočilih. Rezultati so zbrani v preglednici 5. V obeh primerih smo računali, da je antenski snop usmerjen točno proti uporabniku in da oddajnik deluje z vso razpoložljivo močjo.

Preglednica 5. Nekaj izračunov varne razdalje pri različni konfiguraciji baznih postaj 5G

	ICNIRP "delavci"	ICNIRP "splošna populacija"	SLO I.	SLO II.
	Varna razdalja od antene 5G			
Mala celica	0,1 m	0,2 m	0,6 m	0,2 m
Makro celica	9,0 m	20,1 m	63,7 m	20,1 m

Izračun je upošteval najbolj neugodne razmere. Natančne rezultate dobimo z meritvami, kjer moramo upoštevati priporočila, ki so specifična za vsako tehnologijo. Pri 5G se uporablja tehnika OFDM, ustrezne smernice, kako nastaviti spektralni analizator ali ustrezen dekoder, pa najdemo v [11, 12].

4 Sklep

Tehnologija 5G se bistveno ne razlikuje od 4G, vseeno pa vpeljuje nekatere tehnološke novosti. Uporaba višjih delov frekvenčnega spektra zahteva zgostitev mreže baznih postaj, njihovo približevanje uporabnikom in posledično zmanjševanje oddajnih moči. Mehanizem oblikovanja antenskega snopa pomembno zmanjšuje vpliv EMV na okolje. Napreden način "spanja" naprav v času manjšega prometa poleg manjše porabe energije vpliva tudi na manjšo obremenitev okolja. Predvidena je delitev infrastrukture med operaterji – vpliv na okolje v primeru delitve je manjši, kot je seštevek vplivov dveh ločenih infrastruktur. Neposredno po vpeljavi 5G se zaradi vzporednega delovanja več tehnologij predvideva nekoliko večji vpliv na okolje, pozneje pa se bo ustalil na zdajšnjih ravneh ali pa celo znižal. Za tehnologijo 5G veljajo isti standardi kot do zdaj – obstoječi mednarodni in nacionalni standardi niso odvisni od uporabljene tehnologije. Stopnja izpostavljenosti EMV se ne spreminja – oddajniki 5G, ki bodo postavljeni blizu uporabnikom, bodo imeli podobne lastnosti kot domači Wi-Fi. Postopki za varno umeščanje 5G v prostor in ocenjevanje vpliva EMV na okolje so postavljeni in zbrani v smernicah ITU [10] in drugih mednarodnih direktivah.

Literatura

[1] ETSI TS 138 104 V15.5.0, 5G; NR; Base Station (BS) radio transmission and reception, 2019-05

[2] Tomi Mlinar: 5G in okolje, Občni zbor Slovenskega društva za elektronske komunikacije, November 2019

[3] Making 5G NR a Commercial Reality, A unified, more capable 5G air interface, Qualcomm, March 2019

[4] Daniel Ephraty, 5G Fixed Wireless in mmWave, e-zbornik referatov, Seminar radijske komunikacije SRK 2020, Februar 2020

[5] Valentin Starciuc, NG Private Networks, e-zbornik referatov, Seminar radijske komunikacije SRK 2020, Februar 2020

[6] J. Budin, S. Eggert, L. Gregorač, L. von Klitzing, T. Mlinar, Elektromagnetna sevanja, Inštitut za telekomunikacije, 2004

[7] Recommendation ITU-R P.676-10, Attenuation by atmospheric gases, 09/2013

[8] ICNIRP guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz), Health phys 118(5): 483–524; 2020

[9] Uredba o elektromagnetnem sevanju v naravnem in življenjskem okolju, Uradni list RS št. 70/96

[10] ITU-T Recommendations on Human Exposure to Electromagnetic Fields: https://www.itu.int/net/ITU-T/lists/standards.aspx?Group=5&Domain=40

[11] Recommendation ITU-T K.100, Measurement of radio frequency electromagnetic fields to determine compliance with human exposure limits when a base station is put into service, 07/2019

[12] ITU-T K-series Recommendations, Suplement 9, 5G technology and human exposure to radio frequency electromagnetic fields, 05/2019

Modeliranje razširjanja radijskih valov v naravnih jamah

Andrej Hrovat[1], Michel D. Bedford[2], Tomaž Javornik[1]

[1] Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenija
[2] Camborne School of Mines, University of Exeter, Cornwall Campus, Penryn, CornwallUK
E-pošta: andrej.hrovat@ijs.si, M.D.Bedford@exeter.ac.uk, tomaz.javornik@ijs.si

Modeling of the radio waves propagation in natural caves

Abstract. *Paper discusses open problems in the modeling of the radio wave propagation in natural caves with rough walls, floor and ceiling and non-uniform cross section shape and dimension along the cave. The common problems which have to be solved in modeling process in empirical and deterministic channel models are: cave geometry estimation and simplification, estimation and inclusion of the wall roughness in the model and uncertainties in received signal characteristics measurements and their impact on channel model.*

1 Uvod

Razširjanje radijskega valovanja v cestnih in železniških predorih je opisano v številnih študijah za obstoječe in bodoče telekomunikacijske sisteme [1]. Cestni in železniški predori so pravilnih geometrijskih oblik. Stene, strop in tla so relativno gladki, zato ne povzročajo omembe vrednega sipanja radijskega valovanja. Dodatna prometna infrastruktura, kot so prometni znaki, pa ne pomenijo bistvenih ovir za širjenje radijskega valovanja. Za modeliranje razširjanja radijskega valovanja v predorih se uporabljajo naslednji pristopi: izračun nivoja radijskega signala z reševanjem Maxwellovih enačb, modalna analiza, kjer predor modeliramo z izgubnim valovodom, vrsta empiričnih modelov, kjer na podlagi meritev izračunamo potek slabljenja radijskega signala v praznem in/ali polnem predoru, in modeliranje širjenja radijskega valovanja z metodo sledenja radijskim žarkom, ki je še posebno primerna za modeliranje radijskih kanalov sodobnih širokopasovnih in več-antenskih sistemov [1]. Razširjanje radijskih valov v naravnih jamah, starejših predorih izdelanih ročno, kot so rudniški predori in predori za arheološka raziskovanja do sedaj še ni raziskano. Poznavanje lastnosti radijskega kanala v naravnih jamah in rudniških rovih postaja vse pomembnejše zaradi: *(i)* vpeljave številnih brezžičnih senzorjev v podzemno okolje, ki služijo za preverjanje trenutnega stanja v podzemnih jamah kot so temperatura, vlaga ter prisotnost zdravju škodljivih plinov, *(ii)* govornih in podatkovnih komunikacij v primeru nesreč in *(iii)* dostopnost multimedijskih vsebin v dobro obiskanih turističnih jamah.

Vendar se lastnosti naravnih jam bistveno razlikujejo od lastnosti predorov in prehodov zgrajenih na osnovi sodobnih gradbenih standardov. Glavne razlike so: *(i)* prečni presek naravnih jam ni enak po celotni dolžini jame ampak se nepredvidljivo spreminja vzdolž jame, (*ii)* naravne jame niso ravne ali rahlo ukrivljene temveč lahko vsebujejo precej ostrih nepravilnih zavojev, *(iii)* rovi se občasno končajo v velikih dvoranah, *(iv)* stičišča rovov so geometrijsko nepravilna, *(v)* v rovih se naključno pojavljajo vertikalne in horizontalne ovire kot so stebri in pragovi, in *(vi)* stene rovov niso ravne, temveč razgibane zaradi delovanje vode in kapnikov, ki pri tem lahko nastanejo. Zaradi omenjenih lastnosti naravnih jam je modeliranje razširjanja radijskega valovanja v naravnih jamah zelo zahtevno in predstavlja velik raziskovalni izziv.

V nadaljevanju bomo predstavili nekaj problemov pri modeliranju razširjanja radijskega valovanja v naravnih jama, na katere smo naleteli pri moduliranju razširjanja radijskega valovanja v treh predorih dveh naravnih jamah, ki se nahajajo v naravnem parku Yorkshire Dales Veliki Britaniji in sicer, predora Roof Tunnel in the Milky Way Inlet v jami Kingsdale Master in predor v jami Skirwith. Posvetili se bomo glavnim elementom, ki jih potrebujemo za modeliranje razširjanja radijskega valovanje in sicer *(i)* opis okolja v katerem se radijsko valovanje razširja, *(ii)* meritev radijskega signala, *(iii)* določitev modelov razširjanja radijskega valovanja in *(iv)* prilagoditev parametrov modela glede na izmerjene vrednosti nivoja radijskega signala.

Članek vsebuje naslednja poglavja. V prvem poglavju po uvodu opišemo metodo za izdelavo opisa geometrije podzemne jame, nato nadaljujemo s problemi pri meritvah moči sprejetega signala. V četrtem poglavju opišemo dva modela razširjanja radijskega valovanja v naravnih jamah in njihove omejitve. Sledi še zaključek s sklepnimi ugotovitvami in predlogi za nadaljnje delo.

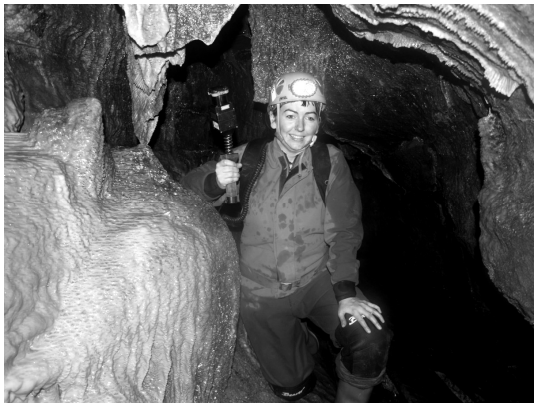
2 Geometrijski opis naravne jame

Geometrijski opis okolja v katerem se širi radijsko valovanje je izrednega pomena za modeliranje širjenja radijskih valov. Digitalni model višin in raba tal sta neobhodna za modele širjenja radijskega valovanje na neposeljenih in vaških okoljih, mesta lahko opišemo s shape-datotekami ali digitalnim modelom stavb, za notranjost stavb pa uporabljamo vrsto vektorskih modelov s katerimi opišemo stene in odprtine v njih kot so okna, vrata, itd. Omenjeni podatki že obstajajo, pridobljeni pa so bili s satelitskim ali letalskim pregledovanjem zemljine površine, ali geodetskimi meritvam stavb, oziroma iz načrtov stavb. Geometrijski opisi podzemnih jam ne obstajajo, zato jih moramo

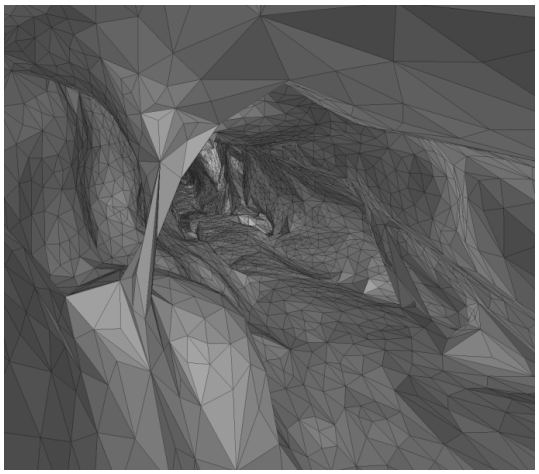
pridobiti sami s 3D laserskim skeniranjem. Laserski skenerji so običajno pritrjeni na trinožna stojala ali na vozilo. Ker tega pristopa v podzemnih jamah ne moremo uporabiti, lahko uporabimo prenosni ročni skener. V primeru pregledovanja naravnih jam smo uporabili skener ZEB1 podjetja GeoSlam, ki je komercialni produkt temelječ na znanstvenih dognanjih [2]. Skener je pritrjen na palico in omogoča pregled okolice na vseh 360 stopinjah. Vsebuje tudi vrsto senzorjev za premikanje v vseh smereh, katerih podatki se uporabijo za zlivanje dveh ali več zaporednih slik. Izhodna datoteka skenerja ZEB1 vsebuje oblak 3D točk, ki se jih z namensko programsko opremo GeoSlam Hub pretvorimo v datoteko standardnega trikotniškega formata (STL). Po podatkih proizvajalca je natančnost skeniranja 2 do 3cm.

Zaradi razgibanosti podzemnih jam so trikotniki, ki opisujejo geometrijo jame razmeroma majhni, zato je število trikotnikov za opis jame precej veliko, kar lahko pri nekaterih pristopih za modeliranje razširjanja radijskih žarkov, na primer z metodo sledenja radijskim žarkom, močno podaljša čas računanja. Število trikotnikov lahko zmanjšamo z združevanjem trikotnikov. Pri tem izgubimo podrobnosti o geometriji jame, še posebno, če uporabimo standardne metode, ki novi večji trikotnik izračunajo tako, da minimizirajo srednjo kvadratno napako med novim združenim trikotnikom in množico manjših trikotnikov. Bolj verno geometrijsko obliko jame dobimo, če poiščemo ekstreme odstopanja ter na osnovi njih zgradimo nov geometrijski model jame. Pri uporabi metode sledenja radijskim žarkom moramo upoštevati, da so trikotniki dovolj veliki, da odbijejo večji del energije radijskega valovanja. Za dober geometrijski model podzemne jame moramo določiti tudi elektromagnetne lastnosti materiala iz katerega je jama zgrajena.

Slika 1 prikazuje postopek skeniranja podzemne jame, Slika 2 pa rezultate skeniranja v obliki STL.



Slika 1. Postopek skeniranja podzemne jame



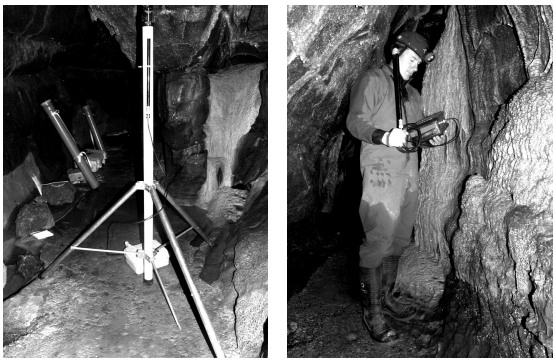
Slika 2. Geometrijski 3D model podzemne jame.

3 Merjenje lastnosti radijskega signala

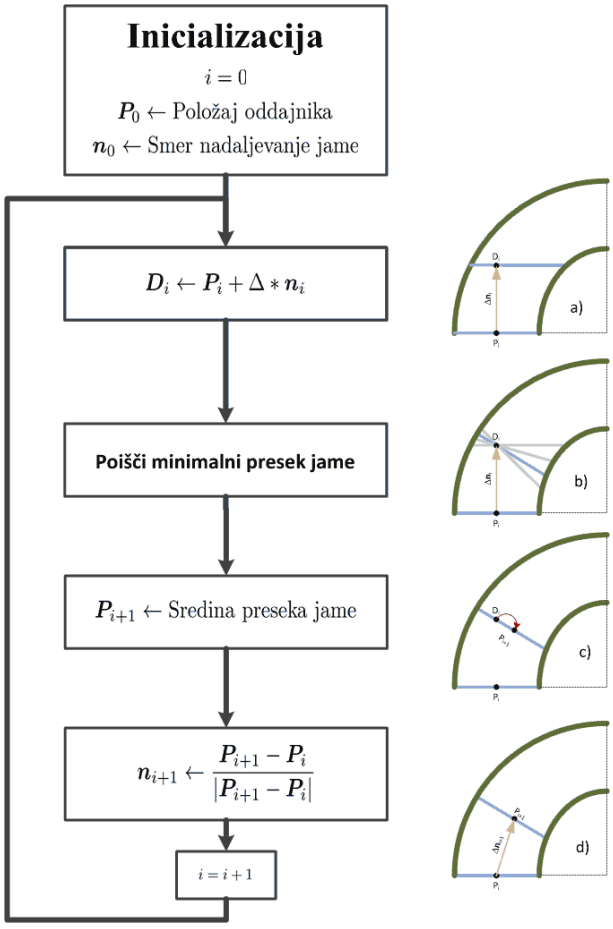
Dober model širjenja radijskega valovanja zahteva umerjanje modela, oziroma preverjanje veljavnosti modela z izmerjenimi rezultati. Vsaka meritev mora poleg radijskih parametrov, na primer nivoja sprejetega signala ali impulznega odziva radijskega kanala, vsebovati tudi položaj oddajnika in sprejemnika, oboje zapisano v lokalnem ali globalnem koordinatnem sistemu. Za določanje lokacije oddajnika in sprejemnika zunaj stavb običajno uporabimo kar globalne satelitske sisteme za določanje položaja, kot sta GPS in GLONASS in umerjen tlačni višinomer za določanje višine oddajnika in sprejemnika. V notranjosti stavb pa položaje oddajnika/sprejemnika vpišemo ročno ali uporabimo metode določanje položaja oziroma sledenja na osnovi sidrnih vozlišč, katerih lokacija je poznana. Višino oddajnika/sprejemnika pa se dokaj enostavno določi z meritvijo razdalje od tal.

V podzemnih jamah so tla razgibana in neravna, mnogokrat nagnjena, prehodi pa ozki in nizki, tako da je edini način, da si zagotovimo dovolj veliko število meritev vzdolž jame ročno prenašanje oddajnika/sprejemnika od točke do točke. Uporaba namenske geodetske opreme za določanje položaja sprejemnika bi bila precej zamudna, zato se običajno uporabljajo nenatančne metode, ki lahko vnašajo v model precej velike napake. Primer izvajanja meritev v jami prikazuje Slika 3. Razdalje med posameznimi meritvami običajno ni težko izmeriti, za kar potrebujemo običajni meter ali laserski merilnik. Obstaja pa dokaj velika negotovost kolikšna sta azimut in elevacijski kot naslednje meritve, kar direktno vpliva na višino sprejemnika in na njegov horizontalni položaj v jami. Za meritve v jamah v Yorkshire smo dobili podatek, da so oddaljene za cca 1 m in da so se meritve izvajale na višini cca 1.5 m, če je bila višina jame stojna, drugače pa na sredini jame. Horizontalni položaj meritve pa je bil vedno v sredini jame. Da bi ocenili položaj sprejemnika v jami, smo načrtali in izdelali algoritem za izračun položaja sprejemnika, ki temelji na

predpostavki, da se merilec giblje v smeri poteka jame, ter se vedno postavi v središče minimalnega prečnega preseka.

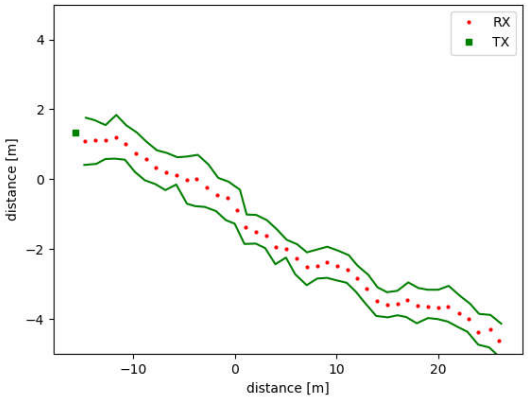


Slika 3. Prikaz izvajana meritev v jami



Slika 4. Blok diagram iskanja poti v jami

Blok diagram algoritma, prikazuje Slika 4. Za začetno točko izberemo položaj oddajnika, smer gibanja $\mathbf{n}_0$ pa predpostavimo v smeri nadaljevanje jame. Korak Δ izberemo tako velik, da lahko sledimo poteku jame. Nato se premaknemo v smeri jame $\mathbf{n}_i$ za korak Δ v točko **D,** okoli katere vrtimo presečno ravnino toliko časa, da dobimo minimalno vertikalno presečno površino jame. Točka na sredini preseka določa naslednjo točko poti $\mathbf{P}_{i+1}$. Smer nadaljevanje jame pa določa smer glede na prejšnjo točko $\mathbf{n}_{i+1}$. Korak Δ izberemo mnogokrat manjši od 1 m, na katerem so bile izvedene meritve. Ker so konfiguracije naravnih jam precej razgibane, pot v jami na koncu filtriramo z drsnim povprečjem. Pot v jami in obrise jame prikazuje Slika 5



Slika 5. Pot meritev jami in stene jame

Negotovost meritev poveča tudi položaj antene in vpliv človeka (merilca) na nivo radijskega signal. Dodatne merilne napake vnašamo z ročnim izvajanjem meritev in položaja antene glede na človeka, ki je odvisen od višine jame na mestu meritve. Slabljenje zaradi prisotnosti človeka znaša med 0 dB in 6 dB, odvisno od tipa sprejemnika in frekvence.

4 Modeli izgube poti v naravnih jamah

Splošni empirični model izgube poti za naravne jame ne obstaja, ker so geometrije in sestave naravnih jam zelo različne. Posledično bi morali za vsako jamo ali značilni tip jame izvesti nov sklop meritev. V [3] so avtorji predlagali empirični model izgube poti, ki temelji na vrsti meritev v naravnih jama na področju Yorkshire. V modelu je podzemna jama razdeljena na sektorje. V sektorju so lastnosti jame, kot so oblika, hrapavost, velikost, nagib, itd, podobne. Izguba poti v posameznem sektorju se izračuna z valovodnim modelom. Slabljenje na meter dolžine za horizontalne in vertikalne valovodne načine se izračuna po enačbah (1,2),

$$L_{mn}^{(h)} = \frac{\lambda^2}{2}\left(\frac{\varepsilon_r^{(h)}(m+1)^2}{w^3\sqrt{\varepsilon_r^{(h)}-1}} + \frac{(n+1)^2}{h^3\sqrt{\varepsilon_r^{(v)}-1}}\right), \quad (1)$$

$$L_{mn}^{(v)} = \frac{\lambda^2}{2}\left(\frac{\varepsilon_r^{(v)}(n+1)^2}{h^3\sqrt{\varepsilon_r^{(v)}-1}} + \frac{(m+1)^2}{w^3\sqrt{\varepsilon_r^{(h)}-1}}\right), \quad (2)$$

kjer sta n in m indeksa valovodnega načina, w in h sta širina in višina valovoda, medtem ko sta $\varepsilon^{(h)}{}_r$ in $\varepsilon^{(v)}{}_r$ relativni dielektrični konstanti tal in sten. Ker stene, tla in strop jame niso ravne, model predvideva še dodatno slabljenje zaradi hrapavosti (3) in slabljenje zaradi valovitosti tal (4), kjer je Δh koren srednje kvadratne hrapavosti in θ koren srednje kvadratne valovitosti tal. Poleg tega model predvideva še dodatno slabljenje v primeru zavoja v jamah in za stičišče dveh ali več rovov.

$$L_{hr} = \frac{\pi^2 \Delta h^2 \lambda}{2}\left(\frac{1}{w^4}+\frac{1}{h^4}\right), \quad (3)$$

$$L_{nagib} = \frac{\pi^2 \theta^2}{2\lambda}. \quad (4)$$

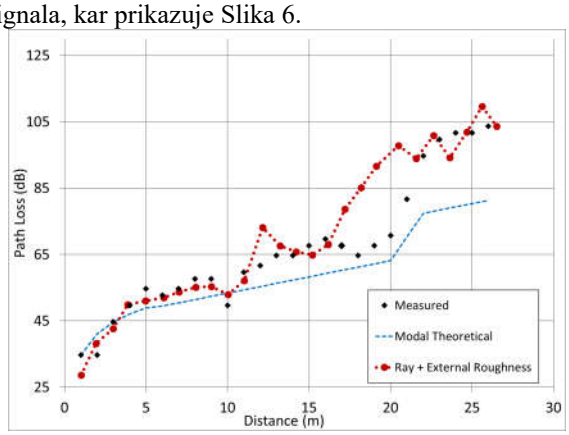
Dodatno se prvi sektor pri oddajniku razdeli na dva podsektorja. Tik ob oddajniku nivo radijskega signala pada kot v praznem prostoru, medtem ko v drugem podsektorju slabljene signala sledi valovodnemu modelu. Meja med prvim in drugim sektorjem se izračuna po enačbi (5), kjer je d_b točka preloma:

$$d_b = \min\left(\frac{w^2}{\lambda}, \frac{h^2}{\lambda}\right). \quad (5)$$

Drugi način modeliranja izgube poti v naravnih jamah je deterministične z metodo sledenja radijskim žarkom. Možnost uporabe metode sledenja radijskim žarkom se je pojavila, ko so na trg prišli prenosni infrardeči skenerji, s katerimi lahko dobro opišemo geometrijo podzemne jame z razmeroma velikim številom povezanih trikotnikov. Za modeliranje razširjanja radijskega valovanja v jami smo uporabili komercialni programski paket Wireless InSite podjetja REMCOM, ki omogoča sledenje radijskim žarkom v treh dimenzijah, upošteva hrapavost površine, uklone na ovirah in odboje na ravnih ploskvah. Program uporablja metodo SBR (Shooting-and-Bouncing-Rays, streljanje in odbijanje žarkov). Zaradi hitrosti izračuna je število interakcij žarka z okolico (število odbojev, uklonov, lomov) omejeno. Pri primerjavi izmerjenih in modeliranih rezultatov z metodo sledenja radijskim žarkom ter empirično metodo, smo ugotovili, da moramo izmerjeni geometrijski model naravne jame ustrezno prilagodi nosilni frekvenci radijskega signala in ustrezno umeriti programski paket. Pri oceni izgube na poti z metodo sledenja radijskim žarkom in primerjavo izmerjene izgube poti z izračunano smo naleteli na naslednje ovire:

- v geometrijskem modelu naravne jame se pojavljajo napake, kot so manjkajoče ploskve, ploskve se popolnoma ne stikajo, kar povzroči uhajanje radijskega žarka iz naravne jame,
- ne poznamo dielektrične konstante tal stropa in sten jame,
- ploskve v modelu naravne jame morajo biti dovolj velike, da omogočijo odboj večino energije vpadnega radijskega valovanje,
- standardne metode za združevanje ploskev niso primerne, ker temeljijo na minimalni srednji kvadratni napaki, medtem ko so za izračun nivoja radijskega valovanja pomembne maksimalne vrednosti izboklin in vdolbin,
- metoda določanja robov na katerih se radijski žarki uklanjajo ni jasno določen, kar vodi v premalo ali preveč uklonjenih žarkov,
- model upoštevanja hrapavosti površine, ki ga uporablja komercialno orodje, je model za urbana okolja, torej za stavbe z ravnimi zidovi in nekaj izboklinami,
- meritve v naravnih jamah je težko izvesti, predvsem pa je težko določiti mesto meritve, višino meritve, lego antene in vpliv človeka na izmerjene rezultate,
- zaradi širjenja radijskega signala po več poteh oziroma presiha radijskega signala in ozkopasovnih meritev, so meritve zelo občutljive na minimalne premike sprejemne antene.

Kljub omenjenim pomanjkljivostim se z ustreznim umerjanjem metode sledenja radijskim žarkom dokaj dobro približamo izmerjenim nivojem radijskega signala, kar prikazuje Slika 6.



Slika 6. Primerjava izmerjenih vrednosti izgube na poti, empiričnega modela in metode slednja radijskim žarkom.

5 Zaključek

Ocena izgube na poti v naravnih jamah z metodo sledenja radijskim žarkom dokaj dobra opiše slabljenje radijskega valovanja, vendar pa zahteva ustrezno umerjanje same metod sledenja radijskim žarkom in pa geometrijskega modela naravne jame. Z bolj natančnim model upoštevanja hrapavosti sten in valovitosti tal bodo rezultati tudi bistveno bolj natančni.

Literatura

[1] A. Hrovat, G. Kandus and T. Javornik, "A Survey of Radio Propagation Modeling for Tunnels," in IEEE Communications Surveys & Tutorials, vol. 16, no. 2, pp. 658-669, Second Quarter 2014.

[2] M. Bosse, R. Zlot and P. Flick, "Zebedee: Design of a Spring-Mounted 3-D Range Sensor with Application to Mobile Mapping," in IEEE Transactions on Robotics, vol. 28, no. 5, pp. 1104-1119, Oct. 2012.

[3] M. D. Bedford and G. A. Kennedy, "Modeling Microwave Propagation in Natural Caves Passages," in IEEE Transactions on Antennas and Propagation, vol. 62, no. 12, pp. 6463-6471, Dec. 2014.

Pametno vinogradništvo: načrtovanje pilota v Šmarjah pri Kopru z uporabo LoRaWAN tehnologije

Aleksander Hrastič [1], Vid Čermelj [1], Emilija Stojmenova Duh[1]

[1]*Laboratorij za telekomunikacije, Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, Ljubljana*
E-pošta: aleksander.hrastic@ltfe.org

Smart viticulture: designing a pilot in Šmarje near Koper using LoRaWAN technology

Abstract. *In this paper we present an architecture for a smart viticulture solution that will be used as a pilot in Slovenian Istria. We discuss and analyze various appropriate IoT connectivity technologies and present the parameters that will be measured. We analyzed various agricultural research and analysis papers and chose the parameters that were presented as important in viticulture.*

Smart viticulture requires sensors and connectivity to work as intended. Therefore, we propose architecture that uses sensor systems and LoRaWAN connectivity to gather vineyard related data. The data gathered from the two pilot vineyards will be used in machine learning algorithms and other analysis with intent to help winegrowers improve wine production and quality.

1 Uvod

V zadnjih letih se je z uporabo različnih brezžičnih senzorjev, dronov, satelitskih slik ter drugih novih merilnih orodij, uveljavil koncept pametnega vinogradništva, ki temelji na opazovanju, merjenju in zbiranju okoljskih parametrov za optimizacijo pridelave vina in zmanjševanja škodljivega vpliva na okolje [1]. Največja slabost pametnega vinogradništva so stroški vzdrževanja in začetni kapital, ki ga mora vinogradnik vložiti v nakup strojne in programske opreme. Poleg reševanja dolgoročnih okoljskih problemov, ki nastajajo s prekomerno porabo pitne vode in agrokemikalij, želimo s pomočjo statističnih obdelav in strojnega učenja na merjenih parametrih pomagati vinogradniku učinkoviteje sprejemati odločitve in s tem izboljšati proizvodnjo in kakovost pridelanega grozdja in kasneje tudi vina.

Digitalne tehnologije, vključno z internetom stvari (IoT) [2], vinogradnikom ponujajo nove rešitve, ki jim lahko pomagajo pri prilagoditvi načina proizvodnje pridelka v njihovih vinogradih.

2 Pregled povezljivostnih tehnologij v internetu stvari

Internet stvari je sistem medsebojno povezanih računalniških naprav, mehanskih in digitalnih strojev, opremljenih z enoličnimi identifikatorji in zmožnostjo prenosa podatkov po omrežju, ne da bi pri tem potrebovali interakcije med človekom in človekom ali človekom in napravo [2].

Povezljivostno tehnologijo arhitekture interneta stvari se uporablja kot sloj za učinkovito povezovanje senzorskih enot v lokalni oblak in posledično s ciljnim strežnikom. Glede na to, da se vinogradi večinoma nahajajo na podeželju in imajo omejen dostop do interneta ter omrežnega napajanja, je smiselno proučiti nekatere povezljivostne tehnologije, ki ustrezajo omenjenim omejitvam in omogočajo pravilno delovanje sistema.

Pri pregledu sorodnih rešitev smo zasledili uporabo tehnologije LoRaWAN, ki je bila uporabljena pri komunikaciji iz vinograda na strežnik [3] in kot del sistema na kmetiji [4].

2.1 LoRaWAN

Lora Alliance je standardizirala LPWAN (angl. Low power wide area network) tehnologijo imenovano LoRaWAN [5]. Komunikacija poteka v nelicenciranih frekvenčnih območjih (868 Mhz v Evropi). Omrežje sestavljajo senzorske enote z LoRaWAN povezljivostjo, LoRaWAN strežnik in LoRaWAN prehodi. Prednost te tehnologije je dolg radiofrekvenčni doseg (na podeželju s primerno anteno: 10-15 km) in majhna poraba električne energije v končnih senzorskih enotah, v zameno za počasnejši prenos podatkov in slabšo odzivnost [6].

2.2 4G

4G je četrta generacija tehnologije širokopasovnega celičnega omrežja. Glavne prednosti te tehnologije so globalna dostopnost, velika skalabilnost ob razširitvi senzorskih enot, širok frekvenčni pas, nizki cenovni razred naprav s 4G tehnologijo in dolga življenjska doba baterije povezanih naprav [7].

2.3 NB-IoT

Ozkopasovni internet stvari (angl. narrowband internet of things - NB-IoT) [8] je LPWAN standardna tehnologija, ki lahko sobiva z LTE (angl. Long term evolution) in GSM (angl. Global System for Mobile Communications) v licenciranih radiofrekvenčnih pasovih. Namenjena je velikemu številu priključenih naprav. V primerjavi s 4G tehnologijo ima krajše časovne intervale komunikacije med napravo interneta stvari in bazno postajo. Posledično je manjša tudi poraba energije povezanih senzorskih enot [9]. NB-IoT povezljivost je možno koristiti kjer je v bližini

postavljena bazna postaja, saj je predviden doseg signala v urbanih področjih 1 km in v podeželskih 10 km [10].

2.4 Sigfox

Ponudniki storitev LPWAN Sigfox tehnologije nameščajo bazne postaje v različnih državah po svetu. Tehnologija prav tako obratuje v nelicenciranih frekvenčnih območjih (868 Mhz v Evropi). Ima majhne nivoje prisotnosti šuma v komunikaciji, kar omogoča nizko porabo baterije v končnih brezžičnih napravah interneta stvari [10]. Sigfox ima omejeno maksimalno število poslanih paketov na 14 na dan in maksimalno velikost posamičnega paketa na 12 bajtov [6]. Uporaba Sigfoxa, kjer je signal na voljo, je možna z nakupom Sigfox modula, registracije v Sigfox omrežje in plačevanjem po številu prenesenih podatkov. Trenutno je Slovenija slabo pokrita s Sigfox omrežjem [11].

2.5 Primerjava povezljivostnih tehnologij

Glavni dve lastnosti, ki sta nam bili pomembni pri izbiri tehnologije sta bili poraba energije in pokritost signala v Šmarjah ter drugod po Sloveniji, saj se vinogradi nahajajo na odtujenih območjih, kjer ni dostopa do omrežnega napajanja in dobrega signala. Hitrost prenosa podatkov in odzivnost pri izbiri povezljivostne tehnologije nista imela velikega pomena. Podatki so prikazani v Tabeli 1.

Tabela 1. Primerjava povezljivostnih tehnologij [6][7][10][11]

Ime tehnologije	Teoretična hitrost prenosa podatkov	Poraba energije	Razpoložljivost signala v Sloveniji
LoRaWAN	50 kb/s	Zelo nizka	Do 15 km od postavljenega prehoda.
4G	1 Gb/s	Visoka	Visoka pokritost
NB-IoT	200 kb/s	Nizka	Delna pokritost, do 10 km od bazne postaje.
Sigfox	100 b/s	Zelo nizka	Slaba pokritost

Izmed predlaganih tehnologij, 4G in NB-IoT nimata zagotovljene pokritosti na območjih, kjer bi želeli imeti postavljene senzorje. S pomočjo LoRaWAN prehoda lahko ustvarimo lastno omrežje in s tem zagotovimo povezljivost senzorjev v internet oziroma do strežnika, kjer drugače to ne bi bilo možno. Pri tem je glavna slabost strošek nakupa in vzdrževanja dodatne strojne opreme (LoRaWAN prehoda).

3 Izbrani merjeni parametri

Ključen del pri izvajanju pilota je zbiranje relevantnih agronomskih parametrov.

Z upoštevanjem preteklih agronomskih raziskav in analiz [12]-[23] smo se odločili, da bomo v pilotnem vinogradu merili 12 parametrov: temperaturo zraka, vlago v zraku, mokroto listov, zračni pritisk, sončno obsevanje, temperaturo zemlje, vsebnost vlage v zemlji, vsebnost kisika v zemlji, prevodnost zemlje, količino padavin ter hitrost in smer vetra.

Z vrednostmi nekaterih parametrov bomo lahko vinogradnikom neposredno podali oceno in svetovali glede ukrepov. Pri preostalih pa bo potrebna bolj poglobljena analiza, saj so vrednosti relevantne v daljšem časovnem obdobju.

Za merjenje parametrov v vinogradu se bodo uporabljali različni senzorji, ki se jih postavi v zemljo, v bližino ali neposredno na trto. Izbrana postavitev senzorja je odvisna od načina njihovega delovanja [1].

3.1 Temperatura zraka

Z višjimi temperaturami ozračja nastane grozdje z drznejšimi okusi, večjo vsebnostjo sladkorja in posledično tudi vino z več alkohola [12].

O. Ashenfelter in K. Strochman sta z analizo ugotovila , da se ob globalnih vremenskih spremembah, ki povzročajo toplejše temperature zraka, kvaliteta grozdja in s tem tudi kvaliteta vina povečata. Posledično se poveča tudi njuna cena [13].

Z uporabo digitalnih temperaturnih senzorjev se bo v vinogradu, s sprotnim merjenjem temperature ozračja, beležilo temperature v obdobju rasti grozdja in na podlagi zbranih podatkov ocenilo kvaliteto ter sladkobo grozdja in vina.

3.2 Vlaga v zraku

Vlaga v ozračju vinograda je v kombinaciji s temperaturo ozračja eden izmed faktorjev, ki vpliva na razvoj škodljivih plesni (npr. Praškasta plesen) na vinskih trtah. Carroll in Wilcox sklepata, da je možnost za razvoj te plesni proporcionalno odvisna od količine prisotne vlage v ozračju vinograda [14].

S sprotnim senzorskim merjenjem in beleženjem vlage, se lahko v kombinacij z izmerjeno temperaturo ozračja predvidi, ali se je v vinogradu v nekem časovnem obdobju povečala verjetnost za razvoj škodljivih plesni. Z izvajanjem te meritve se bo vinogradnikom, na podlagi statističnih podatkov in uporabe strojnega učenja, pomagalo s pravočasnim opozorilom na preventivni ukrep.

3.3 Sončno sevanje

Samo en del sončnega sevanja neposredno doseže zemeljsko površino. Drugi del se v atmosferi razprši in nato doseže zemljino površino kot razpršeno sevanje [13]. Sončno sevanje je eden izmed glavnih pogojev za proces fotosinteze v rastlinah, s katerim se v rastlini proizvajajo sladkorji. Temperatura ozračja je neposredno povezana z energijo sončnega sevanja, saj s soncem segreta zemlja v okolico oddaja toploto.

Sevanje, ki doseže zemeljsko površino, lahko merimo z uporabo piranometra (senzor za merjenje

sončnega sevanja). Z zbiranjem senzorskih podatkov bomo s pomočjo analize ocenili kvaliteto pridelanega grozdja in vina.

3.4 Temperatura prsti

Prst v vinogradu se s sončnim obsevanjem segreva, kar ima različne vplive na biokemijske procese, ki se izvajajo v korenini rastline. Nekateri izmed procesov so vnos vode skozi korenine v rastlino, ki se zmanjša pri nižjih temperaturah, razgradnja organskih snovi, kar vpliva na vrednost PH zemlje in izhlapevanje vode iz zemlje. Poleg naštetih, temperatura zemlje vpliva še na druge pomembne lastnosti tal [15].

Temperaturo prsti se bo merilo s sondo, ki je zakopana v zemljo ob območju koreninskega sistema.

3.5 Vsebnost vlage v zemlji

Količina padavin, ki pade na merjeno območje, neposredno vpliva na vsebnost vlage v zemlji. Ob primanjkljaju vlage v zemlji lahko sklepamo, da je zemlja na območju kjer merimo suha. Glede na zahteve rastline se vinogradniku sporoči, da je na opazovanem območju potrebno zaliti trte.

Merjenje vlage v zemlji je možno izvesti na več načinov. V našem pilotu bomo uporabili senzor s kapacitivnim načinom merjenja [16].

3.6 Električna prevodnost zemlje

Z merjenjem električne prevodnosti zemlje lahko pridobimo informacije o različnih lastnosti tal, ki vplivajo na proizvodnjo končnega pridelka. Nekatere izmed teh so: tekstura tal, zmogljivost izmenjave kationov, drenažni pogoji zemlje, vsebnost organske snovi v zemlji, slanost in druge značilnosti podzemlja [17]. Eden izmed parametrov, ki je povezan s prevodnostjo zemlje, je slanost zemlje. V ne-optimalnih količinah slanost zemlje slabo vpliva na rast trt in kvaliteto pridelanega vina [18].

3.7 Količina padavin

Padavine so bistvenega pomena za preživetje rastline in proizvajanje pridelka v vinogradih. Pomanjkanje vode v vinogradih vpliva na zmanjšanje velikosti končnega grozdnega sadeža [19].

Z merjenjem količine padavin v vinogradih se lahko beleži in ugotavlja, če padavine odstopajo od povprečja in na podlagi tega predvidi porabo vode za zalivanje. Za merjenje količine padavin bomo v vinogradu uporabili ombrometer s prekucno posodo (angl. tripping bucket).

3.8 Mokrota listov

Mokrota listov je merjena kot prisotnost vode na površini lista. V primeru, da je list moker več časa kot traja pragovno obdobje pri določeni temperaturi, je povečana verjetnost nastanka škodljivih bolezni [20].

Z merjenjem mokrote lista bo možno predvideti, če se bo glede na dane pogoje razvila bolezen. Vinogradniku se pomaga s pravočasnim opozorilom na uporabo preventivnih ukrepov.

3.9 Vsebnost kisika v zemlji

Vodilni faktorji, ki vplivajo na vsebnost kisika v zemlji, so poroznost zemlje, tekstura zemlje, struktura zemlje in vsebnost vode v zemlji. Pomanjkanje kisika lahko povzroči, da se v zemlji proizvajajo rastlini škodljive snovi [21].

Z merjenjem vsebnosti kisika v zemlji lahko spremljamo dolgoročne spremembe lastnosti zemlje.

3.10 Smer in hitrost vetra

Neprekinjeni močni vetrovi lahko poškodujejo trto in preprečijo njeno rast. Poleg tega, veter hitreje odnaša toploto s površin listov in zemlje kar pomeni, da bo povprečna temperatura ozračja vinograda nižja. Po drugi strani veter povzroči, da se voda s površine listov hitreje posuši, kar zniža verjetnost nastanka škodljivih bolezni [22].

Z merjenjem hitrosti vetra se lahko ugotovi, če je veter pihal dovolj močno in dovolj dolgo, da so se listi posušili. Te informacije lahko v kombinaciji s preostalimi podatki pomagajo vinogradniku z opozarjanjem na uporabo preventivnih ukrepov.

3.11 Zračni pritisk

Z merjenjem zračnega pritiska bomo preverjali ali ima rastlina primerne pogoje za rast, saj ob večjih odstopanjih ne dosega optimalne višine [23]. Poleg tega, bomo ta podatek uporabili v kombinaciji z vrednostmi drugih parametrov.

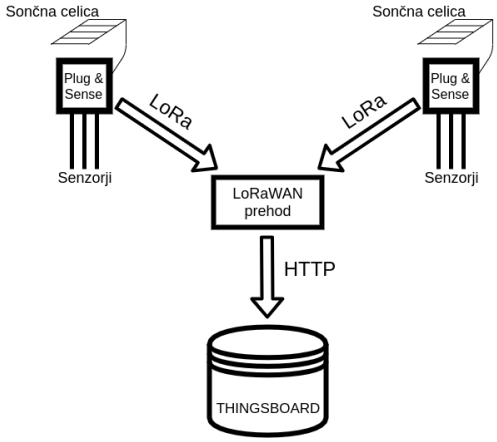
4 Pristop k izvedbi pilota

Z namenom spremljanja v prejšnjem poglavju naštetih parametrov v vinogradih, se na Univerzi v Ljubljani, Fakulteti za elektrotehniko, v Laboratoriju za Telekomunikacije razvija rešitev, ki bo implementirana v pilotnem projektu, ki ga izvajamo v okviru projektov Interreg Alpine Space Smart Villages in Horizon 2020 LiveRur. Rešitev bo v prvi fazi nameščena na dveh vinogradniških kmetijah v Šmarjah pri Kopru. V drugi fazi bo enaka rešitev nameščena na vsaj še treh kmetijah v slovenski Istri ter dveh kmetijah na Štajerskem koncu.

4.1 Arhitektura rešitve

Pri snovanju ustrezne arhitekture je potrebno upoštevati lastnosti lokacije pilotnega vinograda. Vinograda se nahajata v urbanem območju v bližini stavb in sta med seboj oddaljena 100 metrov. Ker na območjih ni možno ustrezno zagotoviti omrežnega napajanja, bosta senzorska sistema napajana prek baterij in sončnih celic. Z upoštevanjem teh lastnosti in z ozirom na lastnosti različnih komunikacijskih tehnologij smo za naše testno območje izbrali tehnologijo LoRaWAN. Z LoRaWAN prehodom bomo ustvarili lastno omrežje, ki bo neodvisno od ponudnikov telekomunikacijskih storitev.

LoRaWAN omogoča uporabo v načinu nizke porabe energije in pošiljanje podatkov na daljše razdalje, kar nam bo olajšalo dodajanje novih senzorskih enot in vinogradov, ki bodo od prehoda bolj oddaljeni. Uporabljena bo zvezdna topologija, kar pomeni, da bodo senzorski sistemi ločeno povezani na en skupen prehod (Slika 1).
Vinograda sta znotraj dosega LoRaWAN signala, zato zadošča, da le z enim prehodom, nameščenim v bližini vinogradov, ustvarimo svoje lokalno povezljivostno omrežje. Prehod se nato preko interneta poveže s strežnikom na katerem je nameščen ThingsBoard.



Slika 1.Arhitektura *rešitve*.

Senzorska sistema iz obeh vinogradov bosta tekom dneva v različnih časovnih intervalih merila parametre in jih na vsakih 60 minut pošiljala v ThingsBoard na strežniku. Prehod v naši rešitvi služi kot vmesna dostopna točka med strežnikom in senzorskim sistemom. Med senzorskim sistemom in prehodom bodo podatki poslani z uporabo LoRaWAN komunikacijskega protokola, med prehodom in strežnikom pa s HTTP protokolom. Maksimalna velikost enega poslanega paketa (med 51 in 222 bajti), energijsko porabo, doseg in hitrost prenosa, je možno spremeniti v nastavitvah protokola [6].

4.2 Izbira senzorskega sistema

Ob izbiri senzorskega sistema nam je bila pomembna možnost, da na sistem namestimo našo programsko kodo. Ker se bo senzorski sistem nahajal v naravi, moramo upoštevati odpornost na vremenske razmere.

Izmed rešitev, ki so trenutno prosto dostopne na trgu, našim kriterijem najbolj ustreza profesionalni senzorski sistem Libelium Plug&Sense Smart Agriculture Xtreme. Omenjen senzorski sistem ima glede na izbiro konfiguracije možnost uporabe več različnih celičnih in LPWAN komunikacijskih tehnologij [24].

Za našo rešitev smo izbrali konfiguracijo, ki omogoča povezljivost senzorskega sistema v LoRaWAN omrežje. Ker na lokacijah senzorskih sistemov ni možnosti omrežnega napajanja, bo napajanje izvedeno z interno polnilno baterijo kapacitete 6600mAh in sončnimi celicami (7V – 500mA).

Senzorski sistem, ki ga bomo namestili v vinogradu, bo sestavljen iz centralne enote in 6 različnih senzorskih enot, ki bodo merile parametre, naštete v tretjem poglavju. Centralna enota bo skrbela za agregacijo in pošiljanje podatkov na centralni strežnik.

4.3 Izbrani senzorji

Na centralno enoto senzorskega sistema je možno priključiti 6 različnih senzorjev iz kataloga, ki ga nudi ponudnik senzorskega sistema Libelium.

Tabela 2. Izbrani senzorji.

Ime Senzorja	Merjeni parametri
TEROS 12	vsebnost vode, prevodnost in temperatura zemlje
Phytos 31	vlažnost listov
SO-411	vsebovanost kisika v zemlji
Temperature, Humidity and Pressure Probe	temperatura, vlažnost in zračni pritisk
SQ-110	sončno sevanje
WS-3000	vremenska postaja za merjenje količine padavin, hitrosti in smeri vetra

Iz vsakega senzorja bomo za vsak parameter pridobili podatek velikosti treh bajtov. Za vseh 12 parametrov bo končni paket koristnih informacij velik 36 bajtov.

4.4 Zbiranje in prikaz podatkov

Podatki bodo tekom pilotnega obdobja zbrani na strežniku, na katerem je nameščen ThingsBoard CE. ThingsBoard je platforma, ki omogoča zbiranje, prikazovanje in nadaljnjo obdelavo prejetih podatkov iz senzorskih modulov [25]. Zbrane podatke bomo uporabili za analizo in pomoč vinogradnikom pri sprejemanju odločitev v zvezi z upravljanjem vinograda. Za lažjo predstavo bodo uporabljeni grafi, ki bodo prikazovali spremembe parametrov v odvisnosti od časa ali drugih parametrov. Uporabljene bodo tudi metode strojnega učenja za iskanje korelacij med parametri in lastnostmi grozdja in vina. Vinogradniki bodo prek platforme lahko spremljali parametre v njihovih vinogradih. Platforma bo samodejno opozarjala na težave in preventivne ukrepe.

5 Zaključek

Pomembno odločitev pri načrtovanju merilnega sistema predstavlja izbira tehnologije, ki se jo uporabi za komunikacijo s strežnikom, na katerega se pošilja podatke. Do sedaj so bile za to najbolj primerne tehnologije LPWAN, v zadnjih letih pa je smiselna tudi uporaba celičnih komunikacijskih tehnologij. V Sloveniji imajo tehnologije z nizko porabo energije slabo pokritost, zato smo se odločili za tehnologijo LoRaWAN.

Z našo rešitvijo interneta stvari želimo na dveh pilotnih vinogradih testirati sistem, ki bo olajšal in izboljšal proces pridelave vina. S pomočjo analize zbranih podatkov bomo lahko vinogradniku pomagali pri sprejemanju pomembnih odločitev v zvezi z vinogradništvom s ciljem, da zmanjšamo stroške uporabe agrokemikalij in vode ter izboljšamo kvaliteto končnega pridelka. S poglobljeno analizo bomo pridobili tudi bolj natančne podatke o pomembnosti vsakega izmed 12 parametrov.

Po končanem pilotu želimo to znanje uporabiti za izboljševanje sistema in nadgrajevanje rešitve, ki jo bo mogoče uporabiti v različnih pokrajinah po Sloveniji in tudi smiselno posplošiti na preostale kmetijske panoge.

6 Literatura

[1] "Smart vineyard: management and decision making support for wine producers," [Online] Dosegljivo: https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/DTM_Smart%20vineyard%20v1.pdf [10.7.2020]

[2] "internet of things (IoT)," [Online] Dosegljivo: https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT [20.7.2020]

[3] D. Davcev, K. Mitreski, S. Trajkovic, V. Nikolovski, N. Koteli, "IoT agriculture system based on LoRaWAN," 2018 14th IEEE International Workshop on Factory Communication Systems (WFCS), Imperia, str. 1-4, 2018.

[4] Codeluppi, G.; Cilfone, A.; Davoli, L.; Ferrari, G. "LoRaFarM: A LoRaWAN-Based Smart Farming Modular IoT Architecture, " Sensors 2020, vol. 20, št. 7, 2020.

[5] Lora Alliance, [Online] Dosegljivo: https://www.lora-alliance.org/ [12.7.2020]

[6] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui and T. Watteyne, "Understanding the Limits of LoRaWAN," *IEEE Communications Magazine*, vol. 55, št. 9, str. 34-40, 2017

[7] J. J. Sara, M. S. Hossain, W. Z. Khan and M. Y. Aalsalem, "Survey on Internet of Things and 4G," 2019 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), Tangerang, Indonesia, str. 1-6, 2019.

[8] "Narrowband - Internet of Things (IoT)," [Online] Dosegljivo:https://www.gsma.com/iot/narrow-band-internet-of-things-nb-iot/ [20.7.2020]

[9] "LTE-M vs NB-IoT – A guide exploring the differences between LTE-M and NB-IoT," [Online] Dosegljivo: https://www.telenorconnexion.com/iot-insights/lte-m-vs-nb-iot-guide-differences/ [16.7.2020]

[10] K. Mekki, E. Bajic, F. Chaxel and F. Meyer, "Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT," 2018 IEEE International Conference on Pervasive Computing and Communications Workshops, Atene, str. 197-202, 2018

[11] Sigfox pokritost, [Online] Dosegljivo: https://www.sigfox.com/en/coverage [18.7.2020]

[12] "YaleEnviroment360, What Rising Temperatures May Mean for World's Wine Industry," [Online] Dosegljivo: https://e360.yale.edu/features/what_global_warming_may_mean_for_worlds_wine_industry [10. 7. 2020]

[13] O. Ashenfelter, K. Strochman, "Using Hedonic Models of Solar Radiation and Weather to Assess the Economic Effect of Climate Change: The Case of Mosel Valley Vineyards," Review of Economics and Statistics, vol. 92, št. 2, str. 333-349, 2010.

[14] J. E. Carroll, W. F. Wilcox, "Effects of humidity on the development of grapevine powdery mildew," Phytopathology, vol. 93, št. 9, str. 1137-1144, 2003.

[15] O. Brown, "Effects of soil temperature on Some Soil properties and plant growth," Journal of Agricultural Science and Technology, vol. 6, št. 3, str. 89-93, 2016

[16] "Using Soil Moisture Sensors for Vineyard Irrigation Management," [Online] Dosegljivo: https://www.vineyardteam.org/files/resources/Using%20Soil%20Moisture%20Sensors%20in%20Vineyards_2014_F.Westover_Final%20Working.pdf [12.7.2020]

[17] R. Grisso, M. Alley, D. Holshouser, W. Thomason, "Precision Farming Tools: Soil Electrical Conductivity," Virginia Cooperative Extension, str. 442-508, 2005

[18] "Vineyard responses to salinity-associated soil chemical properties," [Online] Dosegljivo: https://www.wineladn.co.za/vineyard-responses-salinity-associated-soil-chemical-properties/ [16.7.2020]

[19] Williams, Larry E. "Grapevine water relations." Raisin production manual (Ed. LP Christensen), str. 121-126, 2000.

[20] R. A. Spotts, "Effect of leaf wetness duration and temperature on the infectivity of Guignardia bidwellii on grape leaves," Phytopathology vol. 67, št. 11, str. 1378-1381, 1977.

[21] J.Neira, M. Ortiz, L. Morales, E. Acevedo. "Oxygen diffusion in soils: Understanding the factors and processes needed for modeling," Chilean journal of agricultural research, vol. 75, št. 1, str. 35-44, 2015.

[22] H. Jagoutz, "The effect of wind," Geologische Abhandlungen von Hessen, vol. 114, str. 42-53, 2004.

[23] H. J. Daunicht, H. J. Brinkjans, "Plant responses to reduced air pressure: Advanced techniques and results," Advances in Space Research, vol. 18, št. 4–5, str. 273-281, 1996.

[24] Libelium Plug&Sense, [Online] Dosegljivo: http://www.libelium.com/products/plug-sense [18.7.2020]

[25] ThingsBoard, [Online] Dosegljivo: https://thingsboard.io/ [17.7.2020]

Določanje dejanskega efektivnega lomnega količnika svetlovoda v silicijevem optičnem čipu

Andrej Lavrič, Janez Krč, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko
Tržaška cesta 25, 1000 Ljubljana
E-pošta: andrej.lavric@fe.uni-lj.si

Determination of actual effective refractive index of waveguide in silicon photonic chip

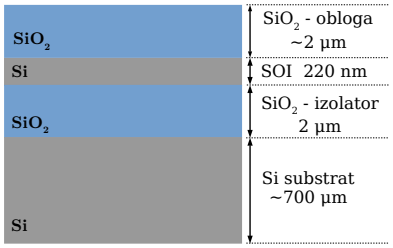
Integrated silicon photonics has gained significant attraction in various technology fields such as optical communications, sensing, data processing and transfer. In this contribution we present a simple optical chip which was designed to perform different tests. In particular we show the results of determination of an actual effective refractive index of an integrated waveguide by using Mach-Zehnder interferometer.

1 Uvod

Silicijeva fotonika predstavlja mlado in obetavno tehnološko področje, ki je svoj zagon dobilo v začetku tisočletja. Z zahtevo po vse večji procesorski moči se je pojavila tudi vse večja potreba po zmogljivosti prenosa podatkov, ki jih omejujejo izgube električnih povezav v klasičnih električnih čipih [1]. Drugi izziv predstavljajo zahteve po večji hitrosti obdelave podatkov in računanja, kjer postaja elektronika omejena. V tej luči se odpira široko področje uporabe optičnih čipov. Kljub ugodnejšim elektro-optičnim lastnostim nekaterih drugih materialov, silicijeva tehnologija omogoča združljivost z zrelo tehnologijo komplementarne kovina-oksid-polprevodnik (angl. Complementary metal-oxide-semiconductor – CMOS). Prav slednja se pojavlja kot glavni motiv za raziskovanje in razvoj optičnih čipov (angl. Photonic Integrated Circuit – PIC) na siliciju [2, 3, 4].

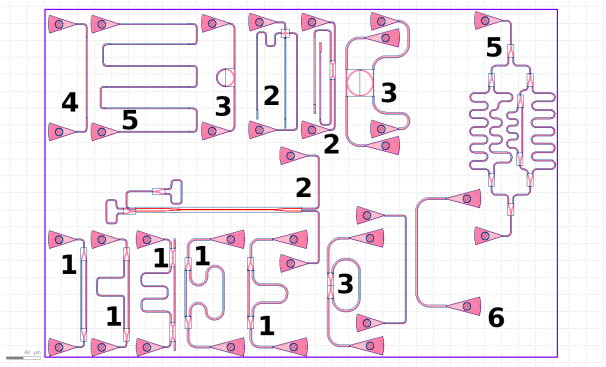
V tem prispevku je najprej na kratko predstavljen preprost testni optičen čip velikosti 600×400 µm^2, ki smo ga načrtovali v tehnologiji silicij-na-izolatorju (angl. Silicon-on-insulator – SOI). V osrednjem delu prispevka se bomo osredotočili na določanje dejanskega efektivnega lomnega količnika svetlovoda na čipu. Svetlovodi nastopajo kot osnovni gradniki vseh ostalih komponent.

Silicijeva rezina tehnologije SOI sestoji iz ∼700 µm substrata silicija (Si), na njem je 2 µm debela plast silicijevega dioksida (SiO_2) – izolatorja, na vrhu pa je 220 nm plast silicija, iz katere se jedka optično vezje. Po končnem jedkanju se na čip napari zgornjo zaščitno plast silicijevega oksida. Presek takšne silicijeve rezine je prikazan na sliki 1. Na našem izdelanem čipu so preproste testne strukture, med njimi uklonske mrežice za sklaplanje svetlobe v čip [5, 6, 7, 8], ravni in zaviti trakasti svetlovodi, Y-spoji [9], smerni sklopniki, obročasti rezonatorji,



Slika 1: Presek rezine tipa SOI, na kateri je izdelan predstavljen optičen čip.

Braggove rešetke. Na osnovi teh je na čipu zgrajenih več Mach-Zehnderjevih interferometrov (MZI), Michelsonovih interferometrov in različnih vrst filtrov. Postavitev čipa z označenimi pasivnimi strukturami je prikazana na sliki 2. Pri vseh komponentah svetlobo pripeljemo in jo zajamemo preko vhodno/izhodnih uklonskih mrežic. Nanos maske na silicijevo rezino je izveden z litogra-



Slika 2: Postavitev (angl. layout) optičnega čipa. Posnetek zaslona iz programa Klayout. 1: Mach-Zehnderjevi interferometri, 2: Michelsonovi interferometri, 3: obročasti resonatorji, 4: Braggova rešetka, 5: optični filtri, 6: uklonska mrežica. Na samem čipu ni optičnih virov in detektorjev, tako da imajo vse strukture na vhodu/izhodu uklonske mrežice za sklopitev svetlobe iz/v optično vlakno. Postavitev je bila narisana z odprtokodnim programom Klayout in uporabo SiEPIC PDK [10].

fijo z elektronskim snopom ločljivosti 6 nm [11]. Celoten postopek izdelave čipa je bil narejen pri Washington Nanofabrication, ki je del University of Washington. Za litografijo z elektronskim snopom so predpostavljene

tolerance -30...+10 nm vzete iz standardnega postopka CMOS. Pri izdelavi so bile uporabljene rezine SOI proizvajalca Soitec (Grenoble, Francija), ki imajo povprečno debelino Si plasti 219,3 nm s standardno deviacijo 3,9 nm. Opisane geometrijske tolerance izdelave optičnega čipa narekujejo odstopanje optičnih lastnosti čipa od načrtovanih vrednosti. To zahteva analizo meritev, ki pokaže ali se lastnosti izdelanega čipa nahajajo znotraj pričakovanega intervala vrednosti. Preprosta ocena intervala pričakovanih vrednosti je kotna analiza (angl. corner analysis) [12], pri kateri izračunamo vrednosti efektivnega in skupinskega lomnega količnika za skrajna odstopanja geometrije.

V tem prispevku niso opisane funkcionalnosti vseh gradnikov na čipu. Osredotočili se bomo le na gradnike MZI in pokazali, kako z njihovo pomočjo določimo dejanski efektivni lomni količnik svetlovoda.

2 Delovanje MZI in pomen efektivnega lomnega količnika

Mach-Zehnderjev interferometer je razmeroma preprosta struktura, ki vhodni optični signal razdeli v dve veji, običajno različnih dolžin. Na izhodu se optična signala spet združita. Ker v splošnem veji interferometra nista enaki, signala prepotujeta različne poti in si na poti pridobita različna fazna zasuka. Ob ponovni združitvi signalov nazaj na izhodu MZI, zaradi razlik v fazi valov pride bodisi do ojačitev ali oslabitev jakosti združenega optičnega signala. V svetlovodih lahko izgube zanemarimo. Če predpostavimo, da izgub ni, lahko opišemo električno poljsko jakost signalov pred združitvijo

$$E_{o1} = \frac{E_i}{\sqrt{2}} e^{-i\beta_1 L_1} \tag{1}$$

$$E_{o2} = \frac{E_i}{\sqrt{2}} e^{-i\beta_2 L_2}, \tag{2}$$

kjer sta L_1 in L_2 dolžini posameznih poti, β_1 in β_2 pa pripadajoči fazni konstanti. Fazna konstanta pa je odvisna od valovne dolžine

$$\beta(\lambda) = \frac{2\pi n_{eff}(\lambda)}{\lambda} = \frac{2\pi}{\lambda_{eff}}, \tag{3}$$

kjer je λ valovna dolžina v vakuumu, λ_{eff} je efektivna valovna dolžina v svetlovodu in n_{eff} je efektivni lomni količnik svetlovoda, ki je odvisen tako od lomnih količnikov snovi kot od geometrije in je v istem svetlovodu različen za različne rodove. Predpostavimo, da so lokalno svetlovodi enaki, kljub tolerancam izdelave, tako da velja $\beta_1 = \beta_2 = \beta$. Ob teh predpostavkah se poti razlikujeta zgolj v dolžini. Prenosna funkcija interferometra je

$$T_{MZI}(\lambda) = \frac{1}{4}\left|1 + e^{-i\beta(\lambda)\Delta L}\right|^2, \tag{4}$$

kjer je ΔL razlika v dolžini obeh vej interferometra. Ko je ta enaka lihim večkratnikom $\lambda_{eff}/2$, ima prenosna funkcija ničle. Torej gre za periodično filtrsko karakteristiko MZI-ja, ki se odraža tudi na karakteristikah prikazanih na slikah 4, 5 in 6. Razdaljo med sosednjimi ničlami (minimumi v prenosni funkciji) imenujemo prosto spektralno področje (angl. free spectral range – FSR), ki je podano kot

$$FSR = \frac{\lambda^2}{n_g \Delta L}, \tag{5}$$

kjer je n_g skupinski lomni količnik. V tem prispevku smo osredotočeni na efektivni lomni količnik, vendar so podane tudi relacije do skupinskega, saj se ta pojavlja kot pomemben parameter pri načrtovanju resonatorjev [13], ki so osnova za sestavljanje kompleksih operacij obdelave optičnega signala. Položaj ničel v prenosni funkciji MZI v enačbi (4) je poleg ΔL odvisen tudi od fazne konstante β, ki pa je odvisna od efektivnega lomnega količnika svetlovoda. Iz prikazanih relacij je razvidno, da lahko preko odziva MZI določimo dejansko vrednost efektivnega lomnega količnika svetlovoda.

3 Načrtovanje efektivnega lomnega količnika

Vrednost efektivnega lomnega količnika svetlovoda lahko predvidimo (načrtamo) z metodo simulacije različnih rodov v strukturi [12, 14]. Takšna simulacija daje vrednosti pri eni sami valovni dolžini in zahteva reševanje valovne diferencialne enačbe, kar je lahko zamudno. Za lažje načrtovanje in kasnejšo analizo meritev si želimo preproste analitične enačbe. Slednjo sestavimo kot polinom nizke stopnje, ki aproksimira rešitve poljskih simulacij v nekaj diskretnih točkah. Efektivni lomni količnik lahko zapišemo kot

$$n_{eff}(\lambda) = n_1 + (\lambda - \lambda_0)n_2 + (\lambda - \lambda_0)^2 n_3. \tag{6}$$

kjer je n_1 efektivni lomni količnik pri λ_0, n_2 in n_3 pa sta koeficienta efektivnega lomnega količnika v polinomski enačbi, ki vsebujeta tudi odvisnost od valovne dolžine. Skupinski lomni količnik n_g je v tesni zvezi z efektivnim

$$\begin{aligned} n_g(\lambda) &= n_{eff}(\lambda) - \lambda \frac{dn_{eff}}{d\lambda} \\ &= n_{eff}(\lambda) - \lambda(n_2 + 2(\lambda - \lambda_0)n_3). \end{aligned} \tag{7}$$

Pri $\lambda = \lambda_0 = 1{,}55$ µm velja kar

$$n_{eff} = n_1 \quad \text{in} \quad n_g = n_1 - 1{,}55 n_2. \tag{8}$$

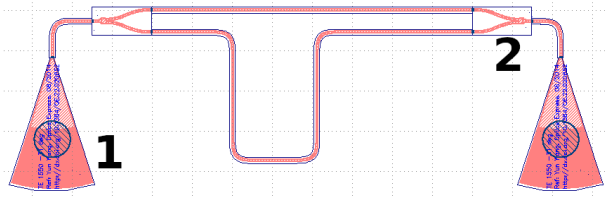
V tabeli 1 so koeficienti prvih dveh rodov valovoda s presekom 500×220 nm. Slednje določimo tako, da najprej v nekaj diskretnih vrednostih napravimo simulacijo rodov, nato pa z metodo najmanjših kvadratov rešitve simulacij aproksimiramo s polinomom nizke stopnje. Koeficienti tega polinoma so koeficienti za naše določanje lomnega količnika v enačbi (6). V nadaljevanju bo predpostavljen postopek določanja dejanskega n_{eff} le za transverzalno električni (TE) rod valovanja. To je rod, ki ima večino električnega polja usmerjenega v horizontalni smeri glede na substrat. Postopek načrtovanja in analize je popolnoma identičen za transverzalno magnetni (TM) rod – rod z večino električnega polja vertikalno glede na substrat. Razlika je le v izbiri koeficientov v enačbi (6) in vseh sledečih izrazih.

Rod	n_1	n_2	n_3	n_g
TE	2,4453	-1,1295	-0,0421	4,1960
TM	1,7727	-1,2795	1,8131	3,7559

Tabela 1: Iz predhodnih simulacij izračunani koeficienti prvih dveh rodov v valovodu s presekom 500×220 nm. Pri analizi bomo osredotočeni zgolj na TE rod.

Kot smo omenili, je n_{eff} odvisen tudi od geometrije svetlovoda. Torej, na njega vplivajo geometrijska tolerančna območja, ki jih dosežemo pri dejanski izdelavi svetlovodov na čipu. V praksi so zato fotonska vezja prepredena z lokalnimi grelnimi elementi [15, 16], kjer s kontroliranimi spremembami v temperaturi vplivamo na snovni lomni količnik in s tem posredno kompenziramo odstopanja n_{eff} zaradi toleranc geometrije.

V našem primeru bomo določevali dejanski n_{eff} (pri sobni temperaturi) na osnovi izbranega MZI. Povečava izbranega MZI iz slike 2 je prikazana na sliki 3 (rotirano).

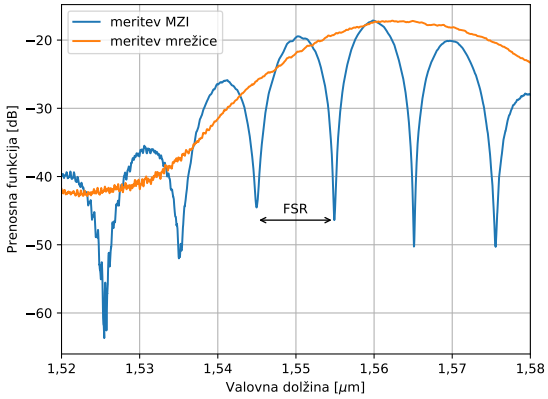


Slika 3: MZI z $\Delta L = 57{,}48$ µm za TE rod (na sliki 2 stoji vertikalno). 1: uklonska mrežica [5, 6, 7, 8], 2: Y-spoj [9].

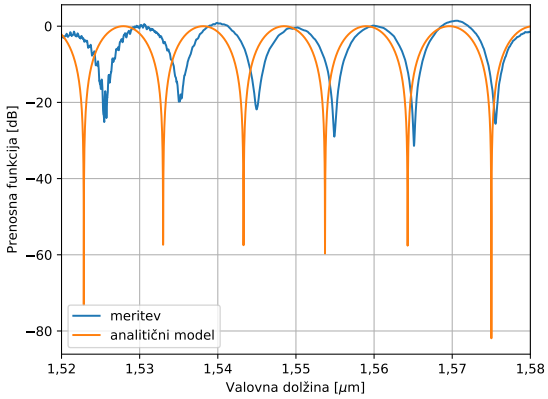
4 Rezultati meritev in analiza

Za meritev čipa je bila uporabljena namenska avtomatizirana merilna postavitev [12] in je bila izvedena v sklopu tečaja UBCx Phot1x na University of British Columbia. Kot svetlobni vir je bil uporabljen nastavljiv laser Agilent 81600B, na detektorski strani pa merilniki optične moči Agilent 81635A. Prelet valovne dolžine laserja je od 1500 nm do 1600 nm s korakom 10 pm.

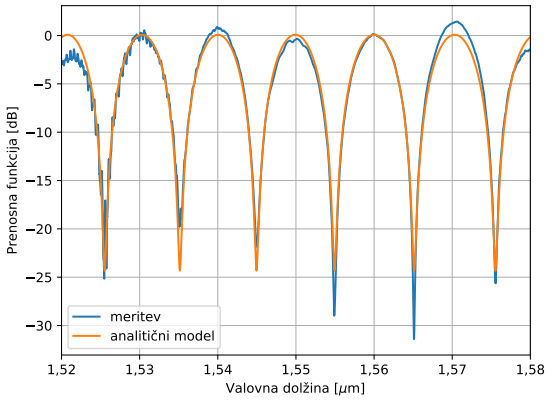
Prvi korak pri analizi meritev je odstranitev frekvenčne odvisnosti uklonskih mrežic pri meritvi MZI. Na sliki 4 je prikazana prikazana meritev prenosne karakteristike izbranega MZI skupaj z vhodno/izhodnimi uklonskimi mrežicami in prenosna karakteristika zgolj vhodno/izhodnih uklonskih mrežic povezanih z enostavnim svetlovodom. V naslednjem koraku meritev MZI normaliziramo, tj. odstranimo odziv uklonskih mrežic, in normalizirano meritev primerjamo z načrtovanim analitskim odzivom MZI. Na sliki 5 so prikazane normalizirane meritve in prenosna funkcija za načrtovan MZI, izračunana po enačbi (4), kjer so za n_{eff} uporabljeni koeficienti za TE rod iz tabele 1. Premik ničel prenosne funkcije potrjuje odmik dejanskega efektivnega lomnega količnika valovoda od načrtovane vrednosti. V zadnjem koraku iz meritev izluščimo dejanski lomni količnik valovoda z uporabno metode najmanjših kvadratov. Koeficiente v enačbi (6) spreminjamo, tako da je odstopanje rezultata enačbe (4) od normaliziranih meritev (glej sliko 5) najmanjše. Na sliki 6 je prikazan končni rezultat analize. Izluščen efektivni



Slika 4: Meritev prenosne funkcije MZI vključno z uklonskimi mrežicami (modra) in meritev odziva uklonskih mrežic brez MZI strukture (oranžna).



Slika 5: Normalizirana meritv prenosne funkcije MZI (modra) in analitska prenosna funkcija za načrtovan MZI (oranžna).



Slika 6: Končni rezultat analize. Normalizirana meritev MZI (modra) in nanjo priležena analitska prenosna funkcija MZI (oranžna), kjer v izračunu uporabimo vrednosti n_{eff}, tako da se analitični rezultat najbolj prilega meritvam.

lomni količnik n_{eff} in skupinski lomni količnik n_g sta skupaj z načrtovanimi oziroma analitskimi vrednostmi in

pripadajočimi odstopanji zapisana v tabeli 2. Obe vrednosti se nahajata znotraj intervala kotne analize. Kljub

lomni količnik	načrtovan	dejanski	Δ	%
n_{eff}	2,4453	2,4537	0,0084	0,34
n_g	4,2044	4,1819	-0,0225	-0,54

Tabela 2: Tabela z efektivnim in skupinskim lomnim količnikom svetlovoda za TE rod. Poleg načrtovane (analitske) vrednosti je v tabeli še vrednost pridobljena z meritvami, nujna razlika in odstopanje v odstotkih.

zelo majhnim odstopanjem dejanskih vrednosti n_{eff} od načrtovanih, se karakteristika MZI premakne za $\sim$1 nm. To v območju okrog 1550 nm pomeni premik za $\sim$125 GHz, kar je za optično obdelavo mikrovalovnih signalov nedopustno odstopanje. Pomembna tehnična ovira optičnih čipov je torej fino nastavljanje, kjer se poleg rešitve z grelci odpira dodatno področje raziskovanja.

5 Zaključek

Na primeru Mach-Zehnderjevega interferometra na izdelanem fotonskem čipu smo določili dejanska lomna količnika svetlovoda n_{eff} in n_g. Zaradi toleranc pri izdelavi je dejanska lomna količnika odstopata za 0,34% (n_{eff}) in -0,54% (n_g), kar je znotraj intervala kotne analize, a vseeno predaleč za nekatere aplikacije. Rezultati kažejo na potrebo po nastavljanju lastnosti takega čipa.

Zahvala

Andrej Lavrič acknowledges the edX UBCx Phot1x Silicon Photonics Design, Fabrication and Data Analysis course, which is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Silicon Electronic-Photonic Integrated Circuits (SiEPIC) Program. The devices were fabricated by Richard Bojko at the University of Washington Washington Nanofabrication Facility, part of the National Science Foundation's National Nanotechnology Infrastructure Network (NNIN), and Cameron Horvath at Applied Nanotools, Inc. Enxiao Luan performed the measurements at The University of British Columbia. Andrej Lavrič acknowledges Lumerical Solutions, Inc., Mathworks, Mentor Graphics, Python, and KLayout for the design software.

Avtorji se zahvaljujejo tudi Javni agenciji za raziskovalno dejavnost Republike Slovenije za financiranje raziskav v okviru programa št. P2-0246.

Literatura

[1] B. Batagelj, V. Janyani, and S. Tomažič, "Research challenges in optical communications towards 2020 and beyond = raziskovalni izzivi optičnih komunikacij do in preko leta 2020," *Informacije MIDEM : časopis za mikroelektroniko, elektronske sestavne dele in materiale*, vol. letn. 44, no. št. 3, pp. str. 177–184, 2014.

[2] B. Jalali and S. Fathpour, "Silicon Photonics," *Journal of Lightwave Technology*, vol. 24, no. 12, pp. 4600–4615, Dec. 2006.

[3] Z. Fang and C. Z. Zhao, "Recent Progress in Silicon Photonics: A Review," *ISRN Optics*, Mar. 2012.

[4] W. Zhang and J. Yao, "Silicon-Based Integrated Microwave Photonics," *IEEE Journal of Quantum Electronics*, vol. 52, no. 1, pp. 1–12, Jan. 2016.

[5] Y. Wang, X. Wang, J. Flueckiger, H. Yun, W. Shi, R. Bojko, N. A. F. Jaeger, and L. Chrostowski, "Focusing sub-wavelength grating couplers with low back reflections for rapid prototyping of silicon photonic circuits," *Opt. Express*, vol. 22, no. 17, pp. 20 652–20 662, Aug 2014.

[6] Y. Wang, H. Yun, Z. Lu, R. Bojko, W. Shi, X. Wang, J. Flueckiger, F. Zhang, M. Caverley, N. A. F. Jaeger, and L. Chrostowski, "Apodized focusing fully etched subwavelength grating couplers," *IEEE Photonics Journal*, vol. 7, no. 3, pp. 1–10, June 2015.

[7] Y. Wang, W. Shi, X. Wang, Z. Lu, M. Caverley, R. Bojko, L. Chrostowski, and N. A. F. Jaeger, "Design of broadband subwavelength grating couplers with low back reflection," *Opt. Lett.*, vol. 40, no. 20, pp. 4647–4650, Oct 2015.

[8] Y. Wang, L. Xu, A. Kumar, Y. D'Mello, D. Patel, Z. Xing, R. Li, M. G. Saber, E. El-Fiky, and D. V. Plant, "Compact single-etched sub-wavelength grating couplers for o-band application," *Opt. Express*, vol. 25, no. 24, pp. 30 582–30 590, Nov 2017.

[9] Y. Zhang, S. Yang, A. E.-J. Lim, G.-Q. Lo, C. Galland, T. Baehr-Jones, and M. Hochberg, "A compact and low loss y-junction for submicron silicon waveguide," *Opt. Express*, vol. 21, no. 1, pp. 1310–1316, Jan 2013.

[10] L. Chrostowski, Z. Lu, J. Flueckiger, X. Wang, J. Klein, A. Liu, J. Jhoja, and J. Pond, "Design and simulation of silicon photonic schematics and layouts," in *Silicon Photonics and Photonic Integrated Circuits V*, L. Vivien, L. Pavesi, and S. Pelli, Eds., vol. 9891, International Society for Optics and Photonics. SPIE, 2016, pp. 185 – 195.

[11] R. J. Bojko, J. Li, L. He, T. Baehr-Jones, M. Hochberg, and Y. Aida, "Electron beam lithography writing strategies for low loss, high confinement silicon optical waveguides," *Journal of Vacuum Science & Technology B*, vol. 29, no. 6, p. 06F309, 2011.

[12] L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*. Cambridge University Press, 2015.

[13] W. Bogaerts, P. D. Heyn, T. V. Vaerenbergh, K. D. Vos, S. K. Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. V. Thourhout, and R. Baets, "Silicon microring resonators," *Laser & Photonics Reviews*, vol. 6, no. 1, pp. 47–73, 2012.

[14] A. Debevc, J. Krč, and M. Topič, "Simulation of optical waveguides in silicon photonic integrated circuits," in *Conference proceedings 2018*. Ljubljana: MIDEM - Society for Microelectronics, Electronic Components and Materials, 2018, pp. str. 87–92.

[15] A. H. Atabaki, E. S. Hosseini, A. A. Eftekhar, S. Yegnanarayanan, and A. Adibi, "Optimization of metallic microheaters for high-speed reconfigurable silicon photonics," *Opt. Express*, vol. 18, no. 17, pp. 18 312–18 323, Aug 2010.

[16] F. Gan, T. Barwicz, M. A. Popovic, M. S. Dahlem, C. W. Holzwarth, P. T. Rakich, H. I. Smith, E. P. Ippen, and F. X. Kartner, "Maximizing the thermo-optic tuning range of silicon photonic structures," in *2007 Photonics in Switching*, 2007, pp. 67–68.

Spremljanje hitrosti in načina gibanja uporabnika multimedijske storitve z visokofrekvenčnimi radijskimi signali

Peter Nimac[1], Urška Levac[2], Nina Gartner[2], Jure Janez Markovič[1], Aleš Simončič[1], Rok Vidmar[1], Amadej Vidic[1], Jakob Gazič[1], Andrej Košir[1], Matija Svetina[2], Urban Burnik[1]

[1] *Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana*
[2] *Univerza v Ljubljani, Filozofska fakulteta, Aškrčeva 2, 1000 Ljubljana*
E-pošta: nimac.peter@protonmail.com

Motion and movement tracking of multimedia service users with super high frequency radio signals

Abstract. *Using the software defined radio technology, we have created a system users can use to receive a recommended route and mode of transport to desired destination. It works on the continuous-wave Doppler radar principle, to measure user's speed of approach to the device. This is used to determine whether the user is in a hurry. The main reason for this approach is to avoid using the camera and thus increase user's privacy and and speed measurement accuracy. We use speech recognition and synthesis in Slovene for human-machine interaction. To create such a system, we reviewed previous psychological studies that link people in a state hurry with those that are under various degrees of stress. From those observations, we designed two experiments in which we tried to provoke users in a state of hurry.*

1 Uvod

Spremljanje telesne aktivnosti uporabnikov v zadnjem obdobju postaja predmet številnih raziskav in praktičnih aplikacij. Zbrane podatke uporabljamo na številnih področjih od spremljanja rekreativnih dejavnosti, zdravstvenega stanja, identifikacije ogroženosti starejših oseb (padec), razpoznave trenutnega konteksta uporabnika, varčevanju z energijo in v številnih drugih aplikacijah informacijsko-komunikacijske družbe. Podatki o gibanju in kretnjah testnih uporabnikov so lahko tudi pomemben vir informacije v podporo starostnikom ali v analizi medijske izpostavljenosti testnih uporabnikov multimedijskih storitev.

Običajno potrebne podatke zbiramo s pomočjo senzorjev gibanja, ki so na posameznega uporabnika pritrjeni, kot na primer s pametnimi zapestnicami ali s senzorskimi oblačili. Težave pri uporabi nastopijo zaradi potrebe po zanesljivem pritrjevanju senzorjev, pri prenosu podatkov in pri zagotavljanju energije za napajanje merilnih naprav. Te težave odpravljajo sistemi na osnovi računalniškega vida, ki pa pri uporabnikih vzbujajo nelagodje zaradi eksplicitne narave uporabe kamere v procesu zajema podatkov in s tem povezanega potencialnega vdora v zasebnost. Alternativo navedenim sistemom predstavljajo sistemi za spremljanje obnašanja opazovanih oseb na podlagi analize radijskega valovanja v prostoru. Tovrstni sistemi so se pričeli razvijati v zadnjem desetletju in v bližnji prihodnosti pričakujemo komercialne rešitve na tej osnovi.

Razvili smo prototipni sistem, ki na podlagi analize visokofrekvenčnih radijskih signalov na neinvaziven način zagotavlja temeljne podatke o gibanju uporabnikov multimedijske storitve. Poleg gibanja, sistem razpoznava socialni signal hitenja uporabnika. Poskrbeli smo za integracijo identificiranih podatkov v večmodalni komunikacijski sistem s situacijskim zavedanjem. Podatki o gibanju in hitenju uporabnika dopolnjujejo obstoječe vire, kot so samodejna identifikacija konteksta uporabe, pogojev izvajanja storitve (trenutno vreme) in temeljno detekcijo prisotnosti uporabnika pred medijskim predvajalnikom.

2 Metodologija

2.1 Programski radio

Radijske postaje so večinoma narejene povsem elektronsko. Če si v mislih celotno radijsko postajo razdelimo na posamezne bloke oz. sestavne dele, bi to bila razna frekvenčna sita, oscilator, mešalnik in ojačevalniki. Vsak od teh sestavnih delov radijske postaje je narejen elektronsko. Radijsko postajo pa lahko izdelamo tudi z uporabo programskega radia oziroma programsko določenega radia (ang. *SDR – Software Defined Radio*). S programskim radiem za opis delovanja posameznih blokov ne potrebujemo električnega vezja ampak jih opišemo s funkcijami v poljubnem programskem jeziku. Edina strojna oprema, ki jo v osnovi potrebujemo za delo s programskim radiem, so frekvenčni uglaševalnik, analogno-digitalni pretvornik in digitalno-analogni pretvornik. Ta enostavnost strojne opreme pri delu s programskim radiem omogoča vsestranskost uporabe za sprejemanje ali oddajanje poljubnega signala. Signal namreč obdelamo na računalniku kot diskreten signal z uporabo diskretne matematike.

2.2 USRP B210

USRP (krajše za *Universal Software Radio Peripheral*) je v času pisanja ena izmed najbolj popularnih družin naprav programskih radiev. Gre za visokozmogljive, večnamenske programske radie, ki so večinoma grajeni modularno.

Model B210 pripada seriji radiev, pri katerih komunikacija z osebnim računalnikom poteka prek serij-

skega vodila USB. B210 pokriva frekvenčno območje od 70 MHz do 6 GHz, omogoča dvo-kanalno komunikacijo (2x RX in 2x TX) v polnem dupleks načinu z možnostjo uporabe večpotja (ang. *MIMO Multiple-In Multiple-Out*). Pasovna širina prenosa podatkov med radiem in računalnikom znaša 56 MHz (v obe smeri) prek serijskega vodila USB 3.0.

Antene, ki smo jih uporabili pri izgradnji radarja so monopolne antene, ki so prilagojene za frekvenčno območje 5 GHz signala WiFi po standardu IEEE 802.11ac. Na postajo USRP so pritrjene s konektorji SMA.

2.3 GNU Radio

Pri uporabi programskega radia obdelavo signalov izvajamo z računalnikom. Za to delo smo uporabili GNU Radio, ki se uporablja predvsem na področju SDR. GNU Radio pa je uporaben tudi za splošno signalno analizo.

Princip dela s tem orodjem si lahko predstavljamo kot risanje poti, po kateri bo potoval signal. To pot imenujemo signalna pot (ang. *flow graph*) in jo izrišemo oziroma sestavimo iz signalnih blokov. Ti bloki predstavljajo frekvenčna sita, signalne ojačevalnike, številčne pretvornike itd. Na začetku signalne poti uporabimo posebne izvorne bloke, iz katerih signal izvira, ter ponorne bloke na koncu, kjer se signalna pot konča.

2.4 Dopplerjev radar

Napravo za zaznavanje hitenja smo v tej fazi zgradili kot Dopplerjev radar. Dopplerjev radar je vrsta radarja, ki pri sprejemu odbitega signala uporablja zakonitosti Dopplerjevega pojava za določitev hitrosti premikajočega objekta. Hitrost premikajočega objekta lahko izračunamo po enačbi (1).

$$v = \frac{|f_0 - f_d|}{f_0}\, c_0 = \frac{\Delta f}{f_0}\, c_0 \tag{1}$$

Kjer je v hitrost premikajočega objekta, c_0 hitrost svetlobe v praznem prostoru, f_0 oddajna frekvenca radarja, f_d sprejeta oz. Dopplerjeva frekvenca in Δf razlika med Dopplerjevo in oddano frekvenco [1].

Dopplerjev radar smo zgradili kot radar CW (ang. *Continuous Wave radar*). Pri tej vrsti radarja tvorimo sinusni signal s frekvenco 500 kHz. Tega nato frekvenčno zamaknemo in oddajamo na frekvenci 5,25 GHz (slika 1). Ko se opazovan objekt (v našem primeru uporabnik) približuje radarju, se, v primerjavi z oddajno frekvenco, Dopplerjeva frekvenca dvigne. V primeru oddaljevanja pa se ta spusti. Frekvenco sprejetega signala izračunamo z algoritmom FFT (ang. *Fast Fourier Transform*), ki računa z vektorskimi bloki signala v velikosti 4096 elementov v frekvenčnem oknu širine $B_{FFT} = 2\,\text{MHz}$. Signal pred Fourierevo transformacijo še decimiramo za faktor 128, da izboljšamo frekvenčno ločljivost f_r (2). Frekvenčna ločljivost po obdelavi znaša $f_r = 3,8\,\text{Hz}$, kar da hitrostno ločljivosti $\Delta v = 0,11\,\text{m/s}$ (3). Rezultat po Fourierevi transformaciji je vektor Dopplerjevih frekvenc.

$$f_r = \frac{B_{FFT}}{128 \cdot 4096} \tag{2}$$

$$\Delta v = \frac{f_r}{2\, f_0}\, c_0 \tag{3}$$

Pythonsko kodo, ki se ustvari na podlagi signalne poti v GNU Radiu, smo dopolnili z enostavno klasifikacijo rezultatov. Predznak Dopplerjeve frekvence poda informacijo o približevanju oz. oddaljevanju uporabnika. Absolutna vrednost Dopplerjeve frekvence pa informacijo o hitenju oz. ne-hitenju. Rezultate po klasifikaciji strnemo v dvobitno število, kjer prvi bit predstavlja približevanje oz. oddaljevanje, drugi bit pa hitenje oz. ne-hitnje uporabnika v danem časovnem trenutku. Ta dvobitna števila pa prek protokola MQTT (ang. *Message Queuing Telemetry Transport*) pošljemo v podatkovno bazo do katere dostopa komunikator.

2.5 Priporočilni sistem za izbiro poti do želenega cilja

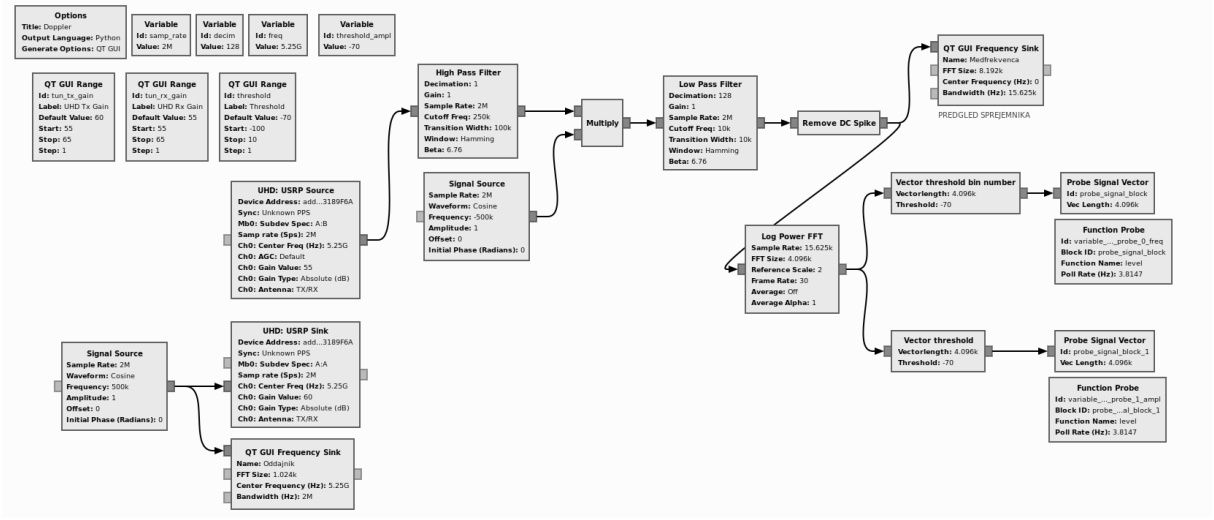
Podatke, ki jih posreduje radar, prevzame glasovni komunikator. Ta deluje po podobnem načelu kot Amazon Alexa, Google Assistant idr. Ti za začetek komunikacije uporabljajo prožilno besedo (frazo) kot je na primer *"Alexa!"* Ali *"Hey Google!"* Mi smo namesto prožilne besede uporabili podatke, ki jih posreduje radar saj z njimi dobimo tudi del socialnega konteksta, torej informacijo o hitenju.

Komunikator je zgrajen na mikroračunalniku Raspberry Pi, saj ta omogoča enostaven dostop do interneta, oblačnih storitev in podpira uporabo zvočnih uporabniških vmesnikov. Ker za slovenščino še ne obstajajo celovite rešitve za razpoznavo in sintezo govora smo to rešili z ločenima tehnologijama. Za razpoznavo govora uporabljamo Googlove oblačne storitve, za sintezo pa oblačne storitve Microsoft Azure. Vremenske podatke pridobivamo od Agencije Republike Slovenije za okolje, za navigacijo pa uporabljamo navigacijske storitve podjetja HERE.

Ko komunikator zazna, da se mu je uporabnik približal, ga pozdravi in vpraša o njegovi destinaciji. Destinacija, ki jo uporabnik sporoči komunikatorju je lahko ali naslov ali pa približna lokacija - to navigacijski sistem nato pretvori v koordinate. Komunikator stalno razpolaga s podatki o vremenu, ki nudijo dodatne informacije, na podlagi katerih izpopolni razumevanje trenutnega socialnega konteksta uporabnika. Komunikator nato, glede na pridobljen socialni kontekst, uporabniku priporoči uporabo specifičnih linij mestnega avtobusa ali pa sprehod do želene destinacije.

2.6 Merjenje hitenja

Psihološki del raziskovanja je bil usmerjen v prepoznavanje vidnih znakov hitenja. Hitenje trenutno še ni dobro raziskano področje. Raziskave na tem področju večinoma potekajo preko časovno omejenih eksperimentov, ki prisilijo ljudi, da pri opravljanju določene naloge hitijo. Ljudje so v časovno omejenih pogojih pod stresom, zato smo predvidevali, da bodo znaki hitenja podobni znakom, ki kažejo na stres. Pri raziskovanju literature smo se tako usmerili tudi na vidne znake stresa.



Slika 1: Signalna pot Dopplerjevega radarja v GNU Radiu

Raziskave so pokazale, da so udeleženci, ki so bili časovno omejeni, čutili več aktivacije, bili bolj vzburjeni in bili pod večjim stresom [2]. Pospešeno je bilo dihanje, povečan premer zenice, zmanjšalo pa se je mežikanje [3]. Povišani frustraciji in vznemirjenosti sledi negativno razpoloženje kot je jeza. Stanje jeze pa je povezano z vedenjem, ki je vidno navzven in s fiziološkimi odzivi. Prepozna se lahko z nesramnimi gestami, kletvicami in kričanjem. Med stresom in fiziološkim odzivom obstaja močna povezava. Visoke ravni stresa med drugim vplivajo tudi na porast krvnega tlaka in višji srčni utrip. Raziskava o vedenju pešcev pri prečkanju ceste je pokazala, da ni bilo razlik med spoloma in glede na starost, pri vseh udeležencih je bilo vedenje pod časovnim pritiskom bolj tvegano. Pri hitenju se ljudje hitreje gibajo [4]. Študija o iskanju znakov preobremenitve je ugotovila, da so simptomi in vedenjski znaki, ki se kažejo pri stresu med drugim izguba fokusa, nepotrpežljivost, težave s sprejemanjem odločitev in tresenje telesa [5].

Za merjenje hitenja, kjer smo želeli preveriti delovanje programskega radia in komunikatorja v ekološko veljavnih situacijah, smo izvedli dva eksperimenta. Ekološka veljavnost je predstavljala velik izziv, saj je avtentično hitenje težko izzvati oziroma zaigrati. Eksperimente smo zasnovali stopenjsko, pri čemer smo upoštevali predvidene zmožnosti programske opreme v danem trenutku. Glede na realen razvoj, smo nato eksperimente prilagajali potrebam in trenutnemu raziskovalnemu problemu.

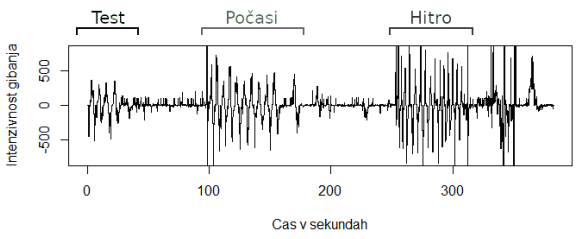
S prvim pilotnim eksperimentom smo želeli identificirati opazljive znake hitenja, ki bi služili kot izhodiščna točka za pripravo primerne opreme na merjenje.
Udeleženci: dva moška in dve ženski (M = 22,8).
Pripomočki: kamera, ključi, denarnica, telefon, jakna, torba, štoparica.
Postopek: Testno situacijo sta sestavljala dva pogoja, brez časovne omejitve in časovna omejenost na 1 minuto. Vsak udeleženec je sodeloval v obeh pogojih, vendar je vrstni red pogojev med udeleženci variiral zaradi možnega vpliva vrstnega reda pogojev. Testni prostor je predstavljala dnevna soba v kateri so bili skriti predmeti (denarnica, ključ in telefon). Naloga udeležencev je bila, da v dnevni sobi najdejo skrite predmete, jih pospravijo v torbo, se oblečejo in odidejo iz prostora. Med eksperimentom je bila z udeležencem v prostoru oseba, ki ga je snemala.
Analiza rezultatov: Zbrane posnetke smo analizirali in jih poslali 10 zunanjim opazovalcem za višjo objektivnost rezultatov. Ti so pregledali neme posnetke udeležencev in odgovarjali na podana vprašanja: Opazuj in opiši vedenje osebe na obeh posnetkih. Katere so razlike v vedenju (obrazni izrazi, gibanje, neverbalna komunikacija)? Zaradi česa je po vašem mnenju prišlo do teh razlik?
Rezultati: Podane odgovore smo semantično uredili in izdelali kategorije opaznih znakov hitenja. V vsako od štirih glavnih kategorij spada več značilnih znakov hitenja: Gibanje telesa (večja intenzivnost gibanja, hitrejše gibanje); obrazni izrazi (namrščeno čelo, beganje z očmi); čustva (jeza, obup, skrb, zmedenost); vedenje (manjša natančnost, nepremišljenost in brezbrižnost).

Drugi eksperiment je potekal od doma in na fakulteti. Od doma smo izvedli pilotno študijo drugega eksperimenta, v kateri sta bila udeležena moški in ženska, stara 23 in 24 let. Študija je bila namenjena preizkušanju delovanja opreme, ustreznosti navodil ter oblikovanja prostora za eksperiment. V nadaljevanju smo z eksperimentom nadaljevali na fakulteti. Želeli smo pridobiti informacije o razliki v gibanju osebe, ki se ji mudi in osebe, ki se ji ne mudi. Iz pridobljenih podatkov smo želeli najti značilne vzorce in parametre prek katerih bi programski radio lahko klasificiral ljudi v ti dve skupini.
Udeleženci: 11 moških in 1 ženska, stari od 21 do 54 let (M = 26,4).
Pripomočki: Kinect, programski radio, 15 plastičnih lončkov, štoparica.
Postopek: Testno situacijo sta ponovno sestavljala dva pogoja (brez časovne omejitve in časovna omejenost na 1 minuto). Vsak udeleženec je sodeloval v obeh pogojih, vendar je vrstni red pogojev med udeleženci variiral.

Na eni strani prostora je na mizi postavljenih 15 lončkov, na drugi strani pa je označen prostor za zidanje piramide (nizka mizica). Kinect in programski radio (radar) snemata prostor iz sprednje strani, tako da se udeleženec neposredno približuje k radarju oziroma oddaljuje stran od njega. Udeleženec se giblje v ravni liniji od prostora z lončki, do prostora za postavljanje piramide, nad katerim sta Kinect in programski radio. Vsako nalogo je udeleženec začel opravljati pri mizi z lončki. Najprej je 30 sekund hodil v ravni liniji z običajnim tempom, kar nam bo služilo kot kontrola hitrosti. Možno je namreč, da udeleženec začne hiteti tudi v pogoju brez časovne omejitve. Nato so udeleženci dobili navodilo, naj iz lončkov zgradijo piramido. Začnejo na strani kjer se lončki trenutno nahajajo, naenkrat lahko do mizice nesejo le en lonček, piramido pa morajo sestavljati sproti. V pogoju s časovno omejitvijo so dobili še dodatno navodilo, naj bodo čim hitrejši, saj imajo časa le eno minuto.

3 Rezultati

S pomočjo drugega eksperimenta smo uspešno pridobili podatke meritev iz obeh trenutno uporabljenih senzorjev, programskega radia in kamere Kinect. Iz podatkov, pridobljenimi s programskim radiem, je jasno razvidno med stanji hitenja in ne-hitenja uporabnika, ko se ta približuje in oddaljuje od senzorjev (slika 2). Intenzivnost hitenja se odraža v amplitudi Dopplerjevega odziva, sama hitrost gibanja pa se odraža v frekvenci Dopplerjevega odziva. Pri vseh udeležencih sta bila opažena povečana intenzivnost in hitrost gibanja v situacijah, kjer smo izzvali hitenje. Kakor smo ugotovili že iz posnetkov prvega eksperimenta. Sočasno pridobljeni posnetki iz Kinecta bodo služili pri pripravi boljše klasifikacije med obema stanjema približevanja ter bodo uporabljeni v nadaljevanju raziskave. Ti bodo uporabni predvsem pri identifikaciji morebitnih nevidnih znakov hitenja.



Slika 2: Graf Dopplerjevega odziva zaradi premikanja uporabnika pred radarjem

4 Razprava

Ta pristop trenutno omogoča le grobo klasifikacijo med hitenjem in ne-hitenjem. Pridobljeni rezultati imajo le omejeno veljavnost. V vseh primerih so se uporabniki napravi približevali neposredno, v ravni liniji. Za Dopplerjev radar je to sicer optimalen kot približevanja ampak v realnem okolju se bodo uporabniki napravi približevali pod različnimi koti. Prav tako je vzorec uporabnikov statistično majhen niti ni dobro uravnotežen po spolu in starosti uporabnikov.

V nadaljevanju raziskave načrtujemo eksperiment, kjer bo komunikator postavljen v živo okolje kot je na primer avla fakultete. V takem okolju bomo lahko zbrali ekološko bolj veljavne rezultate. Tokrat brez uporabe kamere Kinect. Komunikator bo na voljo za uporabo kot priporočilni sistem za načrtovanje poti, hkrati pa bo zbiral podatke o hitenju uporabnikov. Te podatke pa bomo s tehnikami strojnega učenja uporabili za natančnejšo klasifikacijo med uporabniki.

Zahvala

Rezultati tega dela so nastali v okviru projekta PKP 'Spremljanje lege, gibanja in karakterističnih kretenj uporabnika multimedijske storitve z visokofrekvenčnimi radijskimi signali'. Projekt sta sofinancirala Republika Slovenija in Evropska unija iz Evropskega socialnega sklada.

5 Zaključek

Hitenje je psihološko mnogo zahtevnejše za analizo in ga kot tako ne moremo določiti le na podlagi hitrosti približevanja uporabnika k napravi. Na trgu se trenutno pojavljajo rešitve kot so senzorji, ki operirajo z radijskimi valovi v milimetrskem valovnem področju. Ti omogočajo tudi natančno določanje oddaljenosti uporabnika od naprave, kot približevanja in tudi merjenje biometričnih znakov (kot so hitrost dihanja in srčni utrip). Omogočajo pa tudi zaznavanje kretenj. Vse to nam daje dodaten vpogled in informacije, s katerimi lahko naredimo še boljšo klasifikacijo hitenja.

Literatura

[1] L. N. Ridenour, *Radar system engineering*. New York: McGraw-Hill Book Co, 1947. OCLC: 1104722440.

[2] C. Oliveras, M. Cunill, M. Gras, M. Sullman, M. Pedra, and C. Figner, "Effects of time pressure on feelings of stress, activation and arousal, and drivers' risk taking behaviour," *Human Factors in Transportation*, pp. 245–248, 2002.

[3] E. Rendon-Velez, P. M. . van Leeuwen, R. Happee, I. Horváth, W. van der Vegte, and J. de Winter, "The effects of time pressure on driver performance and physiological activity: A driving simulator study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 41, pp. 150–169, Aug. 2016.

[4] B. A. Morrongiello, M. Corbett, J. Switzer, and T. Hall, "Using a Virtual Environment to Study Pedestrian Behaviors: How Does Time Pressure Affect Children's and Adults' Street Crossing Behaviors?," *J. Pediatr. Psychol.*, vol. 40, pp. 697–703, Aug. 2015.

[5] J. H. Amirkhan, I. Landa, and S. Huff, "Seeking signs of stress overload: Symptoms and behaviors.," *International Journal of Stress Management*, vol. 25, pp. 301–311, Aug. 2018.

Nizkocenovna avtomatizacija visokofrekvenčnih meritev

Peter Miklavčič, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: peter.miklavcic@fe.uni-lj.si

Low-cost automatization of radio-frequency measurements

In this paper a Python library in-development is presented that enables low-cost automatization of radio-frequency measurements using different instruments. The library supports instruments with omnipresent buses such as RS-232, USB and Ethernet/LXI. A cheap and practical solution for connecting legacy instruments with the GPIB bus is also discussed and identified. In the beginning, the background for the proposed solution is discussed and general software and hardware proposals are laid out. The proposed approach is then presented in detail, discussing the core library and its topics first and the developed measurement scripts are discussed next, followed by the postprocessing and graphing scripts. Remote and embedded execution are discussed as well.

1 Uvod

Avtomatizacija meritev v smiselni meri je pri raziskovalnem delu vedno dobrodošla. Poleg multiplikativnega prihranka časa pri ponavljajočih meritvah, jasna avtomatizacijska skripta omogoča boljšo ponovljivost in manj napak tako pri posameznih serijah meritev, kakor tudi pri znanstveni ponovljivosti poskusov na dolgi rok.

Radiofrekvenčne meritve pogosto zahtevajo zajem in obdelavo večje količine meritev in to iz različnih enokanalnih ali večkanalnih merilnih instrumentov (na primer visokofrekvenčnega merilnika moči, faznega detektorja, spektralnega analizatorja, ...). Profesionalne rešitve avtomatizacije takih meritev so poleg visoke cene programske in strojne opreme podvržene tudi nekompatibilnosti med programsko opremo različnih proizvajalcev ter različnimi namenskimi vodili.

Pri delu z najrazličnejšimi merilnimi instrumenti in pripomočki za potrebe meritev v visokofrekvenčni, optični in antenski tehniki imamo v laboratoriju potrebo po prilagodljivi, vsestranski in preprosti avtomatizaciji zajema, poobdelave ter prikaza meritev. V nadaljevanju je predstavljen predlog rešitve, naši dosedanji poskusi ter ugotovitve. Predlagana kombinacija programske in strojne opreme omogoča tekoče delo z različnimi merilnimi instrumenti, od profesionalnih komercialnih, do ročno izdelanih, kot so na primer nizkocenovni doma izdelani VF generatorji [1] in detektorji [2, 3] prof. dr. Matjaža Vidmarja. Z nekaj dodatne programske logike in obdelave meritev so taki eksperimentalni instrumenti primerni za grobe visokofrekvenčne meritve. Za meritve smernega diagrama antene se pri svojem delu v veliki meri zanašamo na ročno izdelane merilne vrtiljake, objavljene v [3]. Tudi v članku opisana programska oprema je primarno namenjena antenskim meritvam, tako da je bilo največ razvoja opravljenega v to smer.

2 Podatkovna vodila

Starejši merilni instrumenti tipično omogočajo krmiljenje naprave in branje izmerjenih vrednosti po 8-bitnem vzporednem vodilu IEEE 488, standardiziranem leta 1975. Vodilo je bilo razvito kakih 10 let prej pod imenom HP-IB (ang. *Hewlett-Packard Interface Bus*), po standardizaciji in splošni uporabi pa je postalo znano kot GPIB (ang. *General Purpose Interface Bus*). Danes GPIB nadomešča LXI (ang. *LAN eXtensions for instrumentation*), ki temelji na vodilu Ethernet, IP in protokolu IEEE 1588 za psevdo-realnočasovno uro, omogoča pa tudi pravo realnočasovno signalizacijo z uporabo dodatnih povezav. Pri napravah s podporo LXI je možno hitro in poceni avtomatizirati tudi kompleksnejšo meritev z več instrumenti, in to z uporabo le Ethernet stikala in nadzornega računalnika ali sorodne naprave ter ustrezne programske opreme. Večje težave so pri komercialno dostopnih vmesnikih med vodilom GPIB in modernimi vodili (na primer USB ali Ethernet), ki stanejo nekje od 100€ dalje, kar v določenih primerih lahko predstavlja previsok strošek. Poleg razmeroma visoke cene večina takih vmesnikov potrebuje namenske, včasih slabo podprte gonilnike, ki ne delujejo nujno na obeh glavnih družinah operacijskih sistemov, ki poganjata osebne računalnike in sorodne naprave.

Nekateri instrumenti omogočajo priklop preko drugih standardnih vodil, od katerih zagotovo velja omeniti RS-232 in USB. Pri teh vodilih za razliko od namenskih GPIB in LXI aplikacijska raven (protokol nadzora instrumenta in kodiranje podatkov) ni standardizirana, tako da nekateri instrumenti kljub uporabi standardnega vodila ne omogočajo preprostega in odprtega povezovanja z različno programsko opremo, pač pa v tem primeru zahtevajo (bolj ali manj dobro podprto) namensko programsko opremo z različnimi programskimi vmesniki in možnostmi za nadaljnjo avtomatizacijo meritev.

3 Programska oprema

Med poplavo komercialnih in odprtih programskih orodij ter množico programskih jezikov je bil s ciljem nizkocenovne rešitve izbran pristop z odprtokodnim tolmačitvenim programskim jezikom Python [4]. Ta z odprtim pristopom in široko (tudi akademsko) skupnostjo že leta doživlja hiter razvoj in ponuja vrsto kvalitetnih knjižnic za najrazličnejša opravila. Pri takih možnostih je v luči cilja znanstvene ponovljivosti danes verjetno smotrno izbrati tako ali primerljivo odprto okolje, kakor ga ponuja jezik Python.

4 Strojna oprema

Čeprav bi bilo zaradi znanstvene ponovljivosti verjetno prav, da bi bila odprtokodna strojna oprema v znanosti standard, je realnost toliko drugačna, da so bili kriteriji za primerne merilne instrumente (za razliko od rigoroznega pristopa pri izboru programske opreme) bolj pragmatično zastavljeni. Primerna naprava naj podpira vsaj enega izmed danes razširjenih vodil (RS-232, USB, Ethernet/LXI, in podobno), za naprave z vmesnikom GPIB pa se splača preveriti morebitne nizkocenovne in odprtokodne rešitve pretvornika. Naprava mora imeti dokumentirano ali vsaj znano komunikacijo na aplikacijski ravni. Na dolgi rok je smiselno stremeti k čim bolj odprtim in ponovljivim strojnim rešitvam, za dosego cilja nizkocenovnih ponovljivih meritev pa verjetno v smeri čim bolj univerzalnih merilnih instrumentov.

4.1 Povezava naprav z vmesnikom GPIB

Preizkušenih je bilo več različnih komercialnih in nekomercialnih rešitev pretvornikov za GPIB. Kot pomembna šibka točka so se pokazali že prej omenjeni slabo podprti in pomanjkljivo zasnovani gonilniki, ki onemogočajo zanesljivo delo. Tudi sicer pogosto uporabljena knjižnica PyVISA na primer ni polno podprta na kartičnih računalnikih s procesorji ARM. Kot robustna in preprosta rešitev so se izkazali vmesniki USB-GPIB, ki za povezavo na aplikacijski ravni uporabljajo navidezna serijska vrata. Obstajajo tudi odprtokodni projekti razvoja takih vmesnikov, od katerih se je dobro izkazal vmesnik AR488 [5], zasnovan na mikrokrmilniških razvojnih ploščah Arduino. Ponuja zmogljiv nizkocenoven vmesnik z odprtokodno strojno in programsko opremo.

5 Jedrna knjižnica

Pri razvoju predlagane programske rešitve so glavni cilji preprostost zasnove in uporabe, pravilna in ponovljiva uporaba okolja Python ter kompaktna in modularna programska koda, ki se ravno v pravi meri zanaša na zmogljive zunanje knjižnice. Rešitev naj ne vsebuje nepotrebnih funkcionalnosti, potrebne pa naj abstrahira na smiselnem nivoju za ponovljivo uporabo v različnih merilnih skriptah. Za večino opravil zadostuje tekstovni vmesnik (CLI, ang. *command-line interface*), ki je na voljo povsod, grafični vmesnik (GUI, ang. *graphical user interface*) pa se naj uporablja tam, kjer je to potrebno oziroma smiselno. Knjižnica je namenjena uporabi v okolju Python, zato se ji ni potrebno izolirati in razen kjer je to nujno potrebno ne poskuša ujeti napak, ampak se zanaša na Pythonov izpis in postopanje ob napakah.

Za dosego zadanega cilja je v razvoju nabor skript, od katerih glavna (z delovnim imenom `liblab.py`) vsebuje različne nizkonivojske in visokonivojske funkcije z namenom čim bolj praktične modularizacije programske (merilne) kode. Za shranjevanje meritev je bil izbran berljiv in vseprisoten format CSV (ang. *comma-separated values*), ki pri pravilni implementaciji dela s tekstovnimi datotekami ne predstavlja ozkega grla, poleg tega za delo s tem formatom obstaja standardna knjižnica `csv`. Tekom razvoja se je pokazalo tudi, da se ni smiselno omejevati na eno izmed obeh trenutno aktualnih različic jezika Python, ampak je programsko kodo dokaj trivialno napisati tako, da deluje z obema različicama, s tem pa je možno tudi izkoristiti prednosti posamezne. V praksi je pri pisanju programske kode za obe različici jezika potrebno najti način izpisa poljubnega niza, ki deluje enako pri obeh. V predstavljeni knjižnici je omenjeno rešeno z uporabo funkcij `stdout.write()` in `stdout.flush()`, za kateri je potrebno vključiti standardno knjižnico `sys`. Jedrna knjižnica `liblab.py` med drugim vsebuje:

1. Uporabljene matematične in fizikalne konstante.
2. Funkcije za izpis razhroščevalnih informacij.
3. Funkcije za delo s serijskimi vrati:
   - neblokirno (ang. *non-blocking*) branje serijskih vrat z medpomnilnikom delnih branj,
   - sprejem in obdelava identifikacijskih podatkov naprave, poslanih po serijski povezavi (detekcija tipa in/ali posamezne naprave),
   - določitev unikatnega identifikacijskega niza (sestavljenega iz VID, PID, serijske številke in poslanih identifikacijskih podatkov),
   - izpis serijskih vrat in njihovih lastnosti (namenjen razhroščevanju).
4. Splošne funkcije za delo z merilnimi instrumenti:
   - iskanje želenih naprav med zaznanimi serijskimi vrati,
   - vzporedno čakanje med koncem strojnih in začetkom programskih inicializacij naprav,
   - branje medpomnilnika delnih branj,
   - vpis v medpomnilnik posameznega kanala meritev (s poljubnim povprečenjem posameznega kanala),
   - osnovna obdelava vpisanih meritev (preverjanje območja, normalizacija vrednosti, detekcija ovojnice - minimalnih in maksimalnih vrednosti),
   - programska zaključitev (zapiranje odprtih serijskih vrat in povezav).
5. Namenske funkcije za delo z merilnimi instrumenti:
   - razčlenjevanje vhodnih parametrov meritev (glede na tip instrumenta),
   - inicializacija strojnega in programskega nivoja (glede na tip instrumenta),
   - filtriranje (s pomočjo knjižnice `re`) in razčlenjevanje prebranih vrstic v koristne meritve (glede na tip instrumenta),

- nastavljanje uporabljenih visokofrekvenčnih izvorov in sintetizatorjev,
- nadzor frekvenčnega preleta VF generatorja,
- nastavljanje uporabljenih merilnih vrtiljakov,
- nadzor vrtenja in korakanje koračnega vrtiljaka,
- splošna omejitev hitrosti zajema.

6. Obdelava meritev (anten):
   - izračun veličin, potrebnih za uspešne meritve anten (impedančna prilagoditev, Fraunhoferjeva razdalja, premer prve Fresnelove cone, izgube zaradi razširjanja v praznem prostoru),
   - analiza smernega diagrama ter integracija smernosti.
7. Funkcije za delo z datotekami CSV:
   - branje glave in razčlenjevanje stolpcev,
   - tekstovni vmesnik za izbor želenih stolpcev.
8. Funkcije za generiranje zvoka (samo v Python 3):
   - vključitev zunanje knjižnice `simpleaudio`,
   - (ne)blokirno igranje poljubnega tona,
   - opozorilni pisk.

Ob izvršitvi skripte `liblab.py` je na voljo nekaj osnovnih razhroščevalnih funkcij: izpis serijskih vrat in zaznanih naprav, poskusno branje podprtih instrumentov, preizkus zvoka, sicer pa je skripta (knjižnica) namenjena vključitvi v *višjenivojske* merilne skripte. Pri idealno modularizirani kodi te tako vsebujejo le za posamezno meritev pomembno programsko logiko.

5.1 Vhodni podatki in nastavitve

Jedrna knjižnica in pripadajoče merilne skripte za svoje delovanje potrebujejo določene podatke, kot na primer identifikacijske nize vseh instrumentov, parametre vodil in povezav, privzete nastavitve instrumentov (na primer moč izvora, hitrost vrtenja vrtiljaka, ...), seznam merilnih kanalov posamezne naprave s pripadajočimi merilnimi enotami in območji, in podobno. Našteti podatki so zapisani v datoteki `config.py`, ki je namenjena vključitvi s strani jedrne knjižnice. Na ta način so uporabniški podatki zelo preprosto ločeni od programske kode. Pri zagonu katerekoli merilne skripte uporabnik poda le še obvezne (in neobvezne) vhodne parametre posamezne skripte, ki jih skripte nato razčlenijo s standardno Python knjižnico `argparse`.

6 Zbiranje meritev

Zaenkrat je bilo razvitih le nekaj osnovnejših merilnih skript, ki pa s svojim nastavljivim pristopom in modularizirano kodo omogočajo že kar nekaj različnih merilnih scenarijev v visokofrekvenčni tehniki.

6.1 Časovni prelet: `recorder.py`

Merilna skripta `recorder.py` je namenjena dolgotrajnemu zbiranju meritev iz različnih merilnikov, amplitudnih in faznih detektorjev, dodatno pa omogoča še zajem podatkov iz senzorja Bosch Sensortec BME280, ki zna meriti temperaturo okolice, relativno zračno vlažnost in zračni pritisk. Senzor je na računalnik moč priklopiti z uporabo katere izmed mikrokrmilniških razvojnih plošč [6] in preprostim programom, ki prebrane vrednosti pošilja po (navidezni) serijski povezavi po vodilu USB.

Skripta `recorder.py` ima z namenom dolgoročne stabilnosti preprosto zasnovo in uporablja POSIX funkcijo `os.fork()` za dvonitno delovanje, s katerim usklajuje zajem meritev iz različno hitrih instrumentov. Narejena je kot avtomat končnih stanj (ang. *finite-state machine*). V primarni niti programa se vrši branje vseh instrumentov, razen najbolj počasnega, tekstovni vmesnik, pisanje izhodnih datotek, morebitna omejitev hitrosti zajema in komunikacija s sekundarno nitjo (za kar je uporabljena knjižnica `mmap`). V sekundarni niti se vrši branje najpočasnejšega instrumenta, ki po končanem branju z medprocesno komunikacijo da takt vpisa v izhodno datoteko primarni niti. Skripta se izvaja dokler uporabnik ne pošlje prekinitve (ujete s pomočjo standardne knjižnice `signal`), nakar zaključi izhodno datoteko in zapre povezave do instrumentov. Omogoča tudi nastavljivo povprečenje posameznih kanalov, zajete meritve pa poskuša časovno čim bolj poravnati in jih po potrebi označi z milisekundno uro.

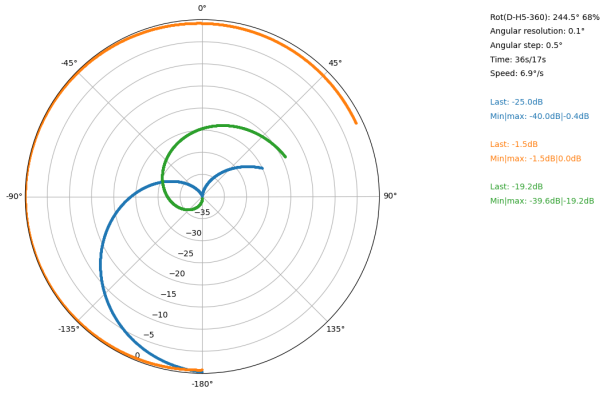
6.2 Frekvenčni prelet: `sweeper.py`

Skripta `sweeper.py` je namenjena zajemu meritev pri diskretnem frekvenčnem in/ali amplitudnem preletu nastavljivega visokofrekvenčnega generatorja. Omogoča zajem meritev individualno izbranih in povprečenih merilnih kanalov v vsaki točki preleta in jih zapiše v izhodno datoteko, med izvajanjem pa izpisuje stanje preleta. Tekom preleta spremlja stanje VF izvora in opozarja na nedosegljive točke merjenja izven frekvenčno-amplitudnega območja generatorja, preverja pa tudi merilno območje izmerjenih vrednosti posameznih merilnih kanalov.

6.3 Prostorski prelet: `rotator.py`

Skripta `rotator.py` je namenjena meritvam z merilnim vrtiljakom. V našem primeru gre največkrat za meritve smernega diagrama antene, katerim je skripta primarno namenjena. Glavna zanka programa upravlja in spremlja vrtenje vrtiljaka, povezanega preko vodila USB. Omogoča nastavljiv korak in sektor preleta. Med vrtenjem skripta zbira želene meritve (kanali se izbirajo enako kot pri skripti `sweeper.py`) in po potrebi nastavlja tudi visokofrekvenčni izvor ali ustavi vrtiljak (na primer za opravljanje frekvenčnega preleta). Poleg zajema meritev v izhodno datoteko in tekstovnega izpisa stanja, pa `rotator.py` za razliko od drugih opisanih skript, vsebuje tudi grafični vmesnik s polarnim izrisom za lažje spremljanje meritve v živo (prikazan na sliki 1) ter zvočno opozorilo ob začetku in koncu meritve.

Grafični vmesnik poleg spremljanja izmerjenih vrednosti izbranih kanalov (podobno kot tekstovni) izpisuje nastavitve in stanje vrtiljaka, pretečeni in preostali čas ter kotno hitrost vrtenja. Na koncu meritve prikaže na maksimalno vrednost poravnan snop in rezultate analize izmerjenega snopa. Ti obsegajo širino glavnega snopa (v primeru usmerjene antene), odklon maksimalne vrednosti od sredine izbranega sektorja in smernost izmerjenega smernega diagrama.



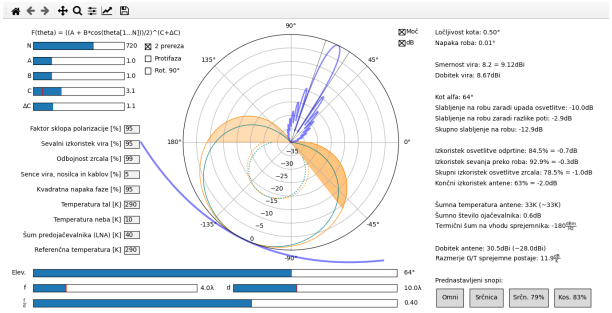
Slika 1: Grafični prikaz meritve s skripto `rotator.py`.

7 Obdelava in prikaz meritev

Knjižnica `liblab.py` v osnovi ni namenjena poobdelavi in prikazu meritev, saj v ta namen obstajajo primerni ter zmogljivi odprtokodni programi, kot je na primer LibreOffice Calc. Prav tako v ta namen jezik Python že ponuja različne zmogljive knjižnice, kot je na primer Matplotlib. Ta omogoča tudi risanje kvalitetnih polarnih grafov, kar se v praksi pokaže kot šibkost drugih odprtokodnih rešitev. Ker se v praksi kaže potreba po programski opremi za risanje polarnih grafov, je bil del razvoja opravljen tudi v to smer. Skripta `csvplot.py` (podobno kot grafični način skripte `rotator.py`) omogoča izpis vrednosti analize izmerjenih smernih diagramov ter polarni izris poljubnih meritev z uporabo knjižnice Matplotlib. Dodatno omogoča tudi grobo in fino rotacijo prikazanih smernih diagramov ter polarni kurzor za hitro odčitavanje meritev.

V drugem načinu delovanja skripta `csvplot.py` izriše kartezični graf, namenjen hitremu vpogledu v daljše datoteke CSV (na primer pri dolgotrajnem zajemu s skripto `recorder.py`). Pri tem vsak kanal (stolpec) izriše na skupni ali svoji ordinatni osi in le-te označi z glavo posameznega stolpca v datoteki CSV. Posebnost tega načina je možnost delovanja s hitro knjižnico za branje tekstovnih datotek `cStringIO`, kar omogoča hiter izris (nekaj sekund za več milijonov vrstic).

Poleg zajema, obdelave in prikaza meritev, je bila za potrebe dela na doktorski nalogi razvita tudi skripta za naprednejšo analizo smernega diagrama kot primarnega vira za osvetlitev simetrične parabolične antene za satelitske komunikacije. Skripta `illumer.py` zaenkrat omogoča branje, sintezo in prikaz smernega diagrama vira, poleg tega shematsko prikaže tudi zrcalo glede na nastavljene vhodne parametre (goriščna razdalja, premer zrcala, elevacija nad obzorjem) in približen izračun končnega smernega diagrama zrcala. Vmesnik je prikazan na sliki 2. Upošteva tudi vhodne izkoristke, faktorje in temperature za izračun šumnih veličin. Poleg analize smernega diagrama vira in zrcala zna izračunati tudi slabljenje na robu zrcala, izkoristek osvetlitve zrcala, šumno temperaturo antene, spektralno gostoto moči sprejetega termičnega šuma in razmerje G/T (dobitek antene proti šumni temperaturi sprejemnega sistema).



Slika 2: Grafični vmesnik programa za analizo osvetlitve zrcalne antene (`illumer.py`).

8 Oddaljen dostop

Ker je primarni uporabniški vmesnik razvitih skript ukazna vrstica, lahko knjižnica `liblab.py` in pripadajoče merilne skripte tečejo na najrazličnejših napravah. Tekstovni vmesnik hkrati omogoča trivialen oddaljen dostop in zagon na katerikoli primerni napravi, na primer s pomočjo SSH (ang. *secure shell*), ki tipično omogoča tudi oddaljeno kopiranje datotek s SCP (ang. *secure copy*). Možnost poganjanja avtomatizirane meritve iz manjše naprave, kot na primer kartičnega računalnika Raspberry Pi, omogoča dodatno svobodo pri postavitvi merilnih instrumentov in ne okupira uporabnikovega računalnika med zajemom meritev.

9 Zaključek

V tem delu je predstavljen koncept knjižnice za odprtokodno avtomatizacijo zajema ter obdelavo in prikaz meritev v radiofrekvenčni tehniki. Predstavljene so programske smernice in ugotovitve pri povezovanju merilnih instrumentov preko standardnih vodil kot na primer RS-232, USB, Ethernet/LXI in GPIB. Taka knjižnica skupaj s primeri merilnih skript je trenutno še v fazi razvoja, z objavo dosedanjih poskusov razvoja pa se preverja odziv znanstvene skupnosti, ki bo usmerjal nadaljnje delo.

Zahvala

Raziskovalni program št. P2-0246 je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

[1] M. Vidmar: Visokofrekvencni vir z ulomkovno zanko, 2017, http://lea.hamradio.si/~s53mv/fpll/fpll.html

[2] M. Vidmar: Lock-in sprejemnik z virom do 12GHz, 2017, http://lea.hamradio.si/~s53mv/lockin/lockin.html

[3] M. Vidmar: Amaterska antenska merilnica, 2017, http://lea.hamradio.si/~s53mv/aam/aam.html

[4] Python.org - Documentation, https://www.python.org/doc/

[5] AR488 Arduino GPIB Interface, https://github.com/Twilight-Logic/AR488/

[6] J. Ivković in J. L. Ivković: Analysis of the performance of the new generation of 32-bit Microcontrollers for IoT and Big Data Application, *Proc. of the International Conference on Information Society and Technology (ICIST)*, 2017

Sistemi za elektronsko bojevanje

Gal Likar, Boštjan Batagelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Laboratorij za sevanje in optiko, Tržaška cesta 25, 1000 Ljubljana, Slovenija

E-pošta: likar.gal.likar@gmail.com

Electronic warfare systems

Abstract. *This paper is an overview of the field of electronic warfare and it's different areas and systems used. It paints a broad picture of what this field of warfare covers while also pointing out a few typical and exciting technologies, some of which are in combat use today. The purpose of this paper is to show the field of electronic warfare, which needs a lot of attention, as technologies and methods of work are continually changing, and the obsolescence of systems can mean a significant weakness in conflict with the enemy.*

1 Uvod

Vojna spremlja človeka že vse od njegovega začetka. Z razvojem človeške družbe in njenim tehnološkim napredkom, so se vzporedno spreminjale tudi oblike vojskovanja. V vojnem konfliktu tehnologija predstavlja pomembno utež, ki lahko prevesi tehtnico razmerja sil v tisto stran, katera jo zna bolje uporabiti. Ob začetku uporabe elektromagnetnega valovanja se je odprla nova dimenzija vojskovanja. Če se je prej vojna bila na zemlji, vodi in v zraku, se je kasneje konflikt razširil še na področje elektromagnetnega spektra (EMS), ki je postal pomembna dobrina, uporaben sprva predvsem za komunikacije, pozneje pa tudi za radarske in namerilne sisteme. Tako kot je raba radijskih komunikacij in drugih naprav, ki uporabljajo elektromagnetni spekter, nepogrešljiva v civilnem svetu, lahko ohromitev takih sistemov povzroči zlom celotne vojaške organizacije, ki je svoje delovanje prilagodila uporabi tovrstnih modernih tehnologij.

2 Elektronski sistemi in elektronsko bojevanje

V NATO slovarju je termin elektronsko bojevanje definiran kot vojaška aktivnost, s katero raziskujemo elektromagnetni spekter in obsega iskanje, prestrezanje, identifikacijo signalov v elektromagnetnem spektru, uporabo elektromagnetne energije (vključno z usmerjeno energijo), s ciljem zmanjšati in preprečiti uporabo elektromagnetnega spektra s strani sovražnika in aktivnosti, ki zagotavljajo učinkovito uporabo elektromagnetnega spektra s strani lastnih sil. [1]

Sodobno elektronsko bojevanje lahko razdelimo v tri večje podskupine; elektronske protiukrepe, elektronske zaščitne ukrepe in podporne ukrepe elektronskega bojevanja. V tem članku so predstavljene posamezne podskupine kot tudi njihove glavne značilnosti. Namen članka je predstaviti sisteme za elektronsko bojevanje in izpostaviti nekaj sistemov, ki so trenutno v uporabi.

3 Elektronska komunikacija

Ena prvih oblik uporabe elektromagnetnega spektra v vojaške namene je bila elektronska komunikacija med enotami in poveljstvom s pomočjo radijskih valov v elektromagnetnem spektru. Elektromagnetne valove, ki jih oddaja radijska postaja, je mogoče prestrezati in motiti, hkrati pa iz njih izluščiti posredovano informacijo. Poleg tega lahko določimo sovražnikove elektronske zmožnosti iz modulacije, smeri sprejema signala, njegove moči in sorodnih parametrov radijske komunikacije.

Sodobne vojaške operacije zahtevajo izmenjavo velike količine informacij med sodelujočimi enotami. Najpomembnejša tipa komunikacij za namene elektronskega bojevanja so taktične komunikacije, kjer so naprave v mreži hkrati oddajniki in sprejemniki, navadno delujejo na isti frekvenci, prenos informacij pa večinoma poteka od nadrejenega k podrejenemu in pa podatkovne zveze med brezpilotnimi letali in nadzornimi centri ali med letali, ki si delijo informacije o letu, po njih pa lahko prenašamo podatke senzorjev, statusne informacije in podatke za določanje lokacije. [2]

4 Radar

Radar (angl. radio detection and ranging) je bil razvit za sledenje zrakoplovov in plovil. Danes jih uporabljamo tudi v civilni sferi za sledenje potniškim in tovornim letalom, določanje vremenskih pogojev, izrisovanje terena, itd.

Radar sestavljata sprejemnik in oddajnik, kot prikazuje slika 1. Oddan signal se odbije od površine katerekoli ovire in se vrne v sprejemnik, kjer je ta signal analiziran. Pogosto sta oddajnik in sprejemnik na isti lokaciji, lahko pa uporabljata tudi isto anteno. Skoraj vedno se uporabi zelo usmerjene antene saj se tako lahko azimut ovire, od katere se je signal odbil, natančneje določi. Razdalja do tarče (d) se izračuna iz časa med oddajo in sprejemom odboja (Δt). Enačba je:

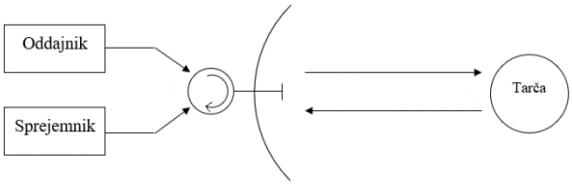
$$d = \frac{c\Delta t}{2} \tag{1}$$

pri kateri je

d = razdalja do ovire (m);

c = hitrost svetlobe (3×10^8 m/s);
Δt = čas med oddajo in sprejemom signala (s).

Če lokacija tarče ni znana, se radarski snop vrti v nekem vzorcu, da pokrije prostor v katerem bi se tarča lahko nahajala. V primeru da je tarča odkrita, ji lahko radar sledi ali pa nadaljuje s skeniranjem prostora.



Slika 1: Poenostavljen koncept delovanja radarja.

Največji sodoben radarski sistem, ki ga uporablja Slovenska vojska je Ground Master GM-403 proizvajalca Tales Raytheon in je sposoben detekcije letal, helikopterjev, manevrirnih raket, brezpilotnih letal in taktičnih balističnih raket na razdaljah do 470 km. Eden izmed takih sistemov se nahaja na Ljubljanskem vrhu. [3]

5 Elektronski podporni ukrepi

Sistemi elektronskih podpornih ukrepov (EPU) zaznavajo signale iz EMS ter določajo njihove lastnosti in možne lokacije oddajnikov. Sistemi EPU so torej tisti, ki 'poslušajo' kaj se na spektru dogaja, da se lahko ostali sistemi elektronskega bojevanja na to pravilno odzovejo. Sisteme EPU delimo na več tipov, med katerimi so najpomembnejši radarski detektorji, sistemi za določanje ciljev (angl. threat targeting sistem) in sistemi za nadzor bojišča (angl. battlefield surveillance systems).

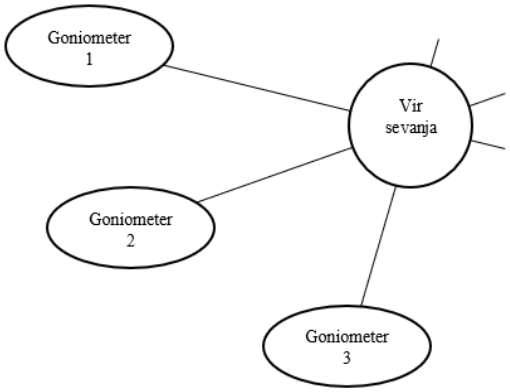
5.1 Radarski detektorji

Kot že samo ime pove, radarski detektorji opozorijo, da je vozilo (najpogosteje letalo) v območju sevanja radarja, kar omogoči pravočasno uporabo protiukrepov ali manevrov za izogib nevarnosti. Z analizo zajetih signalov lahko javljalnik določi tudi tip sovražnega radarja, ali je vir radarskega signala na primer drugo letalo ali pa protiletalski raketni sistem.

5.2 Sistemi za nadzor bojišča

Sistemi za nadzor bojišča sprejemajo in analizirajo vse vrste signalov na bojišču in tako določijo sovražnikove zmožnosti elektronskega bojevanja ter pomagajo pri določanju tarč. Takšni sistemi se lahko namestijo na vse vrste vozil, zrakoplovov in vodnih plovil. V uporabi so tudi prenosni sistemi, katere se lahko prenaša v nahrbtniku, kot je na primer Wolfhound. [4]

Goniometriranje je določanje lokacije vira elektromagnetnega sevanja, med tem ko le-ta oddaja signal. Najosnovnejši način goniometriranja je triangulacija, kjer s pomočjo najmanj dveh goniometrov, ki poznata svojo lokacijo in smer iz katere signal prihaja, lahko določimo območje, kjer se snop križa. S tremi ali več goniometri, kot prikazuje slika 2, je mogoče dobiti natančnejše meritve. Na natančnost določanja lokacije vpliva veliko dejavnikov, med drugim tudi frekvenca na kateri vir oddaja. Pri nižjih frekvencah je točno lokacijo težje določiti, ker se taki elektromagnetni valovi poslužujejo različnih vrst razširjanj, medtem ko je večina visokofrekvenčnih zvez omejena na zvezo z neposredno vidljivostjo (angl. line of sight communications).



Slika 2: Goniometriranje vira sevanja s triangulacijo.

Za določanje lokacije virov, ki oddajajo v kratkovalovnem (HF) območju, je bila razvita metoda SSL (angl. single site location), ki s pomočjo merjenja kota, pod katerim se elektromagnetni valovi odbijejo od ionosfere in znane višine plasti ionosfere lahko poleg smeri določi razdaljo do izvora signala. [5] Za goniometriranje s to metodo zadostuje le en goniometer, ki je po navadi stacionaren. Eden izmed predstavnikov premičnih goniometrov v nemški vojski je KWS RMB (nem. Kampfwertsteigerung Radio Multiband), osnovan na šest kolesnem transportnem vozilu Fuchs. Uporablja se za goniometriranje lokacije radarskih sistemov. [6]

6 Elektronski protiukrepi

Glavni namen elektronskih protiukrepov je aktivno motenje, zavajanje ali uničenje sovražnikovih elektronskih naprav ali sistemov. Nadpomenka elektronskih protiukrepov je elektronski napad, pod katerega štejemo tudi orožja, ki uporabljajo usmerjeno energijo in protiradiacijske izstrelke. Med orožji, ki uporabljajo usmerjeno energijo je še posebej kontroverzen tako imenovani Active denial system, ki ga je razvila vojska Združenih držav in v usmerjenem snopu oddaja signal s frekvenco 95 GHz z močjo 100 kW, ki v koži povzroči podoben učinek kot mikrovalovna pečica – vzburi molekule vode in maščobe, ter jo tako segreje, kar je pri testirancih povzročilo zelo neprijeten občutek. Nihče od tistih, na katerih so preizkusili ta sistem, ni zdržal dlje kot 5 sekund. Pri testirancih pa sistem ni imel

negativnih posledic na njihove oči, splošno zdravje kože ali pa na moške reprodukcijske organe. Prav tako izpostavljenost žarkom na tej frekvenci ni povečala možnosti za raka. [7]

Elektronsko motenje je načrtno sevanje elektromagnetne energije v sovražnikove sprejemnike, z namenom da bi mu otežili ali celo popolnoma onemogočili sprejemanje tistih signalov, ki dejansko nosijo informacijo. Motnje so lahko elektromagnetni šum, neželeni signal ali pa sprememba lastnosti prenosnega medija.

Učinkovitost motilcev komunikacij se meri z razmerjem med močjo motilnega in močjo motenega signala v sprejemniku (angl. jamming-to-signal ratio – J/S). Večje kot je to razmerje, bolj učinkovito je motenje. Moč signala in motenj na sprejemniku je odvisna od izhodne moči oddajnika in motilca, njune oddaljenosti od sprejemnika in terena, ki ga morajo signali premostiti. V praksi je faktor motenja okrog 10 zadosten za učinkovito motenje komunikacij.

S prekrivnim motenjem se znižuje kvaliteta izhodnega signala, tako da se v prostor oddaja moduliran šum, kar znižuje razmerje signal-šum ter ustvarja območje, kjer je sprejem signala zelo otežen. Ta način motenja se skoraj vedno uporablja za motenje komunikacij. Najpreprostejše za izvedbo je širokopasovno motenje, s katerim se prekrije vse možne frekvence, ki bi jih komunikacijski ali radarski sistemi lahko uporabljali. Širokopasovno motenje ima slab izkoristek, saj se večina moči potroši na frekvencah, ki v tistem trenutku sploh niso v uporabi, vendar za učinkovito motenje pri tem ni potrebno vedeti katere frekvence sovražnik uporablja. Boljši izkoristek ima ozkopasovno motenje, pri kateremu se najprej analizira elektromagnetni spekter. S podrobno analizo spektra se pridobi frekvence, ki jih nasprotnik uporablja, nato pa se na teh frekvencah oddaja motilne signale. Seveda je nujno, da vsake toliko časa motenje prekine, ponovno analizira spekter in po potrebi spremeni frekvence na katerih se oddajajo motilni signali. Tretja možnost je tako imenovano brišoče motenje, pri katerem se ozek pas, na katerem se moti, pomika po širšem delu spektra. S tem se nasprotniku povsem ne onemogoči delovanje, lahko pa se ga moti dovolj, da npr. odpovedo avtomatizirani procesi sledenja tarč, tako da mora sledenje ročno izvajati izkušen operater. [8]

Elektronsko zavajanje radijskih komunikacij (angl. spoofing) je oddajanje lažnih informacij sovražniku na isti način kot da bi prihajalo iz njegovega lastnega omrežja. Potrebno je kopirati postopke avtentikacije, ki jih sovražnik uporablja, zavajanje pa postane še posebej zahtevno, če je v zvezah uporabljeno šifriranje. Ena izmed tehnik zavajanja temelji na snemanju govorne zveze določenega poveljnika, ki se mu s posebno programsko opremo kasneje premeče besede, spremeni njihovo intonacijo, doda motnje ipd. nato pa v kritični situaciji, ko operater pozabi na protokole avtentikacije, odda posnetek njegovega poveljnika v kričečem tonu, ki ukazuje določen manever ali ukrep.

Pri motenju radarjev je pomembno razmerje med močjo motenja in močjo odbitega radarskega signala od tarče (na primer od letala), ki doseže radarski sprejemnik. Pomembno je tudi, da upoštevamo nekatere posebnosti radarjev, na primer smernost radarske antene, ki radarju omogoči da bolje sprejema signale iz določene smeri, medtem ko jih sprejema slabše s tiste smeri, kamor antena ni usmerjena. V primeru da motilec ne seva v glavni snop radarske antene, je potrebno motiti z veliko večjo močjo, kot jo ima sprejeti signal. Druga posebnost pa je posledica razširjanja elektromagnetnih valov; moč odboja signala od tarče se namreč do radarskega sprejemnika zmanjšuje s četrto potenco razdalje do tarče, medtem, ko se moč motnje zmanjšuje le z drugo potenco razdalje motilca od radarja. [9]

Najpreprostejša oblika radarskega motenja je samozaščitno motenje, pri katerem se motilec nahaja na tarči sami. Prednost te oblike je, da motilec seva naravnost v glavni snop radarskega sprejemnika, kar pomeni da je motenje učinkovitejše. Na sodobnih nevidnih letalih se taka postavitev ne uporablja, saj se s tem izniči prednosti majhnega radarskega preseka. Alternativa temu je motenje z daljave (angl. standoff jaming), pri katerem se motilec nahaja na drugem letalu, ki ni v dosegu protiletalskih sistemov nasprotnika. Letala s takim tipom motilca so velika in oddajajo z visoko močjo, kar je nujno, če sevamo v enega od stranskih snopov radarske antene. [10]

Da lahko protiletalski sistemi streljajo na tarčo, jim je potrebno najprej slediti preko radarja. Sled sestavljajo vrednosti kotov in razdalj iz katerih se lahko predvidi položaj tarče v prihodnosti. Elektronsko zavajanje radarskih sistemov ustvari lažno sled tarče, s čimer postanejo protiletalski sistemi neuporabni. Za zavajanje radarjev je potrebna natančnost v časovni resoluciji na nivoju manj kot mikrosekunda, zato so praktično uporabni le samozaščitni motilci, montirani na tarči sami. Eden izmed tipov elektronskega zavajanja je tako imenovani range gate pull-off (RGPO), pri katerem tarča sprejme radarski signal in ga odda s kratko zakasnitvijo, s čimer simulira odboj valov. V praksi namreč radarski signal z večjo zakasnitvijo pomeni bolj oddaljeno tarčo, ker pa dobiva več odbojev (prvi je pravi, ostali pa so emisije motilca) se lahko tarča nahaja kjerkoli v tem oknu razdalj. Radar tako ne more določiti točne lokacije tarče in izgubi njeno sled. [11]

Toplotno vodeni protiletalski izstrelki so eno najnevarnejših sodobnih orožij. Uporabljeni so lahko kot izstrelki zrak-zrak ali pa zemlja-zrak, pri slednjih so še posebej pomembni tisti, ki so izstreljeni iz relativno poceni lanserjev, za katere je velikokrat potreben le en operater. Taka orožja so pasivno vodena – senzor v glavi izstrelka sledi vročim delom motorja ali izpuha. Eden najpreprostejših protiukrepov za obrambo pred toplotno vodenim izstrelkom so toplotne vabe, ki jih letalo ali helikopter izstrelita in ki zaradi svojega vročega gorenja vodijo izstrelek stran od tarče. Poznani so tudi IR motilci, pri katerih laser v IR spektru meri v senzorsko glavo izstrelka in z utripajočimi pulzi spremeni preračunano pot izstrelka. Slovenska vojska za zaščito svojih helikopterjev Cougar uporablja sistem ISSYS (angl. Integrated Self-Protection System – ISSYS) Sestavljajo ga trije različni senzorji za zaznavanje valovanja, ki ga oddaja raketni motor, radar ali laserski namerilnik. Senzorji so vgrajeni v strukturo helikopterja tako, da

pokrivajo celotno območje okrog helikopterja. Del ISSYS je tudi sistem za izmet protiukrepov oziroma vab, tako toplotnih kot radarskih trakov. Sistem lahko ob zaznani grožnji avtomatsko spusti vabe, lahko pa le opozori posadko. [12]

Ena prvih oblik elektronskih protiukrepov so bili trakovi za motenje radarjev (angl. chaff), narejeni iz kovinskih lističev, žice ali steklene volne, prevlečene s kovino, ki odbijajo elektromagnetno energijo in tako ustvarjajo lažno radarsko sliko, običajno pa se odvržejo iz zračnega plovila ali pa se izstrelijo iz granat ali raket. Dolžina trakov je odvisna od frekvence, ki jo želijo optimalno odbijati. Če radarski sistem uporablja več frekvenc, se odvrže trakove različnih dolžin.

V široki uporabi so tudi imitacije ali vabe, ki imajo za namene elektronskega bojevanja podobno radarsko odbojno površino in vrnejo enak odbojni impulz, kot objekt ki ga poskušajo imitirati. To se lahko doseže pasivno z uporabo kotnih sejalcev, ki so neke vrste kresnice za radarske valove – signale iz katerekoli smeri, odbijejo nazaj proti viru, zaradi česar imajo tudi do sto krat večjo radarsko odbojno površino, kot bi jo imelo telo z isto fizično površino. Druga možnost so aktivne elektronske imitacije, ki ustvarjajo večjo navidezno radarsko površino s sprejemom in oddajo ojačenih radarskih signalov. Nasprotnik je lahko s tem zaveden, da na relativno poceni imitacijah iztroši svoje drage izstrelke, ob aktivaciji sovražnikovih obrambnih sistemov pa lahko s sistemi za nadzor bojišča določimo elektronske kapacitete in lokacije sovražnih položajev za elektronsko bojevanje.

7 Elektronski zaščitni ukrepi

S pomočjo elektronskih zaščitnih ukrepov se poskuša nevtralizirati sovražnikove elektronske protiukrepe, oziroma elektronski napad, ter s tem zaščititi elektromagnetni spekter za lastno uporabo.

Radarji na primer lahko sledijo viru motenja ter tako sledijo tarči, ki poskuša oddati svoj lažni položaj. Pri radijskih komunikacijah se lahko uporabi katera izmed tehnik oddajanja z nizko verjetnostjo prestrezanja (angl. low probability of intercept), kot je na primer frekvenčno skakanje, ki signal razširi na širok spekter in je tako manj občutljiv na ozkopasovno motenje.

8 Zaključek

Področje elektronskega bojevanja je zelo široka tema, ki zahteva številne strokovnjake tako v civilnem kot vojaškem delu obrambne strukture. V tem članku so predstavljene posamezne podskupine kot tudi njihove glavne značilnosti elektronskega bojevanja. Namen članka je predstaviti sisteme za elektronsko bojevanje in izpostaviti nekaj takih sistemov v trenutni uporabi. Dimenziji elektronskega vojskovanja je potrebno posvečati veliko pozornost, saj se tehnologije in načini dela nenehno spreminjajo, zastarelost sistemov pa lahko pomeni veliko šibkost v konfliktu z nasprotnikom.

Literatura

[1] AAP-06(2019), "NATO glossary of terms", North Atlantic Treaty Organization NATO Standardization Office (NSO) 2019, dostopano 20. 7. 2020

[2] Golob D., Možina U., Frangež Z, Skripta elektronsko bojevanje, Poveljstvo za doktrino, razvoj, izobraževanje in usposabljanje, 2006, str. 19-24, ISBN 10-961-660-04-4

[3] Radarski sistem Ground master GM 403, dostopano dne 31. 8. 2020, dostopno na: http://www.slovenskavojska.si/oborozitev-in-oprema/zracna-obramba/tales-raytheon-ground-master-gm-403/

[4] Wolfhound Handheld Threat Warning System, dostopano 20. 7. 2020, dostopno na: https://www.definitive-design.com/project/wolfhound-handheld-threat-warning-system/

[5] Coetzee, P. J., and du Plessis, W. P. (2016), Definition of a quality factor for single site location estimates, Radio Sci., 51,str. 555– 562, doi:10.1002/2015RS005799.

[6] Transportpanzer Fuchs KWS-RMB, dostopano dne 31. 8. 2020, dostopno na: https://www.bundeswehr.de/de/ausruestung-technik-bundeswehr/landsysteme-bundeswehr/transportpanzer-fuchs-kws-rmb

[7] Patrick A. Mason, Thomas J. Walters, John DiGiovanni, Charles W. Beason, James R. Jauchem, Edward J. Dick, Jr, Kavita Mahajan, Steven J. Dusch, Beth A. Shields, James H. Merritt, Michael R. Murphy, Kathy L. Ryan, Lack of effect of 94 GHz radio frequency radiation exposure in an animal model of skin carcinogenesis, Carcinogenesis, Volume 22, Issue 10, oktober 2001, str. 1701–1708, doi:10.1093/22.10.1701

[8] Adamy D. L., Tactical Battlefield Communications Electronic Warfare, Artech House Inc., 2009, str. 251-256, ISBN-13: 978-1-59693-387-3

[9] D. Adamy, "Introduction to Electronic warfare Modeling and Simulation", Artech House Inc., 2003,str. 42-53, ISBN 1-58053-495-3

[10] S. Vardhan and A. Garg, "Information jamming in Electronic warfare: Operational requirements and techniques," 2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE), Hosur, 2014, str. 49-54, doi: 10.1109/ICECCE.2014.7086634.

[11] Neri F., Introduction to Electronic Defense Systems 2nd edition, Artech House Inc., 2001, str. 373- 435, ISBN 1-58053-179-2

[12] Podgoršek B., 2019, Sistem samozaščite helikopterja z metalci toplotnih in radarskih vab, Revija SV, številka 10, str. 16-17

Multimedija

Multimedia

Vrednotenje uporabniške izkušnje inteligentnih tipk

Gašper Kastelic

Fakulteta za elektrotehniko, Univerza v Ljubljani
gasper.kastelic97@gmail.com

Evaluating the user experience of intelligent switches

Abstract. *This paper addresses a potential way of evaluating and measuring the user experience of intelligent switches used to control smart-home systems. New smart-home functionalities require a redesigned controlling interface which has to be easy to use and effective. Therefore, a user study was conducted in which four different controlling interfaces were tested. The article includes a summary of the methodology that was used in this specific study. In the last part results of the user experience testing are provided. It has been discovered that the user experience of intelligent switches is comparable to user experience of traditional wall switches.*

1 Uvod

Napredek informacijsko-komunikacijskih tehnologij je omogočil življenje v pametnih domovih. Koncept pametnega doma združuje dom z električnimi sistemi, kot so razsvetljava, ogrevanje in druge električne naprave, ki jih lahko povežemo v internet oz. jih lahko nadziramo na daljavo z uporabo pametnih telefonov. Z višanjem ozaveščenosti ljudi o pametnih domovih in širjenjem pripadajočega trga se povečuje tudi težnja po merjenju in konstantnemu izboljševanju uporabniške izkušnje pri uporabnikih pametnih domov. Pomemben del uporabniške izkušnje predstavlja interakcija uporabnika s pametnimi sistemi in njihovo upravljanje [1]. Osnoven pristop upravljanja tradicionalnih in tudi pametnih domov je uporaba klasičnih stikal. Ta so v splošnem vsem zelo dobro znana, saj jih lahko najdemo praktično v vsaki zgradbi oz. domu. Dejstvo je, da imajo klasična stikala v splošnem omejeno funkcionalnost, vendar so za uporabo v navadnih domovih povsem zadovoljiva. Uporaba klasičnih stikal v pametnih domovih bi lahko že v bližnji prihodnosti postala neprimerna, predvsem zaradi omejenih funkcionalnosti, obenem pa klasična stikala ne sledijo napredku pametnih domov.

Zato obstaja težnja po uvajanju in uporabi tako imenovanih inteligentnih stikal oz. tipk. Glavni prednosti inteligentnih tipk sta programljivost oz. personalizacija nastavitev in večji nabor funkcionalnosti. Inteligentne tipke različnih proizvajalcev imajo lahko integrirane različne senzorje npr. senzor za merjenje temperature zraka, kakovost zraka, senzor bližine ... Tipke imajo v nekaterih primerkih vgrajen "touch-screen" zaslon, kar pomeni, da ni več potreben fizični pritisk tipke, temveč samo blagi dotik. Fizični »feedback« ob pritisku oz. dotiku tipk je lahko zamenjan s haptično povratno informacijo (vibracijo) ali svetlobnim učinkom. Za upravljanje pametnega doma se uporablja tudi glasovne sisteme, kot sta Amazon Alexa in Google Voice Assistant. Glasovni sistemi v večini primerov ne predstavljajo samostojnega upravljanja, temveč služijo le kot dopolnilo ostalim načinom upravljanja. Morda najbolj znan način upravljanja pametnih domov je upravljanje z uporabo pametnih telefonov oz. namenskih mobilnih aplikacij. Ti omogočajo popoln nadzor nad pametnim domom tudi na daljavo in za uporabnike predstavljajo dodatno fleksibilnost. Različni načini upravljanja pametnega doma torej od uporabnikov zahtevajo različne načine interakcije. Posamezen pristop ima svoje prednosti in tudi svoje slabosti. Te se kažejo predvsem v različni uporabniški izkušnji uporabnikov pametnih domov.

Zato v nadaljevanju sledi predstavitev pregleda področja, načrtovanja, izvedbe in vrednotenja uporabniške izkušnje pri upravljanju pametnega doma z različnimi uporabniškimi vmesniki. Poudarek je na upravljanju z inteligentnimi tipkami. V zadnjem delu so povzeti rezultati testiranja in glavne ugotovitve uporabniškega testiranja inteligentnih tipk, namenjenih za upravljanje pametnega doma.

2 Pregled področja

Osnovne funkcije upravljanja pametnega doma so: nadzor razsvetljave, nadzor ogrevanja in hlajenja prostorov, nadzor okenskih žaluzij in senčil. Rezultati ankete, opravljene med potencialnimi uporabniki pametnih domov, kažejo, da so poleg osnovnih funkcionalnosti zaželene tudi funkcije zmanjševanja porabe električne energije gospodinjstev in dodatni varnostni sistemi [2]. Koncept pametnega doma omogoča lažji nadzor nad električnimi napravami, kar posledično pomeni lažje sledenje in analizo porabljene električne energije. Drugi vidik zmanjševanja porabe energije je nadzor na daljavo. Ta omogoča boljšo in natančnejšo optimizacijo porabe, npr. prižig ogrevanja prostorov preko mobilne aplikacije, preden uporabnik prispe domov. K zmanjševanju porabe električne energije pripomore tudi uporaba scen oz. urnikov prižiganja in ugašanja različnih električnih naprav in sistemov. Popularna zahteva uporabnikov pametnih domov je tudi zagotavljanje popolnega zasebnega nadzora in kontrole nad celotnim sistemom. Tu se med uporabniki pojavlja problem zasebnosti, nezaupanja

novim tehnologijam in problem internetne varnosti. Povprečni uporabnik si namreč želi dom, kjer bi se počutil varnega in užival maksimalno udobje doma [3].

Pametni domovi bi lahko odigrali pomembno vlogo tudi pri zagotavljanju zdravja ljudi. Osnoven pristop zagotavljanja zdravja uporabnikov bi lahko bil konstanten nadzor in zbiranje podatkov o zdravstvenem stanju posameznika. Z uporabo senzorjev in opazovanjem posameznikovih vedenjskih vzorcev bi lahko pametni dom preventivno opozarjal na morebitna zdravstvena tveganja in služil kot alarmni sistem v primeru nenadnih zdravstvenih dogodkov. Pametni domovi bi lahko še posebej olajšali življenje invalidom in starejšim, ki bi z manj truda in na enostavnejši način lahko upravljali z bivalnimi prostori. Z vsemi zdravstvenimi in drugimi funkcionalnostmi pametnega doma pa je potrebno zagotoviti tudi enostavno in primerno interakcijo s pametnimi sistemi. Raziskave namreč kažejo, da predstavljajo zapleteni upravljalski vmesniki pri starejših uporabnikih velik problem [4]. Zato je ključno, da so uporabniški vmesniki pravilno načrtovani s ciljem, da uporabniku ponudijo jasno, enostavno in prijetno uporabniško izkušnjo.

Če je bil koncept pametnega doma pred leti predstavljen samo s povezanostjo doma v internet, je danes temu nekoliko drugače. Trendi namreč kažejo, da v ospredje prihaja težnja po avtomatizaciji hišnih električnih sistemov. Danes imajo nekatere rešitve že integrirano umetno inteligenco, ki spremlja uporabnikove bivalne navade in na podlagi teh spreminja razmere v domu, bodisi razsvetljavo ali pa npr. ogrevalne sisteme. Opazen je tudi trend povečevanja števila senzorjev, ki bodo poleg umetne inteligence v prihodnosti odigrali pomembno vlogo pri avtomatizaciji domov. Zanimiv je tudi koncept pametnih energetskih omrežij, ki bi lahko z usmerjanjem in deljenim upravljanjem električne energije pametnih domov pripomogel k optimizaciji porabe električne energije gospodinjstev, na drugi strani pa pomagal elektrodistributerjem zagotoviti večjo zanesljivost in prilagodljivost energetskih sistemov [4]. Avtomatizacija domov bo vplivala tudi na razvoj upravljalskih uporabniških vmesnikov. Ti bodo morali slediti potrebam in željam uporabnikov.

3 Opis projekta

Razvoj novih funkcionalnosti pametnega doma je prinesel tudi uporabo nove vrste upravljavskih vmesnikov - inteligentnih tipk. Te se po dizajnu skušajo približati klasičnim stikalom. Prednosti teh so dodatne funkcionalnosti, kot je npr. integriran termostat, senzor prisotnosti in prilagajanje osvetlitve. Inteligentne tipke so trenutno sorazmerno nov element upravljanja pametnega doma, ki šele dobro prodira na trg pametnih domov. Zato je težko določiti, kako bodo uporabniki sprejeli koncept inteligentnih tipk. Za lažje in natančnejše vrednotenje smiselnosti uporabe inteligentnih tipk smo zato izvedli uporabniško testiranje inteligentnih tipk podjetja Entia, ki se ukvarja z razvojem in implementacijo pametnih domov. Osnovni cilj projekta je bilo ugotoviti, kakšna je prejeta kakovost storitve, ki jo ponujajo inteligentne tipke. Fokus je bil predvsem na odkrivanju in razumevanju vpliva novega dizajna in dodatnih funkcionalnosti na mnenje uporabnikov. Pomemben del projekta pa je bila tudi primerjava uporabniške izkušnje med upravljanjem z inteligentnimi tipkami in klasičnimi stikali.



Slika 1: Inteligentne tipke podjetja Entia

Raziskovalna vprašanja so bila razdeljena na dva dela in sicer glede na dizajn in funkcionalnost inteligentnih tipk. S testiranjem uporabniške izkušnje smo želeli ugotoviti ali so inteligentne tipke za nadziranje razsvetljave, okenskih senčil in termostata dovolj intuitivne. Zanimalo nas je, kakšen je vpliv povratne informacije tipk na uporabnika oz. ali haptični (vibracijski) odziv zadovolji pomanjkanje fizičnega premika tipke. Glede funkcionalnosti nas je zanimalo ali bi uporabniki v prosti izbiri med različnimi upravljalskimi vmesniki najprej posegli po uporabi inteligentnih tipk. Raziskovali smo ali je upravljanje z inteligentnimi tipkami časovno primerljivo z uporabo klasičnih stikal. Prav tako smo preverjali smiselnost implementacije uporabniških scen, ki združujejo več prednastavljenih ukazov v en sam ukaz oz. pritisk na tipko.

Največji delež testiranja je bil torej izveden na inteligentnih tipkah podjetja Entia. Uporabljeni sta bili dve verziji inteligentnih stikal z osmimi posameznimi tipkami. Prva verzija tipk, je bila črne barve in je vsebovala haptično povratno informacijo ob posameznem pritisku. Druga verzija tipk je bila bele barve in haptične povratne informacije ni vsebovala. Vsaka posamezna tipka je bila osvetljena, osvetljen pa je bil tudi zadnji del celotnega stikala, ki služi za lažjo lociranje stikala ob npr. nočnih urah, ko ni svetlobe. Tipke so vsebovale tudi termostat in temperaturno skalo za nastavljanje in določanje željene temperature. Obe verziji tipk sta bili na začetku ustrezno sprogramirani glede na specifične potrebe testiranja. Omogočale so uporabo dveh vrst pritiskov in sicer kratkega pritiska in dolgega pritiska.

4 Metodologija in potek vrednotenja

Za vrednotenje uporabniške izkušnje je bila uporabljena »within-subjects« metoda. Vsi testni uporabniki so

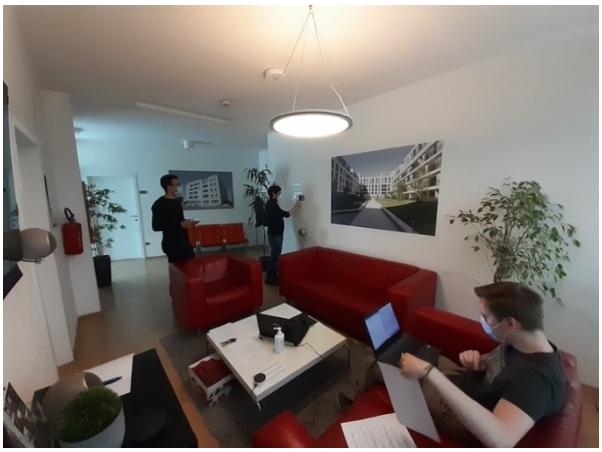
testirali vse štiri različice upravljavskih vmesnikov in sicer: inteligentne tipke brez haptike, inteligentne tipke s haptiko, klasično stikalo in mešano upravljanje z mobilno aplikacijo oz. glasovno pomočnico Amazon Alexo. Testiranje je bilo izvedeno v testni sobi podjetja Entia. Glede na omejitve prostora je bilo moč nadzorovati: spreminjanje temperature prostora, prižiganje, ugašanje in prilagajanje razsvetljave, nastavljanje lege okenskega senčila in upravljanje prednastavljenih scen. Pri testiranju je sodelovalo 33 testnih uporabnikov. Nagovarjanje ljudi k sodelovanju je potekalo ustno oz. po elektronski pošti. Povprečna starost testnih uporabnikov je bila 30 let.

Vsak posameznik je bil ob začetku testiranja seznanjen s potekom testiranja in naprošen k podpisu soglasja o sodelovanju in uporabi podatkov eksperimenta v raziskovalno znanstvene namene. Sledilo je izpolnjevanje začetnega vprašalnika, ki je spraševal po demografskih podatkih in o poznavanju tehnologij in koncepta pametnega doma. Nato je sledilo izpolnjevanje vprašalnika PANAS, ki meri trenutno počutje posameznika. Vprašalnik PANAS [5] je pomemben, ker meri vpliv izvedbe eksperimenta na počutje testnih uporabnikov ter posredno tudi na rezultate testiranja. Po izpolnitvi vprašalnika PANAS je vodja testiranja testnemu uporabniku razložil delovanje upravljavskega vmesnika in odgovoril na možna zastavljena vprašanja s strani uporabnika. Potem je vodja testiranja pričel z branjem scenarijev. Scenariji so predstavljali logična navodila, ki so testnemu uporabniku sporočila, katere ukaze mora izvesti na upravljalskem vmesniku. Primer dela scenarija:

»Predstavljajte si, da se nahajate v vaši dnevni sobi. Odločili ste se, da si boste na televiziji ogledali vašo najljubšo serijo. Zaradi jutranjih sončnih žarkov, ki sijejo direktno na ekran televizije, je gledanje serije nemogoče. Spremenite razmere v sobi.

a. Izklopite obe glavni luči.
b. Izklopite okroglo luč, ki se nahaja nad mizo.
c. Spustite okensko senčilo do približno polovice višine okna.
d. Vklopite ogrevanje.
e. Na glas povejte, na kakšno temperaturo je trenutno nastavljen termostat.
f. Temperaturo termostata povišajte na 22°C.«

Pomočnik vodje testiranja si je ob testiranju beležil čas, ki ga je uporabnik porabil za izvedbo posameznih ukazov oz. nalog ter si zapisoval napačne pritiske uporabnika na upravljavskem vmesniku. Po končanem scenariju je sledilo izpolnjevanje vprašalnika o mnenju uporabnika o upravljalskem vmesniku, ki ga je uporabnik ravnokar testiral. Uporabljeni so bili standardni vprašalniki za merjenje uporabniške izkušnje (UEQ in SUS). Uporabnik je nato po istem principu nadaljeval s testiranjem ostalih treh upravljavskih vmesnikov. Po vsakem testiranju posameznega vmesnika je torej izpolnjeval ločen vprašalnik. Za testiranje inteligentnih tipk s haptiko, inteligentnih tipk brez haptike in klasičnega stikala se je uporabljal enak scenarij. Za testiranje mešanega upravljanja z mobilno aplikacijo in glasovno pomočnico Alexo pa se je uporabljal nekoliko skrajšan in prilagojen scenarij, saj je bil glavni fokus na testiranju inteligentnih tipk v primerjavi s klasičnimi stikali. Z metodo latinskega kvadrata smo vsakemu posameznemu testnemu uporabniku določili vrstni red testiranja posameznih vmesnikov, ter se na ta način izognili vplivu naučenosti uporabnikov, zaradi ponavljanja enakih scenarijev oz. nalog. Po testiranju uporabnika vseh štirih vmesnikov je sledilo izpolnjevanje zaključnega vprašalnika. Ta je vključeval vprašanja glede celotnega poskusa kot celote ter vprašanja o počutju PANAS. Na samem koncu je sledila še izvedba pol-strukturiranega intervjuja, ki jo je vodil tretji član raziskovalne ekipe. Namen intervjuja je bil identificirati glavne prednosti in slabosti posameznih upravljavskih vmesnikov ter identificirati ideje testnih uporabnikov, ki jih je bilo nemogoče pridobiti pri strukturiranih vprašalnikih. S tem je bilo testiranje za uporabnika zaključeno.



Slika 2: Testiranje uporabniške izkušnje v testni sobi

Potrebno je omeniti, da so testiranja potekala v času izrednih razmer zaradi virusa Covid-19. Izvedba testiranja je bila zato nekoliko otežena. Izvajalci testiranja so pozorno skrbeli za doseganje vseh varnostnih standardov. Ob prihodu vsakega novega testnega uporabnika je sledilo razkuževanje površin vseh upravljavskih vmesnikov. Veljala pa je tudi obvezna uporaba zaščitnih mask tako za tesne uporabnike kot tudi za vse izvajalce testiranja.

5 Diskusija in ugotovitve

Pred začetkom glavnega testiranja sta bili izvedeni dve pilotski testiranji. Ti sta pokazali, da je besedilo scenarijev in število zahtevanih nalog preobširno. To se je pokazalo tudi pri testnih uporabnikih pilotskega testiranja, saj je bilo zaradi prekomernega ponavljanja navodil in ukazov pri uporabnikih moč zaznati upad zanimanja in koncentracije. Do glavnega testiranja je bila zato izvedena optimizacija scenarijev in celotnega poteka testiranja. Zaradi izrednih razmer je bilo iskanje primernih testnih uporabnikov zahtevna naloga. Posledica je manj optimalna in neenakomerna porazdelitev starosti testnih uporabnikov, spola uporabnikov in njihove izobrazbe. Zato je nadaljnja analiza podatkov glede na starost, spol oz. izobrazbo nemogoča in nesmiselna, saj bi prišlo do naključnih

ekstremnih vrednosti rezultatov. V idealnih razmerah bi bila dobrodošla tudi višja povprečna starost testnih uporabnikov, saj bi na ta način zajeli rezultate ciljne skupine potencialnih uporabnikov pametnih domov.

Glede izrednih razmer in obveznega nošenja mask med testiranjem se ocenjuje, da ta vidik bistveno ni vplival na počutje testnih uporabnikov ter posledično na rezultate testiranj. Najbolj izpostavljena naloga je bila glasovno upravljanje ob uporabi zaščitne maske, vendar glasovno upravljanje ni bil najpomembnejši del tega testiranja, zato to na rezultate bistveno ne vpliva. Zaradi določenih karakteristik elektroinštalacije testnega prostora inteligentnih tipk in klasičnega stikala ni bilo možno pritrditi na enako višino. To bi potencialno lahko vplivalo na rezultate testiranja, vendar je vpliv tega vidika težko ugotoviti.

V nadaljevanju so predstavljene glavne ugotovitve testiranja uporabniške izkušnje inteligentnih tipk. Podatki uporabniškega testiranja so bili v fazi analize obdelani s pomočjo programa Tableau, glavna analiza podatkov pa je bila izvedena v programskem okolju SPSS. Rezultati kažejo, da je bilo uporabnikom najljubše upravljanje z mobilno aplikacijo, kar pa ni presenečenje, glede na to, da je vsakodnevna uporaba pametnih telefonov pri testnih uporabnikih prej pravilo kot pa izjema. Upravljanje pametnega doma z inteligentnimi tipkami je bilo približno enako zadovoljivo kot s klasičnimi stikali. Rezultati kažejo tudi, da so se statistično pomembne razlike med inteligentnimi tipkami in klasičnimi stikali pokazale le pri nalogi, ko je bilo potrebno upravljanje upravljalskega vmesnika brez pogleda. Tu se je klasično stikalo pokazalo kot izrazito boljše. Ta rezultat je morda ob vsakodnevni rabi stikal oz. tipk v realnih primerih lahko nepomemben.

V splošnem rezultati indicirajo, da so si inteligentne tipke in klasična stikala tako po intuitivnosti kot težavnosti upravljanja enakovredna. Haptični odziv ob pritisku na inteligentne tipke je, kot kaže, dovolj dobra zamenjava za fizični premik stikala. Dejstvo je, da je bil haptični odziv med testnimi uporabniki pozitivno sprejet in prepoznan kot potreben, saj je kot pričajo rezultati upravljanje inteligentnih tip brez haptične povratne informacije nekoliko nerodno in oteženo. Moč oz. stopnja haptične vibracije pa ostaja prosta izbira vsakega posameznika, saj je nemogoče določiti stopnjo, ki bi ustrezala vsem uporabnikom. Med nekaterimi uporabniki se je pojavila tudi želja po spremenjeni teksturi samega mesta pritiska. Na ta način bi lažje locirali mesto pritiska tudi s pomočjo tipa. Prednastavljene scene so bile dobro sprejete in se uvrščajo kot zanimiva funkcionalnost nadzora pametnega doma. Kot prednost klasičnih stikal se je pokazalo dejstvo, da so uporabniki navajeni uporabe le teh. Glede na analizo intervjujev pa je moč zaznati, da je pomembna prednost inteligentnih tipk napram klasičnim stikalom modern dizajn, ki pri testnih uporabnikih vzbuja občutek prestiža.

6 Zaključek

Popularnost pametnih domov se v zadnjem času vztrajno povečuje. Z razvijanjem novih tehnologij se razvijajo tudi nove funkcionalnosti, ki uporabnikom pametnih domov omogočajo večje udobje, prijetnejše bivanje in nižje stroške porabljene energije. Integracija novih funkcionalnosti pa zahteva tudi nov koncept uporabniških vmesnikov, ki morajo kljub zapletenim procesom, ki se odvijajo v ozadju, uporabniku nuditi enostaven in učinkovit nadzor in upravljanje celotnega sistema pametnega doma. Že poznana klasična stikala počasi nadomešča nov koncept upravljalskega vmesnika, imenovanega inteligentne tipke. Te v svoji osnovi ponujajo prenovljen dizajn, integrirane senzorje in nov način upravljanja z vpeljavo kratkega in dolgega pritiska na tipko. Z novim principom upravljanja pametnega doma pa se pojavljajo nova vprašanja o uporabniški izkušnji, ki jo nudijo inteligentne tipke.

Zato je v prispevku predstavljen način vrednotenja uporabniške izkušnje inteligentnih tipk, ki je bil izveden v obliki uporabniškega testiranja s poudarkom na primerjavi uporabniške izkušnje inteligentnih tipk in tradicionalnih klasičnih stikalih. Podana je kritična ocena celotnega poteka testiranja. Na koncu so predstavljeni še rezultati izvedenega testiranja uporabniške izkušnje inteligentnih tipk s podanimi ključnimi ugotovitvami. Glede na rast trga pametnih domov se zdi, da bodo uporabniška testiranja v prihodnosti odigrala pomembno vlogo pri dizajniranju in načrtovanju upravljalskih vmesnikov pametnih domov.

Zahvala

Zahvaljujem se prof. ddr. Iztoku Humarju in doc. dr. Jožetu Guni za mentorstvo in vodenje projekta. Zahvale gredo tudi podjetju Entia, ki je omogočilo testno okolje, ter vsem ostalim študentom, ki so sodelovali pri projektu. Hvala tudi testnim uporabnikom, ki so si vzeli čas in testirali uporabniško izkušnjo inteligentnih tipk.

Literatura

[1] D. W. Seo, H. Kim, J. S. Kim, J. Y. Lee: Hybrid reality-based user experience and evaluation of a contex-aware smart home, 2015

[2] B. Schiffhauer, J. Bernotat, F. Eyssel, R. Bröhl, J. Adriaans: B. Schiffhauer, J. Bernotat, F. Eyssel, R. Bröhl, J. Adriaans: Let the user decide! User prefrences regarding functons, apps and interfaces of a smart home and service robot, 2016

[3] Y. Zhao, X. Zhang, J. Crabtree: Human-computer interaction and user experience in smart home research: A critical analysis, 2016, Volume 17, Issue 3, pp. 11-19

[4] M. R. Alam, M. A. M. Ali, M. B. I. Reaz: A rewiev of smart homes: Past, present and future, 2012, IEEE Transacions on systems man and cybernetics part C

[5] D. Watson, L. A. Clark, A. Tellegen: Development and validation of brief measures of positive and negative affect: the PANAS scales, 1988, *54*(6), 1063

Javni štipendijski, razvojni, invalidski in preživninski sklad Republike Slovenije
REPUBLIKA SLOVENIJA MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST IN ŠPORT
EVROPSKA UNIJA EVROPSKI SOCIALNI SKLAD NALOŽBA V VAŠO PRIHODNOST
Naložbo sofinancirata Republika Slovenija in Evropska unija iz Evropskega socialnega sklada.

Raznolikost voznikovih odzivov na zahtevo za prevzem vodenja pogojno avtomatiziranega vozila

Timotej Gruden, Grega Jakus

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, Ljubljana, Slovenija
E-pošta: timotej.gruden@fe.uni-lj.si

Diversity of drivers' reactions on take-over request of a conditionally automated vehicle

Abstract. *This paper presents different users' reactions on take-over request (TOR) of a conditionally automated vehicle. The data was collected with a user study in a driving simulator with 24 participants. They were exposed to six different take-over requests, two of them required to stop the vehicle and four of them required to bypass the obstacle.*

We observed the drivers' reactions before and after the take-over was performed. The attention time was normally distributed with an average of 1,49 s ± 0,53 s. Majority of drivers used the brake pedal prior to turning the steering wheel. However, some cases where drivers did not use the brake or did not react to the TOR at all were also noted. Knowing the drivers' reactions can therefore help TOR user interface designers to actively encourage safe driving.

1 Uvod

Mednarodno združenje inženirjev s področja avtomobilizma SAE (ang. Society of Automotive Engineers) je že leta 2014 različne stopnje avtomatizirane vožnje razvrstilo po šest-stopenjski lestvici [1]. Prve tri stopnje (0, 1 in 2) predstavljajo klasično, ročno vožnjo, ki je v prvi, še v večji meri pa v drugi stopnji, dopolnjena z naprednimi podpornimi oz. asistenčnimi sistemi (ang. ADAS – Advanced Driver Assistance System). Preostale tri stopnje (3, 4 in 5) predstavljajo postopen prehod do popolnoma avtonomne vožnje.

Čeprav je razvoj avtomatizacije vožnje razmeroma hiter, pa popolnoma avtonomna vožnja zaradi številnih razlogov še dolgo ne bo mogoča. Namesto nje lahko srednjeročno pričakujemo soobstoj ročne in različnih stopenj avtomatizirane vožnje. Trenutno je aktualen prehod v tretjo stopnjo, ki ji pravimo tudi pogojno avtomatizirana vožnja. Ta omogoča avtonomno vožnjo le v nekaterih voznih okoljih, kot je na primer avtocesta; voznik lahko med tem svojo pozornost preusmeri drugam, na primer uporabi zabavno-informacijskih sistemov. Ko se vozilo znajde v okoliščinah, za katere nima predvidenega odziva, mora predati vodenje vozniku.

Ključnega pomena pri načrtovanju omenjene interakcije predaje vodenja so dobri (učinkoviti, zmogljivi in zadovoljivi [2]) uporabniški vmesniki [3]. Ti morajo namreč v omejenem času na pravilen način (navadno čim hitreje) posredovati ravno pravo količino in vsebino informacije vozniku. S tem želijo sprožiti čim boljši odziv voznika in posledično preprečiti morebitne kritične situacije.

Zeeb idr. so v svojem modelu reakcijskega časa (od izražene zahteve za prevzem vodenja vozila do dejanskega prevzema vodenja) kot prvi voznikov odziv po zahtevi za prevzem vodenja predpostavili pogled na cesto, za tem (ali izjemoma sočasno) naj bi vozniki vzpostavili stik z vozilom (ang. »motor readiness«), po razmisleku (kognitivnem procesiranju) pa sledi dejanski prevzem nadzora [4]. Endsley v svoji teoriji zavedanja situacije pred samim pogledom na cesto predvideva še krajši časovni interval namenjen zamenjavi nalog [5], a ga v praksi težko zaznamo ločeno od dejanskega pogleda na zaslon. Celoten čas od podane zahteve za prevzem vodenja do dejanskega prevzema, ki vključuje tako pogled na cesto kot vzpostavitev stika z vozilom, je navadno med 2 in 3,5 sekundami [6].

Manj raziskani so odzivi uporabnikov po prevzemu nadzora. Mnoge študije navajajo, da je pri načrtovanju dobrega uporabniškega vmesnika za podajanje zahteve za prevzem vodenja potrebno poleg samega reakcijskega časa do predaje upoštevati tudi kakovost predaje [7]–[9]. Köhn idr. so oceno kakovosti (samozavesti, zaupanja) voznikom merili subjektivno z uporaba vprašalnika [7]. Zeeb idr. so kot mero kakovosti upoštevali odmik od sredine voznega pasu in prečni pospešek [8]. Gold idr. so med modeliranjem prevzema vodenja opazili statistično pomembne razlike v uspešnosti med vozniki, ki so uporabili zavoro, in vozniki, ki niso zavirali [10].

V tem prispevku želimo predvsem prikazati množičnost različnih odzivov različnih uporabnikov na podano zahtevo za prevzem vodenja. Natančneje, odgovoriti želimo na raziskovalno vprašanje: *»Kako lahko poznavanje odzivov voznikov na podano zahtevo za prevzem vodenja pogojno avtonomnega vozila pripomore pri načrtovanju uporabniških vmesnikov?«*

Naslednje poglavje predstavlja uporabljeno metodologijo – način pridobivanja podatkov ter upoštevane različice odzivov. V tretjem poglavju so grafično podani rezultati (tj. tipični odzivi različnih voznikov), zadnje poglavje pa je namenjeno komentarjem rezultatov.

2 Metodologija

2.1 Nabor podatkov

Za proučevanje različnih odzivov smo izvedli uporabniško študijo na podlagi pilotne raziskave

uporabniških vmesnikov za prevzem vodenja pogojno avtonomnega vozila [11]. Sodelovalo je 24 voznikov prostovoljcev med 19. in 50. letom starosti (povprečna starost 28,4 let), od tega 58% moškega spola.

V visoko zmogljivem simulatorju vožnje Nervtech^{TM} [12] jih je med avtonomno vožnjo po zamegljeni tripasovni avtocesti sistem za nadzor avtonomne vožnje na šestih različnih lokacijah pozval k prevzemu vodenja vozila zaradi ovire na cesti (Slika 1). V štirih primerih se je od uporabnikov pričakovalo, da oviro obvozijo, v dveh primerih pa so morali vozilo zaradi popolne zapore ustaviti. V primeru obvoza, je sistem uporabnikom podal tudi informacijo o smeri ovire (levo oz. desno). Polovici uporabnikov je bila zahteva za prevzem vodenja posredovana preko taktilno-vizualnega vmesnika, drugi polovici pa preko zvočno-vizualnega uporabniškega vmesnika. Za zaznavanje točke voznikove pozornosti smo uporabili sledilnik pogleda Tobii Pro Glasses 2 [13].



Slika 1. Voznik v simulatorju vožnje med prevzemanjem vodenja vozila, kar nakazuje rdeče obarvan LED trak pod desnim delom srednjega zaslona

Med avtonomno vožnjo so bili vozniki mentalno zaposleni z igranjem iger na mobilnem telefonu. Ob podani zahtevi za prevzem vodenja s strani sistema avtonomne vožnje je bil voznik odgovoren za uspešen prevzem in varno razrešitev situacije na cesti. Po tem je ponovno vklopil avtonomno vožnjo in nadaljeval z igranjem igre.

2.2 Odzivi pred prevzemom vodenja

Šest sekund pred potencialnim trkom [6] je sistem avtonomne vožnje sprožil zahtevo za prevzem vodenja (ang. TOR – »take-over request«). Od te točke dalje smo merili naslednje časovne intervale:

- *pozornost* – trajanje med TOR in prvim pogledom na zaslon,
- *zaviranje* – trajanje med TOR in prvim pritiskom na zavorni pedal, večjim od 2,5% hoda stopalke,
- *zavijanje* – trajanje med TOR in prvim zasukom volana, večjim od 0,1 rad.

Manjšega izmed zadnjih dveh časovnih intervalov (zaviranje ali zavijanje) smo proglasili za *reakcijski čas* (RT), tj. čas od TOR do uspešnega prevzema vodenja vozila. Iz omenjenih časov smo prepoznali tudi uporabnikovo *prvo reakcijo*, ki je lahko bila zaviranje, zavijanje ali sočasno zaviranje in zavijanje (zaviranje in zavijanje znotraj časovnega intervala 350 ms).

2.3 Odzivi po prevzemu vodenja

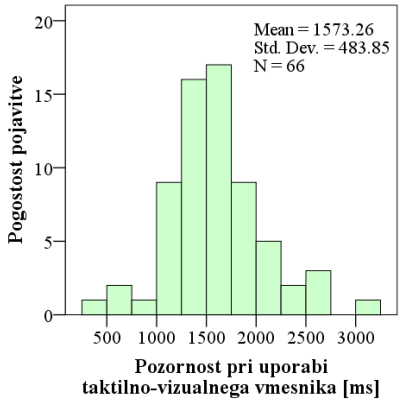
Po prevzemu vodenja smo opazovali:

- ali je voznik med reševanjem kritične situacije zaviral,
- ali je voznik zavil v pravo smer v izogib oviri,
- trajanje prvega zavoja (od začetka zavoja do ponovnega obrata volana preko mirovne lege),
- najnižjo doseženo hitrost vozila.

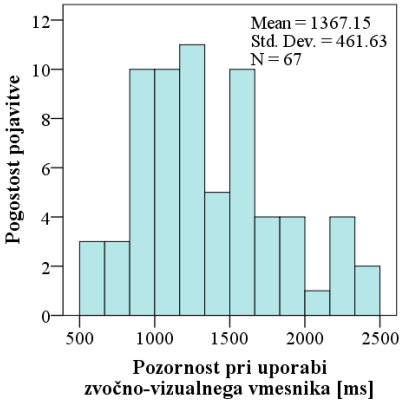
3 Rezultati

Za vsakega izmed merjenih odzivov smo izrisali histograme. Statistično pomembna razlika med uporabo taktilno-vizualnega zvočno-vizualnega vmesnika se je pokazala le pri rezultatih pozornosti ($t(131) = 2{,}51$, $p = 0{,}013$), zato so pri prikazu ostalih rezultatov uporabniki obeh vmesnikov združeni.

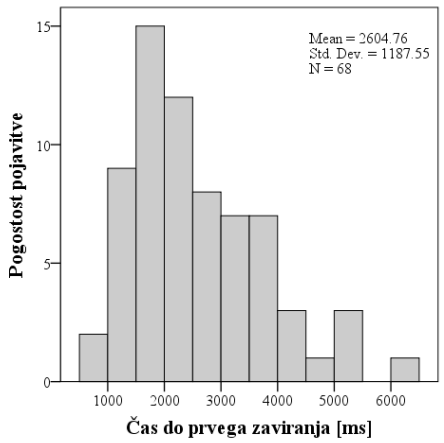
3.1 Odzivi pred prevzemom vodenja



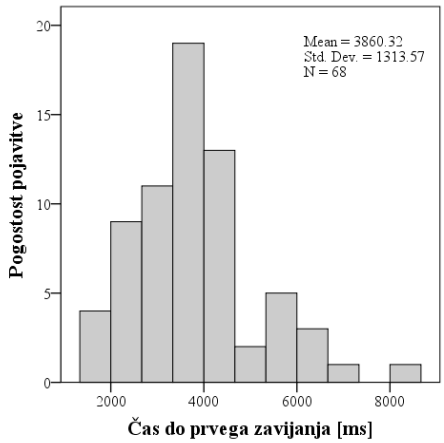
Graf 1. Histogram trajanja od TOR do prvega pogleda na zaslon pri uporabi taktilno-vizualnega vmesnika



Graf 2. Histogram trajanja od TOR do prvega pogleda na zaslon pri uporabi zvočno-vizualnega vmesnika



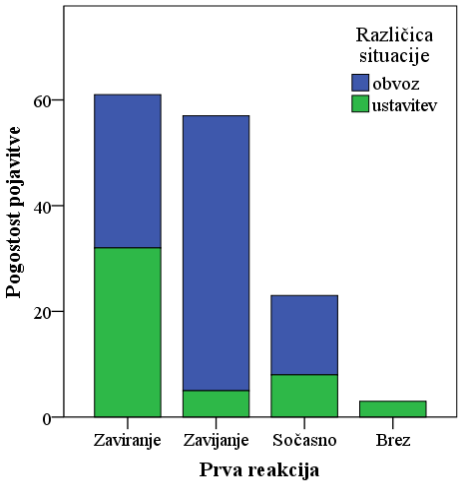
Graf 3. Histogram časov od TOR do prvega pritiska na zavoro



Graf 4. Histogram časov od TOR do prvega zasuka volana

Čas, ki so ga vozniki porabili za preusmeritev pogleda od sekundarne naloge na cesto, je bil med 350 ms in 3 s (Graf 1, Graf 2). Graf 3 prikazuje porazdelitev časov do prvega zaviranja, Graf 4 pa do prvega zavoja.

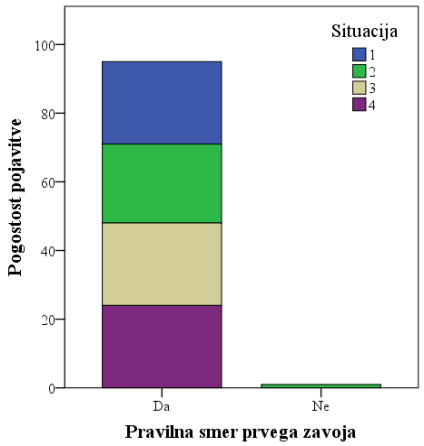
Razporeditev voznikovih prvih reakcij po zahtevi za prevzem vodenja je prikazana na Grafu 5.



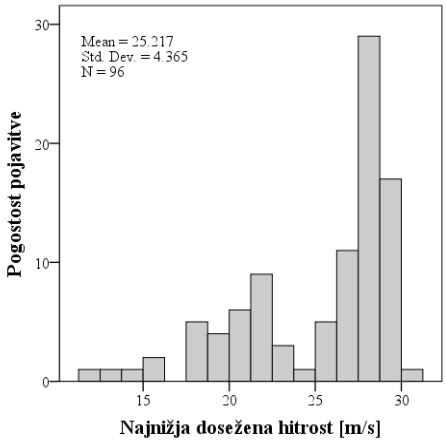
Graf 5. Porazdelitev prvih voznikovih reakcij glede na različico kritične situacije

3.2 Odzivi po prevzemu vodenja

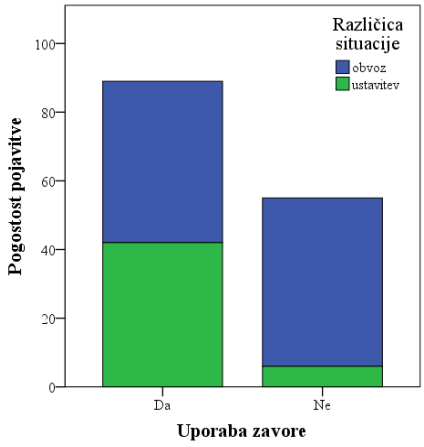
Graf 6 prikazuje pravilnost izbire smeri zavijanja glede na mesto ovire. Podane so tudi razlike glede na zaporedno situacijo posameznega uporabnika – vsak se je štirikrat srečal s situacijo, v kateri je moral obvoziti (zaviti) oviro na cesti.



Graf 6. Porazdelitev smeri prvega zavoja glede na zaporedno kritično situacijo



Graf 7. Porazdelitev najnižje dosežene hitrosti vozila med situacijo, ki je zahtevala obvoz ovire



Graf 8. Uporaba zavore med razreševanjem kritičnih situacij

Graf 7 podaja porazdelitev zabeležene najnižje hitrosti vozila v času med TOR in ponovnim vklopom

avtonomne vožnje. Na Grafu 8 vidimo, koliko voznikov je v tem času sploh uporabljalo zavoro.

4 Razprava in zaključek

Iz opazovanja odzivov voznikov med TOR in dejanskim prevzemom vodenja lahko opazimo, da je večina voznikov za preusmeritev pozornosti od sekundarne naloge k vožnji potrebovala med eno in dvema sekundama. Vozniki z zvočno-vizualnim vmesnikom so za preusmeritev potrebovali manj časa (povpr. 1,37 s ± 0,46 s) kot vozniki s taktilno-vizualnim vmesnikom (povpr. 1,57 s ± 0,48 s). Do prvega zaviranja so v povprečju potrebovali še sekundo dlje, do prvega zavoja pa še dobro sekundo več. Pri načrtovanju uporabniškega vmesnika za podajanje zahteve za prevzem vodenja je torej ključno, da upoštevamo minimalne časovne potrebe, ki jih uporabnik potrebuje za preusmeritev pozornosti in reakcijo. Večina uporabnikov je po izraženi zahtevi za prevzem vodenja najprej zavirala, kar lahko nakazuje na previdnost oz. upoštevanje preventivnih ukrepov. Kljub temu se je trikrat zgodilo, da voznik ob podani zahtevi sploh ni reagiral.

Po prevzemu vodenja so vozniki v veliki večini zavili v pravo smer glede na podano informacijo o lokaciji ovire na cesti. Največ napak je bilo v prvem primeru srečanja z oviro, kar lahko pripisujemo novosti sistema oz. nenavajenosti voznikov na simulator vožnje. V zadnji situaciji, kjer je bilo potrebno obvoziti oviro, napak o smeri ni bilo več. Pred aktivno uporabo podobnega sistema v vozilu bi bilo torej smiselno uporabnike eksperimentalno podučiti o nekaj tipičnih situacijah.

Iz Grafa 7 lahko opazimo dve vrsti voznikov – tiste, katerih najnižja hitrost je bila vseskozi okoli 30 m/s (cca. 110 km/h oz. hitrost vozila v avtonomnem režimu), ter tiste, ki so zmanjšali hitrost vožnje na okoli 20 m/s (cca. 70 km/h). Le manjše število voznikov je hitrost med obvozom ovire zmanjšalo pod 10 m/s (cca. 35 km/h). Podobne ugotovitve lahko razberemo tudi iz Grafa 8, kjer opazimo, da kar nekaj voznikov (približno tretjina) sploh ni uporabljala zavore, nekateri niti ko bi morali vozilo popolnoma ustaviti. Ker je obvoz ovire z nezmanjšano hitrostjo lahko zelo nevaren, bi bilo voznika smiselno aktivno vzpodbujati k uporabi zavore, ne glede na značilnosti kritične situacije. To lahko storimo npr. z dodatnim opozorilom, če voznik ne zavira v določenem času ter samodejnim zaviranjem po preteku kritičnega časa.

Povzamemo lahko, da so si odzivi uporabnikov na zahtevo po prevzemu vodenja pogojno avtonomnega vozila različni. Čas od podane zahteve do prve reakcije se lahko giblje vse od ene sekunde do preko štirih sekund. Z uporabo zvočnih vmesnikov lahko pripomoremo k hitrejši preusmeritvi pozornosti na vožnjo. Večina voznikov ob zahtevi najprej zavira, a obstajajo tudi taki, ki zavore do konca razreševanja situacije sploh ne uporabijo. V kolikor voznik ne zavira, bi ga lahko s primerno načrtovanim uporabniškim vmesnikom dodatno opomnili na ta varnostni ukrep. Zavedanje različnosti vozniških odzivov lahko torej pozitivno pripomore k načrtovanju uporabniških vmesnikov v pogojno avtonomnih vozilih.

Literatura

[1] „SAE International Releases Updated Visual Chart for Its “Levels of Driving Automation” Standard for Self-Driving Vehicles“. https://www.sae.org/ (pridobljeno jul. 19, 2019).

[2] E. Frøkjær, Erik, M. Hertzum, Morten, K. Hornbæk, in Kasper, „Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?“, jan. 2000, doi: 10.1145/332040.332455.

[3] V. Melcher, S. Rauh, F. Diederichs, H. Widlroither, in W. Bauer, „Take-Over Requests for Automated Driving“, *Procedia Manufacturing*, let. 3, str. 2867–2873, jan. 2015, doi: 10.1016/j.promfg.2015.07.788.

[4] K. Zeeb, A. Buchner, in M. Schrauf, „What determines the take-over time? An integrated model approach of driver take-over after automated driving“, *Accident Analysis & Prevention*, let. 78, str. 212–221, maj 2015, doi: 10.1016/j.aap.2015.02.023.

[5] M. R. Endsley, „Toward a Theory of Situation Awareness in Dynamic Systems“, *Situational Awareness*, jul. 05, 2017. https://www.taylorfrancis.com/ (pridobljeno dec. 13, 2019).

[6] A. Eriksson in N. A. Stanton, „Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control“, *Hum Factors*, let. 59, št. 4, str. 689–705, jun. 2017, doi: 10.1177/0018720816685832.

[7] T. Köhn, M. Gottlieb, M. Schermann, in H. Krcmar, „Improving take-over quality in automated driving by interrupting non-driving tasks“, v *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray, California, mar. 2019, str. 510–517, doi: 10.1145/3301275.3302323.

[8] K. Zeeb, A. Buchner, in M. Schrauf, „Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving“, *Accident Analysis & Prevention*, let. 92, str. 230–239, jul. 2016, doi: 10.1016/j.aap.2016.04.002.

[9] J. Radlmayr, V. Weinbeer, C. Löber, M. Farid, in K. Bengler, „How Automation Level and System Reliability Influence Driver Performance in a Cut-In Situation“, v *Advances in Human Aspects of Transportation*, 2018, str. 684–694.

[10] C. Gold, R. Happee, in K. Bengler, „Modeling take-over performance in level 3 conditionally automated vehicles“, *Accident Analysis & Prevention*, let. 116, str. 3–13, jul. 2018, doi: 10.1016/j.aap.2017.11.009.

[11] T. Gruden, M. Hribernik, E. Keš, J. Sodnik, in G. Jakus, „Taktilno-vizualni uporabniški vmesnik za voznikov prevzem upravljanja vozila pri prekinitvi avtonomne vožnje“, v *Zbornik osemindvajsete mednarodne Elektrotehniške in računalniške konference*, Portorož, Slovenija, 2019, str. 53–56.

[12] „Nervtech Simuation Technologies“, *Nervtech*. https://www.nervtech.com (pridobljeno apr. 29, 2019).

[13] „Tobii Pro Glasses 2 wearable eye tracker“, jun. 25, 2015. https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/ (pridobljeno feb. 06, 2020).

Interaktivna razstavna vitrina – orodje za muzejske zbirke prihodnosti

Simon Kolmanič[1], Niko Lukač[1], Selma Rizvić[2], Borut Žalik[1]

[1]Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, Maribor, Slovenija

[2] Univerza v Sarajevu, Fakulteta za elektrotehniko Sarajevo, Zmaja od Bosne, bb, Sarajevo, Bosna in hercegovina

E-pošta: simon.kolmanic@um.si

Interactive showcase – a tool for the museum collections of the future

Abstract. *In this article, an interactive showcase, based on holographic pyramid projector and Leap Motion sensor is presented. It can be used in modern museums for the exhibition of different types of artifacts.*

The idea to use the holograms in museums for the preservation of cultural heritage is not new. However, we have extended that idea by combining Unity 3D, a pyramid holographic projector and the Leap Motion sensor. As a result, a cost-efficient and attractive showcase has been obtained. The presentation in this new type of showcases can be controlled by the visitors' hand movements. The users can "hold" the artifacts, rotate them, and examine them in a way unthinkable in the real world. For the demonstration of our idea, the presentation of Roman heritage in Balkan has been developed, where digitized artifacts from the Roman era have been used.

.

1 Uvod

Ideja o uporabi hologramov za prikaz grafičnih podatkov je stara vsaj 30 let [1]. Takšna predstavitev podatkov je omogočala natančen prikaz računalniško generiranih objektov, ki pa je bil statičen. Idejo o kombinaciji hologramov in interaktivne računalniške grafike prvič zasledimo v letu 2004 [2], kjer je raziskovalec Bimber tako kombinacijo predstavil kot sprejemljiv kompromis med kakovostjo vizualne predstavitve in interaktivnostjo. Holograme je v svojem delu videl kot možen način predstavitve muzejskih artefaktov, ki bi jih z računalniško grafiko dodatno obogatili. V svoji viziji je predstavil možnosti uporabe take tehnologije v industriji, znanosti in izobraževanju.

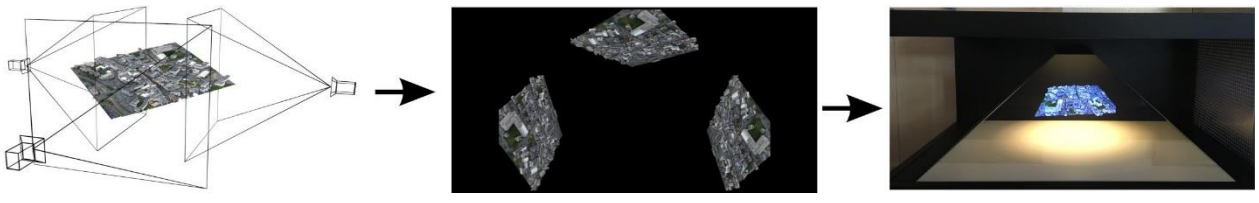
Veliko njegovih idej se je do danes že uresničilo. Barvni holografi OptoClone [3] so tako trenutno že v uporabi v vsaj štirih muzejih po Evropi in dajejo impresivne rezultate, vendar so, tako kot v svoji osnovni zasnovi izpred tridesetih let, statični.

Veliko obetavnejši v smislu interaktivnosti so se pokazali nekoliko enostavnejši holografski projektorji s piramido [4], kjer se hologram ustvari s pomočjo slike na zaslonu projektorja, ki se projicira na steklene stranice piramide (glej sliko 1). Kot lahko vidimo na sliki 1, objekt senčimo s treh različnih smeri (pri svojih poskusih smo namreč uporabljali tristrano piramido), nato pa se slike prezrcalijo na površje piramide, s čimer dobimo iluzijo, da objekt lebdi v notranjosti piramide. Ker lahko slike, ki se preslikajo na površje piramide, generirano v realnem času, lahko tak hologram odreagira na aktivnosti uporabnika.

Pri kontroliranju delovanja hologramа običajnih vhodnih naprav, kot sta miška in tipkovnica, ne moremo uporabiti. V [5] so poskusili holograme kontrolirati s kretnjami, ki so jih zajemali s pomočjo treh pravokotno postavljenih kamer. Sami smo podobne rezultate dosegli s pomočjo uporabe vhodne naprave Microsoft Kinect v2, žal pa se je njeno delovanje izkazalo za dokaj nezanesljivo, proizvajalec pa naprave tudi več ne podpira.

Zaradi omenjenih pomanjkljivosti, smo napravo Kinect zamenjali s senzorjem Leap Motion [6], [7], ki se je izkazal za veliko boljšega. Precej boljša kot pri senzorju Kinect je tudi povezljivost senzorja z grafičnim pogonom Unity, ki je v tem trenutku eno najbolj priljubljenih orodij za razvoj vsakovrstnih multimedijskih aplikacij.

S kombinacijo holografskega projektorja s piramido in senzorja Leap Motion, smo dobili interaktivni hologram. Z namenom, da testiramo primernost te naprave za predstavitve muzejskih zbirk, smo izdelali

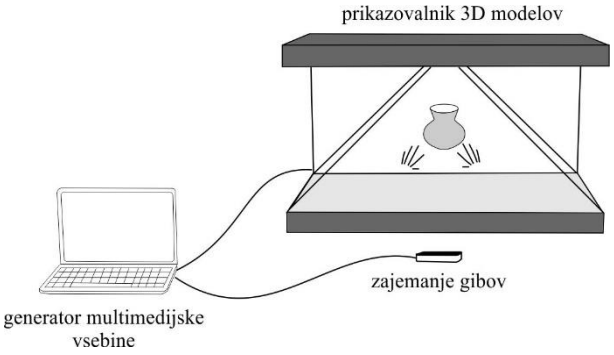


Slika 1. Proces generiranja slike v holografski piramidi.

aplikacijo za predstavitev rimskih artefaktov, najdenih na štirih najdiščih na Balkanu. Testi so pokazali, da je interaktivni hologram primeren za interaktivno razstavno vitrino, delovanje je bilo zanesljivo, naprava pa je razen krmiljenja predstavitve s kretnjami rok omogočala tudi interakcijo uporabnika s samimi razstavnimi eksponati.

2 Zasnova interaktivne razstavne vitrine

Sistem interaktivne vitrine sestavljajo holografski prikazovalnik, senzor Leap Motion in dovolj zmogljiv računalnik za generiranje vsebine predstavitve (slika 2). V našem poskusu smo uporabljali komercialni holografski projektor Dreamoc™ XL3[1] s tristrano piramido. Projektor je zasnovan okrog televizorja z ločljivostjo slike 1920 x 1080 točk in osveževanjem 30 slik/s. Ta je montiran nad holografsko piramido in skrbi za prikaz slike, ki se nato projicira na stene piramide. Z izbiro televizije za prikaz slike, so proizvajalci rešili dve težavi, holografski projektor je lahko samodejno predvajal animacije s pomnilnika USB, s čimer za svoje delovanje v osnovi ni potreboval posebnega računalnika, avtomatsko pa je imel podprt tudi zvok.



Slika 2. Shematski prikaz zasnove interaktivne vitrine.

Za zajemanje kretenj rok, smo uporabljali senzor Leap Motion. Gre za cenovno ugoden senzor, ki prihaja z dobro podporo za grafični pogon Unity. Namenjen je razpoznavi gest tako v aplikacijah navidezne resničnosti kot tudi v namiznih aplikacijah. Omogoča širok nabor funkcionalnosti, pri čemer lahko uporabnik uporablja obe roki.

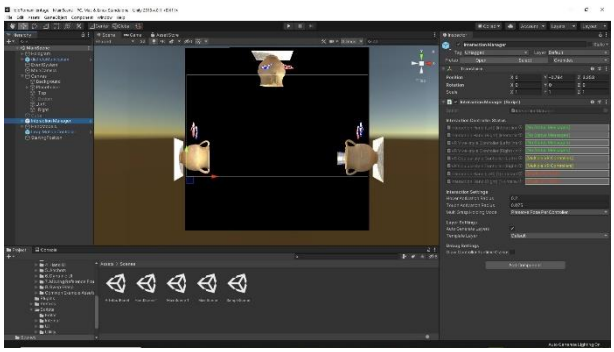
Za generiranje slike in drugih multimedijskih vsebin je potreben dovolj zmogljiv računalnik. Iz slike 1 lahko namreč vidimo, da je potrebno sceno senčiti s treh različnih pogledov. V kolikor bi naša piramida imela štiri stranice, pa bi potrebovali še en pogled več.

Grafični pogon Unity[2] nam omogoča, da pogled z vsake kamere uporabimo kot teksturo, s pomočjo katere sestavimo končno sliko, ki jo pošljemo projektorju. Čeprav je naš holografski projektor uporabljal frekvenco osveževanja z zgolj 30 slik/s, pa smo vseeno želeli zagotavljati hitrost izrisa s 60 slik/s, kar je frekvenca osveževanja klasičnega računalniškega prikazovalnika. Pri tem je treba upoštevati, da imajo objekti, ki jih pričakujemo pri razstavnih eksponatih v muzejih zelo veliko točk, kar dodatno podkrepi zahtevo po zmogljivejšem računalniku.

Ob uporabi senzorja Leap Motion se v sceni pojavi tudi par virtualnih rok, ki zelo dobro sledijo premikom naših rok v realnem svetu. S temi rokami lahko primemo objekte, jih rotiramo, ali pa z njimi krmilimo potek naših predstavitev.

3 Implementacija osnovnih funkcionalnosti interaktivne vitrine

Za izgradnjo programske opreme interaktivne razstavne vitrine, smo uporabili grafični pogon Unity. Pri svojem delu smo uporabljali verzijo 2019.4.0f1. Čeprav je bil igralni pogon Unity v začetku namenjen predvsem hitremu razvoju 2D in 3D iger, pa ga danes vse bolj uporabljamo tudi za razvoj številnih drugih aplikacij [8], z močno multimedijsko komponento. Posebej priljubljen je za razvoj aplikacij za navidezno in mešano resničnost [9], kjer je v tem trenutku prva izbira.



Slika 3. Razvoj interaktivne razstavne vitrine v grafičnem pogonu Unity.

Na sliki 3 vidimo pogled glavne kamere na platno s tremi teksturami rimskega vrča. Teksture v realnem času generirajo tri kamere, postavljene, kot to prikazuje slika 1. Tako sam objekt kot tudi preostale tri kamere so v primerjavi s platnom tako majhne, da jih glavna kamera ne zazna, ampak vidi samo platno s teksturami. Tak sistem prikazovanja je bil v Unity razvit zaradi implementacije daljnogledov na strelnem orožju v strelskih igrah, ko je potrebno oddaljene objekte z daljnogledom močno približati. Cena tovrstnega prikazovanja je, še posebej v našem primeru, v precejšnjem povečanju obremenitve računalnika, saj mora sceno senčiti večkrat. Obremenitev lahko zmanjšamo, če namesto treh različnih pogledov na sceno, na stene piramide projiciramo en sam pogled, česar v večini primerov uporabniki niti ne opazijo. Žal je v

[1] https://www.realfiction.com/solutions/dreamoc-xl3, zadnji obisk 6. 7. 2020

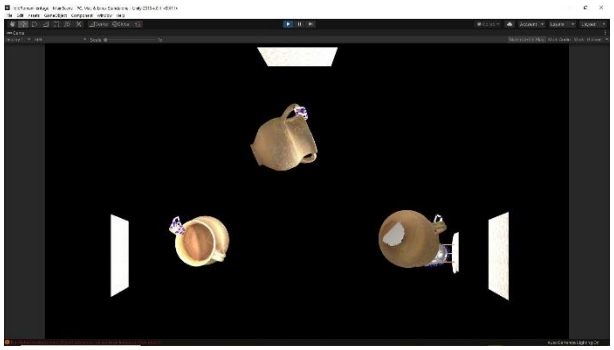
[2] https://unity.com/, zadnji obisk 6. 7. 2020

našem primeru zaradi interakcije uporabnika z objekti to neizvedljivo, saj z objektom upravlja en sam uporabnik. V kolikor bi uporabili samo en pogled, bi bilo to za neaktivne uporabnike moteče, saj bi bili postavljeni v center dogajanja, nad katerim pa ne bi imeli nadzora.

Postavitev slik objekta z različnih pogledov in njihovo velikost je potrebno prilagoditi konfiguraciji holografskega projektorja, pri čemer zasledujemo kompromis med velikostjo prikazanega objekta in možnostjo prekrivanja slik proti vrhu piramide. Do prekrivanja prihaja, ker se piramida proti vrhu oži in se tako del slike projicira na sosednjo ploskev. Tega efekta ne moremo povsem odpraviti, lahko ga zgolj omilimo z uporabo določenih mask, ki pa morajo sovpadati z geometrijo stranic holografske piramide. Drugi način odprave tega efekta je, da omejimo premikanje objekta na samo tisti del vidnega volumna, kjer do težav s prekrivanjem ne prihaja.

Zaradi omejenega prostora, ki ga imamo na razpolago pri postavitvi scene, je uporaba tabel s tekstovnimi opisi razstavnih eksponatov zelo otežkočena. Tabla namreč ne sme v celoti zapolniti vidnega polja kamere, saj v tem primeru izgubimo iluzijo lebdenja objekta v notranjosti piramide. V sami sceni mora namreč prevladovati črna barva, ki pri hologramih pomeni prozorno barvo.

Problem rešimo tako, da namesto pisnih opisov uporabimo zvočno predstavitev artefaktov, kar pa je lahko moteče, v kolikor bi imeli v prostoru več tovrstnih vitrin. Predstavitve so se močno izboljšale tudi z uporabo glasbe v ozadju. V našem testnem primeru smo uporabili glasbo, ki bi naj bila značilna za rimsko obdobje.



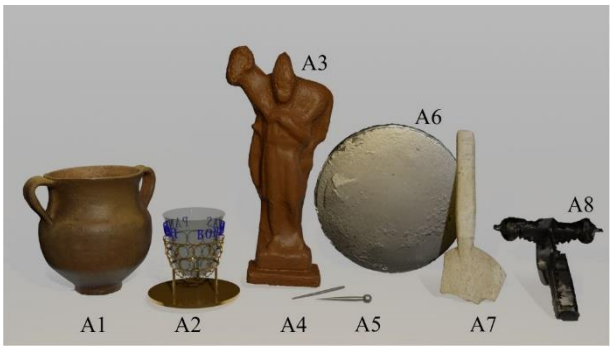
Slika 4. Interakcija uporabnika z objektom v okolju Unity. Uporabnik objekt drži z roko.

Posebna prednost interaktivne vitrine pa je možnost interakcije uporabnika z objektom. Ob objektu se namreč prikaže model rok, s katerim lahko objekt primemo, ga poljubno obračamo ter z njim zadevamo ob druge objekte v sceni. Ker roke v hologramu sledijo našim rokam, dobimo občutek, kot da bi objekt dejansko držali v rokah (slika 4). Ta funkcionalnost bo dodana vrednost h klasičnim muzejskim razstavam, kjer lahko objekte zgolj opazujemo, sedaj pa se jih lahko dotaknemo brez vsakega strahu pred njihovimi poškodbami ali krajo.

Za razliko od očal za navidezno in obogateno resničnost, gre v tem primeru za tehnologijo, ki ne zahteva fizičnega kontakta z uporabnikom, zato je zelo primerna za uporabo tudi v času širjenja virusa Covid 19.

4 Priprava predstavitve rimske dediščine na Balkanu

Za testiranje funkcionalnosti interaktivne razstavne vitrine smo pripravili predstavitev rimske kulturne dediščine Balkana, ki zajema artefakte s štirih nekdanjih rimskih naselbin: Viminacium in Municipium v Srbiji, Aquae v Bosni in Hercegovini ter Dyrrachium v Albaniji. Zbirko sestavlja skupno osem digitaliziranih artefaktov, ki jih lahko vidimo na sliki 5.



Slika 5. Rimski artefakti, najdeni v štirih najdiščih na Balkanu.

Število točk in trikotnikov, ki sestavljajo mreže posameznih objektov, lahko vidimo v tabeli 1.

Tabela 1. Število točk in trikotnikov v mrežah objektov iz slike 5.

ime	št. točk	št. trikotnikov
A1	1.635.245	3.269.837
A2	193.895	386.184
A3	225.241	449.457
A4	63.007	125.916
A5	98.387	196.160
A6	46.912	92.712
A7	73.741	146.808
A8	212.460	424.181

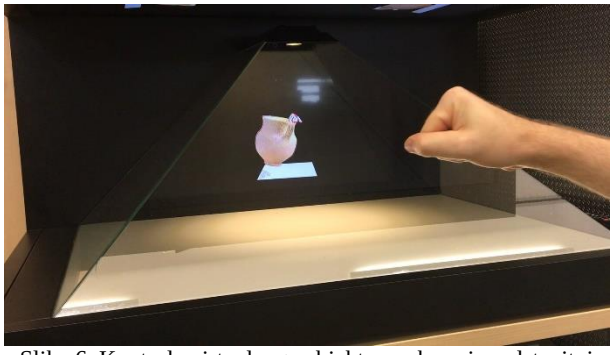
Skupno so objekti na disku zasedali 512 MB prostora, medtem ko je zvokovna predstavitev objektov znašala zanemarljivih 2.4 MB.

Tako objekte kot tudi njihove predstavitve je bilo potrebno uvoziti v grafični pogon Unity, treba je bilo definirati tudi materiale objektov, saj tisti iz animacijskih paketov niso primerni za realno-časovno senčenje. Za vsakega izmed objektov smo ustvarili ločeno sceno, saj smo lahko zaradi pomanjkanja prostora v holografski piramidi, prikazovali naenkrat zgolj en objekt. Vse scene so ponujale enako funkcionalnost, se pravi, da se je ob pritisku na objekt sprožila njegova predstavitev, ki smo jo lahko s ponovnim pritiskom na artefakt zaustavili. Objekte smo lahko tudi »prijeli« in obračali z gibi rok v realnem svetu. S kretnjami rok smo se lahko premikali med posameznimi artefakti. Celoten projekt predstavitve rimske dediščine na Balkanu v pogonu Unity je na disku zasedal 1.6 GB prostora.

5 Rezultati in diskusija

Predstavitev rimske dediščine na Balkanu je bil realen test interaktivne razstavne vitrine. V naših poskusih smo uporabljali prenosnik s procesorjem i7-8750 s šestimi jedri, 16 GB pomnilnika in diskretno grafično kartico NVIDIA GeForce GTX 1060. Uporabljeni objekti so bili po zahtevnosti enaki tistim, ki jih lahko pričakujemo v tovrstnih muzejskih zbirkah. Testiranje je potekalo v okviru laboratorija, testni uporabniki pa so morali po kratki demonstraciji samostojno upravljati s predstavitvijo. Ob demonstraciji smo poskusili uporabniku dati toliko informacij o upravljanju s predstavitvijo, kot bi jih dobil ob opazovanju drugih uporabnikov predstavitve. Pri tem smo opazovali, koliko funkcionalnosti, ki jih predstavitev ponuja, so uporabniki preizkusili.

Predstavitev je tekla gladko in brez zapletov. Kljub temu so testni uporabniki potrebovali nekaj časa, da so se naučili zagrabiti objekt. Ko so objekt enkrat »držali v rokah«, uporabili so lahko tako desno kot levo roko, so z njim manipulirali brez težav. Primer take manipulacije lahko vidimo tudi na sliki 6, kjer uporabnik »v roki drži« rimski vrč. V kolikor uporabnik objekt dvigne iz podlage in ga izpusti, objekt pade iz predstavitve, a se čez eno sekundo ponovno postavi na svojo izhodiščno mesto.



Slika 6. Kontrola virtualnega objekta z roko pri predstavitvi rimske dediščine na Balkanu.

Možnost uporabe predstavljene interaktivne vitrine je ogromna. Poleg same uporabe v muzejih, jih lahko uporabimo v izobraževanju, marketingu, predstavitvah novih produktov, v turizmu in še kje. Vitrina omogoča predvajanje multimedijskih vsebin, ima pa tudi zabavno komponento interakcije z objekti, kar jo dela še posebej privlačno za otroke.

Kot morebitne omejitve lahko navedemo, da je holografski projektor velika in težka naprava, kar pomeni, da je njen transport lahko težaven, zahteva pa tudi veliko prostora za postavitev. Moteče je lahko tudi to, da je senzor Leap Motion z računalnikom povezan preko kabla, s katerega dolžino smo lahko dodatno omejeni pri sami postavitvi celotnega sistema.

Čeprav smo v sami predstavitvi uporabljali zgolj manjše predmete, pa lahko v vitrini prikazujemo tudi stavbe, celotna mesta, uporabniku pa lahko omogočimo tudi veliko večjo stopnjo interakcije z objektom. Nadgradnja trenutne predstavitve bi lahko bil prikaz rekonstrukcije štirih rimskih naselbin, kjer bi lahko uporabnik klikal na posamezne zgradbe, s čemer bi lahko dostopal tako do posameznih podatkov o stavbah, pa tudi predmetih, ki so jih v teh stavbah uporabljali pri vsakodnevnem delu. S tem bi dobili zanimivo interaktivno predstavitev, ki bi bila veliko zanimivejša od klasičnega pristopa.

Vitrino bi lahko dodatno izboljšali z uporabo večjega holografskega projektorja, saj bi lahko bili v tem primeru prikazani objekti večji, kar bi bilo gotovo koristno, odprlo pa bi še dodatne možnosti uporabe. Poiskati pa bo potrebno tudi načine, kako za prikaz najboljše izrabiti celoten volumen holografske piramide, da pa pri tem slika objekta ne bi prehajala med posameznimi pogledi in se tako preslikala na stranico piramide na katero ne sodi. To bo naše delo za prihodnost.

Zahvala

Avtorji izjavljamo, da je raziskavo finančno podprla Javna agencija za raziskovalno dejavnost Republike Slovenija v okviru programa P2-0041 in projekta BI-BA/19-20-003.

Literatura

[1] M. Yamaguchi, N. Ohyama, and T. Honda, ‘Holographic 3-D printer’, in *Practical Holography IV*, 1990, vol. 1212, pp. 84–92.

[2] O. Bimber, ‘Combining optical holograms with interactive computer graphics’, in *ACM SIGGRAPH 2005 Courses*, 2005, pp. 4–es.

[3] A. Sarakinos and A. Lembessis, ‘Color Holography for the Documentation and Dissemination of Cultural Heritage: OptoClonesTM from Four Museums in Two Countries’, *Journal of Imaging*, vol. 5, no. 6, p. 59, 2019.

[4] T. Yamanouchi, N. Maki, and K. Yanaka, ‘Holographic Pyramid Using Integral Photography’, in *2nd World Congress on Electrical Engineering and Computer Systems and Science (EECSS’16)*, 2016, pp. 3–6.

[5] D. Patel and P. Bhalodiya, ‘3D Holographic and Interactive Artificial Intelligence System’, in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2019, pp. 657–662.

[6] W. Lu, Z. Tong, and J. Chu, ‘Dynamic hand gesture recognition with leap motion controller’, *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1188–1192, 2016.

[7] X. Min, W. Zhang, S. Sun, N. Zhao, S. Tang, and Y. Zhuang, ‘VPModel: High-fidelity product simulation in a virtual-physical environment’, *IEEE transactions on visualization and computer graphics*, vol. 25, no. 11, pp. 3083–3093, 2019.

[8] C. Cheng, L. Kang, S. Cai, J. Li, and R. Sun, ‘Virtual Display and Interactive Experience Platform of Farming Culture Based on Unity3D’, *IFAC-PapersOnLine*, vol. 51, no. 17, pp. 637–642, 2018.

[9] S. L. Kim, H. J. Suk, J. H. Kang, J. M. Jung, T. H. Laine, and J. Westlin, ‘Using Unity 3D to facilitate mobile augmented reality game development’, in *2014 IEEE World Forum on Internet of Things (WF-IoT)*, 2014, pp. 21–26.

Platforma za razpoznavo 3D gest s pomočjo električnega polja

Aljaž Blatnik[1]

[1] *Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana*
E-pošta: aljaz.blatnik@fe.uni-lj.si

Platform for recognizing 3D gestures using an electric field

Abstract. *The office environment presents a unique challenge in designing a system for recognizing 3D gestures of the user. A person usually sits at a table, with his hands on a keyboard or mouse and limited space to perform gestures.*

The work presents the design and development of a prototype of a 3D gesture recognition platform that obtains information from the environment by changing the electric field, without the use of process-intensive image processing, or the need for visibility with a sensor. The connection to the computer is made via a USB bus.

Implemented and presented are 6 gestures that allow background management with various daily programs on the computer. A study of the usability of the product on 6 people is presented.

Ključne besede: 3D geste, upravljanje z gestami, električno polje

1 Uvod

Čeprav so pogosto vir raziskav [1], se zračne geste v vsakdanjem življenju uporabljajo pravzaprav precej redko. Svoje veliko tržišče so našle v igračarskem svetu navidezne 3D resničnosti, pri vsakdanjih opravilih pa uporabniki raje posežejo po uveljavljenih uporabniških vmesnikih, kot so fizična stikala in zasloni na dotik. Raziskave uporabniške izkušnje kažejo [2, 3, 4], da uporabniki želijo čim hitrejši odziv na dano zahtevo, torej hitrejšo interakcijo, hkrati pa se niso pripravljeni učiti novih veščin za uporabo tehnologije [5], če se le ta ne prilagaja njihovim potrebam in željam, ter hitro izgubijo potrpežljivost in interes za rokovanje, v primeru neodzivnosti sistema za razpoznavanje [6].

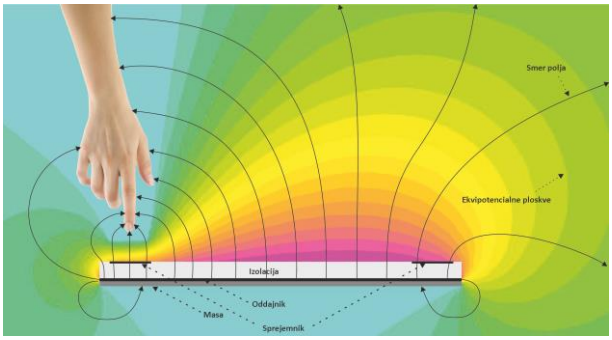
Avtorji se pri razvoju sistema za upravljanje z gestami v veliki večini primerov poslužujejo dveh tehnologij. Neposredna namestitev senzorjev na uporabnika (zapestnice, ure), ali detekcija s pomočjo video kamere in razpoznave z namensko programsko opremo (Xbox Kinect, Nintendo Wii) [7, 8]. Uporabniki navadno ne želijo nositi namenske opreme za razpoznavo, ali si jo nameščati pred uporabo, zato ta pristop ni širše sprejet med njimi [9]. Avtorji kot razlog uporabe video detekcije in razpoznave navajajo nizko ceno senzorjev in podprtost z razvojnimi orodji. Vse to seveda ni dovolj dober razlog, za proglasitev izbranega načina kot najustreznejšega, kot to lahko zasledimo pri K. Djadu [8].

Poleg cene ustrezne opreme, ter zahtevnosti računske obdelave, dodatno težavo predstavlja vsakodnevna uporaba. Kamera mora biti nameščena na statični lokaciji, brez neposredne sončne svetlobe, ki lahko moti ustrezno razpoznavo, pred začetkom uporabe pa mora biti izveden kalibracijski postopek [2]. Uporabnik mora stati na določeni razdalji od kamere in biti za dobro razpoznavo obrnjen proti njej. Takšne zahteve lahko uporabnika odvrnejo od uporabe sistema in s tem upravljanja z gestami [6].

Namen tega dela je prikaz novega, drugačnega pristopa pri razpoznavi gest in interakciji z uporabnikom, ki si še mora zagotoviti svoje mesto na tržišču. Podan je primer v pisarniškem okolju z upravljanjem funkcij v ozadju (glasnost predvajanja, izbira pesmi, listanje).

2 Delovanje senzorja

Osnovna zamisel pri izgradnji in načrtovanju platforme za zaznavo gest je uporaba električnega polja, oziroma njegove spremembe za detekcijo položaja objekta v prostoru. S tem izločimo potrebo po video razpoznavanju, ter se izognemo težavam s prekrivanjem leč ali senzorja z umazanijo, ovirami, tudi slabimi pogoji za zajem (slaba osvetlitev, močno sonce).



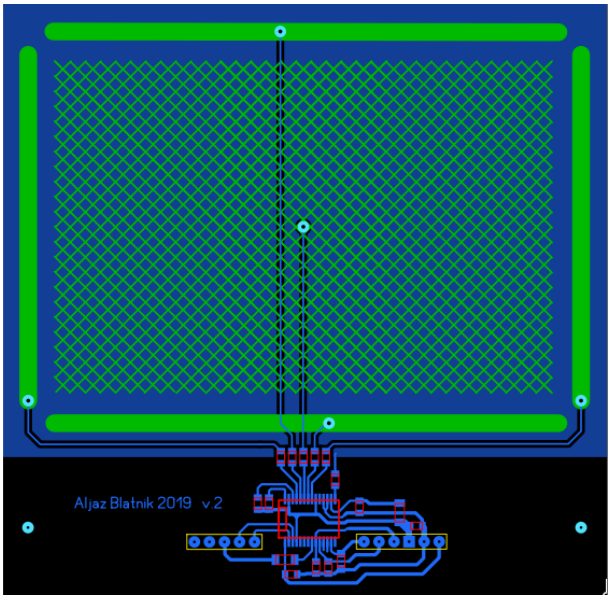
Slika 1: Prikaz vpliva roke na električno polje.

Osnovni princip delovanja prikazuje Slika 1. Senzor vsebuje eno oddajno in več sprejemnih elektrod. Oddajno elektrodo, ki je od sprejemnih ločena z izolativno plastjo, vzbujamo z znano frekvenco in amplitudo. Sprejemne elektrode skupaj z maso tvorijo kondenzator zelo majhne vrednosti. Kondenzator z aktivnim vezjem tvori nihajni krog, čigar časovna konstanta je primarno odvisna od velikosti izražene kapacitivnosti. Kadar v bližino senzorja pride človeška roka, ta predstavlja dodatno kapacitivnost v okolici, s tem pokvari električno polje v bližini senzorja in spremeni časovno konstanto nihajnega kroga dane elektrode. Tako lahko na tehnološko precej enostaven način zaznamo prisotnost človeške roke.

Na podoben način delujejo tudi zasloni s kapacitivnim zaznavanjem dotika, le da tam zaznavamo

dotik v dvo-dimenzijski ravnini, pri 3D senzorju pa smo sposobni zaznavati tudi oddaljenost telesa od površine.

Z ustrezno razporeditvijo elektrod in uporabo več vzorčevalnih kanalov lahko tvorimo senzor, ki je sposoben zaznave položaja roke v prostoru, seveda na določeni razdalji od senzorske naprave. Pri tem je priporočljiva uporaba 5. sprejemnih elektrod za razpoznavo osnovnih gest (levo desno, gor, dol, rotacija), od katerih sredinska elektroda služi namenu boljše kalibracije in detekcije prisotnosti objekta.



Slika 2: Končni izgled dvoslojne tiskanine senzorja.

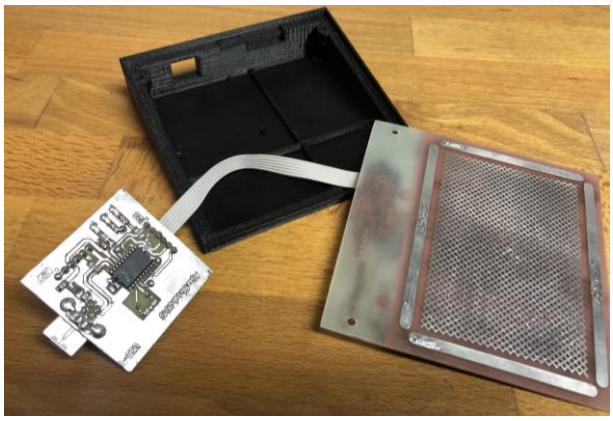
Končno zasnovo sprejemnih elektrod z namenskim čipom prikazuje Slika 1. Sredinska elektroda je izdelana s karo vzorcem, saj tako dobimo približno vsoto površine bakra preostalih 4. elektrod in olajšamo izvedbo meritve (večja površina pomeni večjo kapacitivnost, s tem manjšo resolucijo zaznave objekta). Izolacijsko območje tvori dielektrična plast laminata FR4, na spodnji strani tiskanine pa je izvedena oddajna elektroda.

Čeprav lahko sprejemno/obdelovalno vezje zgradimo sami z uporabo povsem analognih gradnikov, je na tržišču za ta namen na voljo integrirano vezje, ki v svoji notranjosti vsebuje tako analogni del za zajem podatkov, kot preprost mikrokrmilnik za grobo obdelavo podatkov, shranjevanje parameterizacije in komunikacijo preko vodila. V prototipu je uporabljeno osnovno namensko integrirano vezje podjetja Microchip MGC3030 [10], saj je bilo v času razvoja edino prosto dobavljivo, ustreza vsem pogojem za zagotavljanje skladnosti s predpisi in vsebuje algoritem za izločanje motilnih statičnih pojavov (prisotnost objektov, motilni signali).

3 Zajem podatkov in povezava preko USB

Izdelan detektor brez namenske programske opreme ne deluje, saj detektorsko integrirano vezje za prvo delovanje potrebuje vpis parameterizacijskih podatkov in zadnje različice programske kode, ki jo prosto dostopno ponudi proizvajalec. V ta namen je potrebno izdelati USB-I2C vmesnik, ki kasneje služi za komunikacijo z računalniškimi programi in napajanje same naprave. Pri tem ohranjamo vsebino podatkov, ki jih detektor posreduje uporabniku in s tem omogočamo razvoj poljubne programske opreme na osebnem računalniku.

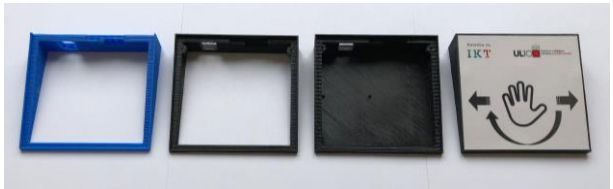
Poleg surovih podatkov (vrednost deviacije signala na sprejemnih elektrodah) lahko iz namenskega čipa prejmemo že obdelan podatek o zadnji izvedeni gesti. S tem prihranimo precej obdelovalne moči na osebnem računalniku, kjer se posvetimo le upravljanju programov in prikazu podatkov. Končen izgled senzorskega in komunikacijskega dela pred namestitvijo v ohišje prikazuje Slika 3.



Slika 3: Končna povezava elementov pred montažo v ohišje.

4 Ohišje in kalibracija

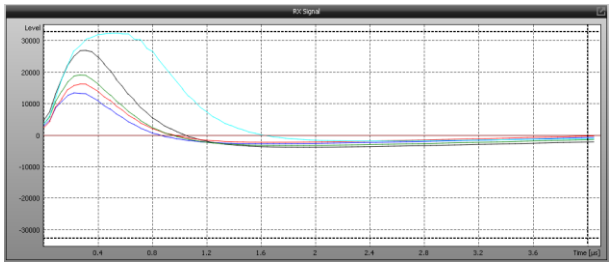
Ohišje senzorja poleg estetskega vidika in mehanske trdnosti, služi tudi kot pomoč pri uspešni zaznavi gest. Tablica je namenjena uporabi na pisalni mizi, torej ko uporabnik sedi. Takrat je gibanje rok omejeno na prostor nad mizo, ravno tako pa zaradi normalnega položaja sedenja gibi niso vedno vzporedni s površino mize. Naklon zaznavne površine senzorja za 10° se je skozi več prototipov pokazal za najprimernejšega v večini primerov, ko uporabnik sedi in upravlja z gestami. Razvoj nosilnega ohišja prikazuje Slika 4.



Slika 4: Razvoj ohišja končne naprave.

Pred začetkom uporabe senzorja, moramo celotno napravo kalibrirati. To je potrebno storiti le enkrat, torej ko tablico sestavimo oziroma vanjo vpišemo program (tovarniški postopek), nato se naprava samodejno prilagaja spremembam v okolici, da detekcija gest ves čas deluje pod enakimi obratovalnimi pogoji. Za ta namen proizvajalec ponuja programsko orodje, ki nam omogoča meritev parametrov in njihov vpis v pomnilnik integriranega vezja.

Kalibracija je potrebna zaradi vpliva ostalih tiskanih vezji v napravi in samega ohišja, na razporeditev električnega polja okoli sprejemnih elektrod. S tem izničimo neenakomerne amplitude v sprejetem signalu, ki niso posledica prisotnosti objekta, ampak motilnih elementov v okolici. Slika 5 prikazuje postopek merjenja kalibracijskih faktorjev pred vpisom v pomnilnik.



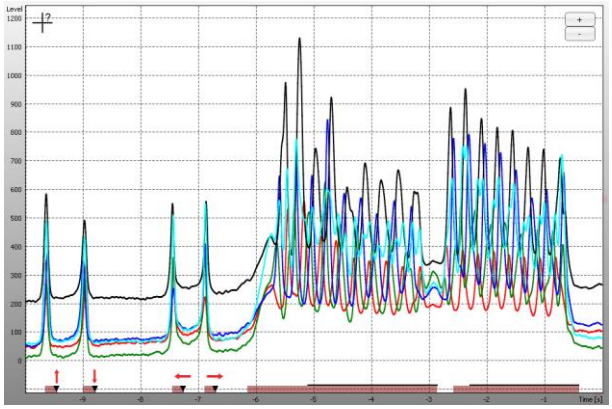
Slika 5: Določanje kalibracijskih vrednosti.

Iz grafa lahko opazimo lastnost velikih elektrod (cian barva krivulje), kjer sredinska elektroda kljub manjšanju površine s pomočjo karo vzorca tvori precej večjo kapacitivnost z maso, kot preostale elektrode. Za boljše delovanje bi bilo potrebno njeno površino še nekoliko zmanjšati, s kalibracijo pa uspemo težavo delno omiliti.

5 Razpoznava gest in delovanje 3D tablice

Program znotraj integriranega čipa na senzorski tiskanini preko oddajne elektrode proizvaja električno polje v frekvenčnem območju nekaj 10kHz. S spreminjanjem frekvence lahko manjša ali veča zaznavno polje in s tem omogoča razlikovanje med prisotnostjo objekta (človeške roke) ali napake kot posledica motečih pojavov (zunanje elektronske naprave, elektorstatika).

Posamezna gesta je razpoznana s pomočjo končnih stanj namenske programske kode naložene v notranjost procesne enote senzorja. Ta deluje precej preprosto. Pomik roke iz leve strani proti desni pomeni detekcijo signala na zahodni elektrodi, nato na vzhodni, medtem ko se severna in južna v jakosti signala skozi čas bistveno ne spreminjata. Podobno velja za preostale tri geste premika roke v smeri neba. Detekcijo in klasifikacijo geste v časovnem poteku signala prikazuje Slika 6.



Slika 6: Signali elektrod v časovni domeni (razpoznava gest).

Senzorska procesna enota lahko razpozna več različnih gest, tudi prisotnost roke, dotik površine in dvojno ponovitev gest. Uporabnik lahko kreira nove geste (npr. poševni zamah z roko), jih klasificira in vpiše v pomnilnik, vendar se to izkaže kot precej nezanesljiv način detekcije. Proizvajalec je v svoj nabor že vključil najbolj optimalne geste, ki jih je mogoče razpoznati v različnih primerih, pod različnimi pogoji. Slika 6 prikazuje šest različnih tipov gest, ki jih razpozna prototipna tablica.



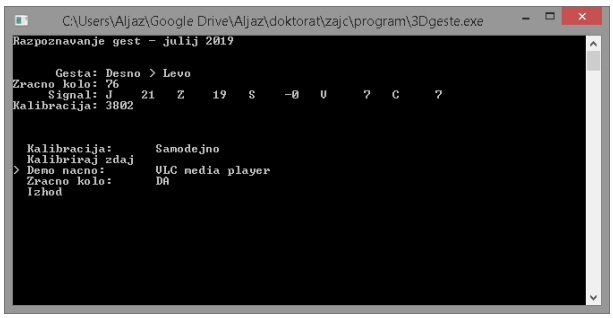
Slika 6: Različni tipi gest.

Za namen prikaza uporabnosti danega prototipa, je bila razvita namenska preprosta programska oprema za osebni računalnik s sistemi Windows. Program v ukazni vrstici prikazuje trenutne vrednosti signala na elektrodah, ter s simulacijo pritiska hitrih tipk, omogoča uporabo v ozadju za tri demo primere.

Besedilni način ob vsaki uspešni detekciji v kateri koli program za urejanje besedila izpiše besedo, ki predstavlja zadnjo prepoznano gesto.

VLC media player način omogoča nadzor nad multimedijskim predvajalnikom VLC. Zamah levo in desno izbira med pesmimi iz seznama predvajanja, zamah gor prične s predvajanjem, zamah dol ga ustavi. Zračno kolo v smeri urnega kazalca veča glasnost predvajanja, zračno kolo v nasprotni smeri pa jo niža.

Demo različica Acrobat reader omogoča upravljanjem s pregledovalnikom dokumentov .pdf. Zamah gor odpre celozaslonski način branja, zamah dol preide v način pregledovanja, zamaha levo in desno izbirata po straneh. Zračno kolo deluje različno v dveh različnih pogledih. V pregledovanju počasi drsi po strani dokumenta, v celozaslonskem načinu pa hitro izbira med stranmi.



Slika 7: Izgled programa za upravljanje z gestami.

Slika 7 prikazuje končni izgled terminalskega programa za prikaz upravljanja s 3D gestami. Podatki, ki jih tablica pošilja preko vodila USB so nespremenjeni in

ustrezajo strukturi podatkov, ki jih je predvidel proizvajalec integriranega vezja. S tem delujejo tudi vsi primeri in koda, ki jo na svoji spletni strani ponuja podjetje Microchip.

6 Uporabniška izkušnja

Uporabnost upravljanja z gestami, je bila preizkušena na 6 različnih osebah, ki so poskusile z upravljanjem računalniških programov s 3D gestami. 4 osebe so bile računalniško dobro pismene, dve osebi pa računalnika ne uporabljata pogosto.

Uporabnik je najprej preizkusil detekcijo gest z besedilnim načinom, da se je spoznal z delovanjem naprave in načinom zaznave. Nato je preizkusil upravljanje multimedijskega predvajalnika glasbe in branja PDF dokumentov na osebnem računalniku. Vsi uporabniki so sedeli za mizo, pred seboj imeli računalnik, desno ali levo od njih pa tablico za upravljanje.

Z opazovanjem je bilo ugotovljeno, da uporabniki hitro osvojijo geste zamaha (levo, desno, gor, dol), tudi, če niso vešči upravljanja z računalnikom, težava pa jim povzroča gesta zračnega kolesa. V 66% primerov je bilo uporabnikom potrebno najprej demonstrirati uporabo in način izvajanja geste, šele nato so razumeli, kako jo uporabljati. Povprečen čas za osvojitev geste je bil približno 3-4min, nato so uporabniki samostojno nadaljevali z upravljanem. Izkazalo se je tudi, da je najlažji način učenja s prikazom drsenja .pdf dokumenta in ne besedilnim načinom, kot je bilo sprva predvideno.

Na 3. uporabnikih je bil test ponovljen po enem tednu. Ugotovljeno je bilo da si osvojeno gesto zračnega kolesa uspešno zapomnijo in lahko takoj pričnejo z upravljanjem.

Uporabniki so tudi navedli občutljivost sistema kot negativno lastnost, torej da lahko gesto sprožijo pomotoma med delom za pisalno mizo. Težavo je možno odpraviti na dva načina. Prvi omogoča zmanjšanje občutljivosti, vendar potem ne moremo več zagotavljati pravilnosti delovanja v vseh primerih (npr. kadar senzor založimo s kupico papirja), druga možnost je detekcija prisotnosti in nastavljanje ustreznega pragu. Torej je gesta veljavna le, kadar se nad celotno površino nahajala roka, v nasprotnem primeru geste ne prepoznamo. S hitrimi ponovitvenimi testi se je to sicer izkazalo kot delno uspešno, saj uporabniki geste s časoma začnejo izvajati površno (hiter, nepopoln zamah), sistem pa jih v teh primerih ne prepozna kot veljavne.

7 Zaključek

3D geste in upravljanje z njimi brez uporabe video kamere in obdelave posnetka za detekcijo, predstavljajo nov način razpoznavanja na tržišču in odpirajo številne možnosti. Ker je senzorska površina lahko povsem skrita za okrasnimi elementi ali ohišjem, je vgradnja možna na številnih mestih, predvsem tam, kjer ostale rešitve niso mogoče zaradi različnih dejavnikov (sterilnost kirurških dvoran, umazanija avtomehaničnih delavnic). Z razvojem prototipa je bila zgrajena platforma za spoznavanje s 3D načinom gest, testiranjem njihove uporabnosti in enostavnosti vpeljave v široko-potrošno proizvodnjo. Za namen preverjanja uporabniške izkušnje je bil izdelan preprost program, ki lahko teče v ozadju osebnega računalnika in omogoča upravljanje različnih programov z gestami. Uporabnost je bila preizkušena na več prostovoljcih, ki so sistem uspešno osvojili.

Po koncu razvoja prvega prototipa, je proizvajalec na tržišče poslal drugo različico čipa, ki omogoča detekcijo na več kanalih hkrati in hitrejšo sprotno obdelavo s kombinacijo 2D in 3D razpoznave. Tako se odpira možnost izdelave sledilne ploščice (track pad), ki v osnovi deluje enako kot obstoječi izdelki na tržišču (uporabniki torej znajo upravljati z njimi), dodaja pa novo možnost upravljanja z gestami, torej dvig roke od senzorske površine in upravljanje v prostoru. Za razvoj takega prototipa bi bil potreben dodaten čas, tudi industrijska izdelava tiskanine, namesto izdelave v domači delavnici.

8 Literatura

[1] T. S. Daniel Ashbrook, „MAGIC: A Motion Gesture Design Tool,“ v *CHI 2010: Everyday Gestures*, Atlanta, 2010.

[2] O. J. Fails, „A design tool for camera-based interaction,“ v *Proc. CHI*, 2003.

[3] V. D. K. N. Y. Chin, „Development of an instrument measuring user satisfaction of zhe human-computer interface,“ v *In Proc. CHI*, 1988.

[4] J. C. F. I. Y. Cui, „A cross culture study on phone carrying and physical personalization,“ v *HCI International*, 2007.

[5] Y. K. Y. Q. a. P. O. K. Miguel A. Nacenta, „Memorability of pre-designed & user-defined gesture sets,“ v *CHI 2013*, Paris, 2013.

[6] A. G. W. S. E. J. d. M. F. Ahmed Sabbir Arif, „Error behaviours in an unreliable in-air gesture recognizer,“ v *One of a CHInd*, Toronto, 2014.

[7] V. G.-P. M. A. S. Arnaud Ramey, „Integration of a Low-cost RGB-D sensor in a social robot for gesture recognition,“ v *HRI'11*, Lausanne, 2011.

[8] C. C. P.-L. S.-C. M. D. Khalid Djado, „Gesture interface for an interactive kiosk,“ v *ACM*, Shenzhen, 2014.

[9] R.-D. Vatavu, „Fundamentals of gesture production, recognition and analysis,“ v *CHI 2017*, Denver, 2017.

[10] Microchip Inc., „MGC3030,“ 2019. [Elektronski]. Available: http://ww1.microchip.com/downloads/en/DeviceDoc/MGC3030-3130-3D-Tracking-and-Gesture-Controller-40001667F.pdf. [Poskus dostopa 2019 7 4].

Opis delokroga za produkcijo 360-stopinjskega glasbenega videospota
Description of a framework for the production of a 360 degree music video

Veronika Malačič[1], Marko Cafnik[2], Helena Gabrijelčič Tomc[1]

[1]Univerza v Ljubljani, Naravoslovnotehniška fakulteta, Oddelek za tekstilstvo, grafiko in oblikovanje, Snežniška 5, 1000 Ljubljana

[2] Multiversum GmbH, Kauergasse 10/7, Dunaj, Avstrija

E-pošta: malacic.veronika@gmail.com

Abstract. *The aim of this contribution was to describe the optimal framework for the production of a 360-degree music video on the basis of a script proposal. Within the scope of the research work, the concept of the creative integration of classical videos, 360° videos and virtual reality was developed and implemented. The result was a 360° audiovisual medium in which a combination of three technologies was refined into a new form of digital mixed medium. The experimental part presents the design of the sequence of work phases from the planning of the video, ideation, the preparation, the production and the final editing and export, which were optimized to an efficient framework during the pre-production, production and post-production process. The discussion critically evaluates the work stages in which the production of the medium was the greatest challenge and which can serve as a guide and recommendation for the creators who will take the path of producing a 360-degree video.*

Key words: 360° music video, video, virtual reality, immersion, production

1 Uvod

Virtualna resničnost je področje, ki se v zadnjem času izjemno hitro razvija. Gre za medij kot nosilec resničnosti, ki lahko izkustvo videospotov ponese na povsem nov nivo. Ena izmed vej virtualne resničnosti je tudi 360-° video, ki udeleženca prestavi v virtualni svet z možnostjo ogleda okolja na vse strani sfere. Na področju glasbene industrije omogočata tehnologiji virtualne resničnosti in 360-° videa prehod iz pasivnega opazovanja videospotov na celostno izkušnjo, prisotnost in raziskovanje v prostoru medija, interakcijo in potopitev. 360-stopinjski video, imenovan tudi panoramični, sferični ali video virtualne resničnosti, je vrsta multimedijskega sporočanja, ki zajame celotno 360-° sferično območje prostora medija. Omogoča nam bogatejšo izkušnjo, saj v primerjavi s tradicionalnimi 2D videji uporabniku omogoča večjo potopitev v virtualno okolje. Uporabnik izkustveno popolnoma zamenja pozicijo objektiva kamere, s pogledom zajame celotno 360-° sferično območje videa okoli nje in tako sam z usmerjenostjo svojega pogleda in svojim doživljanjem režira video. Na tak način uporabniku omogočimo občutek prisotnosti [1-3]. V virtualni resničnosti se uporabljata tridimenzionalna slika in video- ali računalniška grafika za ustvarjanje in sintetiziranje novega, realnosti bolj ali manj podobnega sveta, ki naj bo tudi prepričljiv za uporabnika [4]. Tehniki 360-° posnetkov in virtualne resničnosti dajeta gledalcu popolno svobodo, da gleda okoli sebe, obenem pa zmanjšujeta možnost, da filmski ustvarjalci usmerjajo pozornost gledalca, kar je vzrok za iskanje drugačnih načinov za uspešno pripovedovanje zgodbe, načrtovanje in produkcijo v primerjavi s klasično videoprodukcijo [5, 6].

Ključni izziv za učinkovito pripovedovanje zgodb s pomočjo 360-° videoposnetkov je, kako usmeriti pozornost gledalcev na pomembne vsebine ne da bi jim pri tem odvzeli svobodo razgledovanja po prostoru. Ker je ključna prednost 360-° videoposnetka sposobnost gledalčevega nadzora nad lastnim pogledom, moramo gledalcu omogočiti, da ohrani sposobnost samoupravljanja in usmeri pogled z uporabo prefinjenih, nevsiljivih tehnik [5, 7]. Newton in Soukup predstavita zanimivo lastnost pripovedovanja zgodbe v 360-° maniri. Zaključek njune študije pravi, da mora pripovedovalec razmišljati ne le iz perspektive režiserja, temveč mora prevzeti vlogo matadorja z mahanjem v smeri, v katero si želi, da občinstvo teče, tako da občinstvo oblikuje zgodbe glede na svoje doživljanje [7]. Primer dobre prakse VR pripovedovanja je nagrajeni animirani glasbeni video *Pearl*. V videu uporabnik nastopa kot duh gledalca, ki sedi v avtu in opazuje, kako poskuša samski oče vzgojiti svojo mlado hčer. Režiser Patrick Osborne nas vodi skozi zgodbo s pomočjo glasbe, ki jo ustvarjata oče in hči, preko enotne perspektive s sovoznikovega sedeža pa lahko gledalec usmerja svoj pogled na vse dogajanje [8]. Kot drugi primer dobre prakse lahko navedemo vodenje pogleda s pomočjo dinamične koreografske uprizoritve. Koreografinja Anthony Van Laast in režiser Luis Blackaller sta v 360-° videospotu skladbe *Waterloo* s pomočjo plesa gledalca vodila po prostoru in v celoti izkoristila 360-° pripovedovanje zgodbe. Gledalec je postavljen v vlogo glavne igralke in skozi zgodbo se spreminja tudi njegova pozicija [9].

S pregledom referenc ugotavljamo, da lahko z združitvijo glasbenega videospota s 360-° tehnologijo ustvarimo globljo potopitev udeleženca, hkrati pa to za ustvarjalce pomeni nove oblike kreativnega izražanja. Namen prispevka je bil določiti optimalni delokrog produkcije 360-° glasbenega videospota za skupino Žavbi Brothers

s kombinacijo klasičnega videa, 360-° videa, 3D računalniško generiranih objektov in virtualne resničnosti. Pri tem je bil fokus tudi na pripovedovanju zgodbe, ki je morala biti prilagojena 360-° tehnologiji, s čimer je bil razvit 360-° način pripovedništva, načrtovanja in produkcije.

2 Eksperimentalni del

Načrtovanje delokroga je zaradi predhodnih izkušenj z videoprodukcijo temeljilo na klasičnem načrtu, kar pomeni, da smo začeli s fazo predprodukcije in načrtovanjem izdelka. Sledilo je snemanje oz. faza produkcije, zadnja faza pa je bila obdelava materiala ali postprodukcija, v kateri smo pregledali, zrezali in sestavili posnet material. Bistvo nadgradenj delokroga je bilo v posameznih fazah, ki so pri našem delu in opirajoč se na narativo pesmi temeljile na 360-° načrtovanju, kot prikazuje slika 1. Kot načrtovalci smo za vsako fazo morali preklopiti na 360-° perspektivo organizacije samega dela, testnih posnetkov, priprave snemanj, snemanja in postprodukcije.

2.1 Predprodukcija

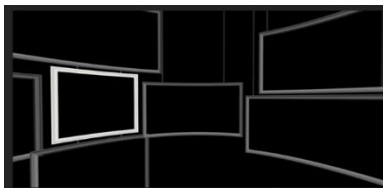
Ideja. Sestavljanje zgodbe videospota in pisanje pesmi sta nastajala vzporedno. Ker smo se odločili za predstavitev zgodbe v sferi, smo želeli tudi pesem prilagoditi in ustvariti na tak način. Ustvarili smo zgodbo, ki nas vodi skozi življenje dveh mladih ljudi, para, ter njun krog življenja in ljubezni. Dramska napetost se tekom videospota stopnjuje in doseže vrh s selitvijo glavne junakinje, razrešitev pa predstavljata njeno spoznanje, kako močno pogreša bližino druge osebe, in njena želja po ponovnem srečanju. Pripovedna struktura je klasična in je sestavljena iz treh osnovnih dejanj.

Scenarij in produkcijski plan. Pri pisanju scenarija za videospot smo upoštevali ritem, časovno razporeditev in besedilo skladbe. Poskušali smo tako vizualne elemente kot dinamiko posnetkov podrediti skladbi. Obliko scenarija za snemanje 360-° videospota smo morali modificirati, saj ne vsebuje dialogov, poleg tega pa se zgodba dogaja na dveh različnih pozicijah v sferi z dvema različnima igralcema.



Slika 1. Produkcijski delokrog.

Posledično oblika napisanega scenarija ni bila podobna tradicionalnim scenarijem, ampak avdio-vizualnemu (AV) formatu. Na osnovi scenarija je nastal produkcijski plan za vse snemalne dni. ***Lokacija in priprava prostora.*** Končni izdelek – 360-° videospot – je kombinacija poustvarjenega navideznega prostora in 360-° snemanja v prostoru. Prvi prostor, namenjen za uporabo v 360-° maniri, v katerem se gledalci med gledanjem videospota nahajajo, smo izdelali s pomočjo programa Adobe Premiere Pro. Pri izdelavi prostora smo bili pozorni na število okvirjev in njihovo velikost ter postavitev. Za sobo, polno okvirjev, smo se odločili, ker bodo videji v okvirjih simbolizirali dogodke iz preteklosti in na ta način v treh minutah predstavili zgodbo para. V zadnji sceni okvirji zbledijo in gledalci bodo ponovno postavljeni v sedanjost, v 360-° prostor – galerijo. Slika 2 predstavlja osnutek 360-° virtualnega prostora, kjer so bili okvirji 3D zmodelirani in uporabljeni kot prehod med različnimi tehnikami.



Slika 2. Osnutek poustvarjenega prostora s slikami.

2.2 Produkcija

Pri snemanju smo uporabili: brezzrcalni digitalni fotoaparat Sony α7 III, objektiv Sony FE 24–70 mm F2,8 GM, objektiv Sony FE 35 mm F1,8, objektiv Sony Sonnar T* FE 55 mm F1,8 ZA, stativ Manfrotto MVK504AQ, Hoya Variabilni ND filter, monitor Feelworld F5, aputure LS C300d, aputure Light Dome II, 5x SanDisk Extreme Pro 128GB, 360-° kamero Insta360 Pro in brezpilotni letalnik DJI Mavic 2 Zoom. Nastavitve za snemanje 2D posnetkov so bile Full HD (100 fps) in za snemanje s 360° kamero pa 8k in 30 fps.

2.3 Postprodukcija

Ker smo se v fazi produkcije posluževali različnih tehnik snemanja, je bila faza postprodukcije izjemnega pomena, saj smo posneti tehniki združili in ustvarili izdelek z obogateno sporočilnostjo. 360-° videospot smo sestavili iz posebej posnetih videov formata 16:9, ki smo jih umestili v okvirje, ter zadnje scene, ki je bila posneta s pomočjo 360-° kamere. Pomembno je bilo, da smo v nastavitvah sekvence omogočili VR projekcijo in uporabo monoskopskega pogleda ter 360-° pogleda v vodoravni in 180-° pogleda v navpični smeri.

Groba montaža. Pri grobi montaži smo izbrane posnetke in oblikovane okvirje uvozili v program ter jih razdelili po mapah. Najprej smo na časovno premico porazdelili okvirje v smiselno kompozicijo. Okvirji so bili v velikostnem razmerju 4:3 – standardni, 16:9 – širokozaslonski in 1:1 – kvadratni.

Fina montaža. V tej stopnji smo poskrbeli za sinhronizacijo in ritem. Uredili smo dolžino vseh sekvenc in kadrov. Pomagali smo si z markerji, saj smo lahko na podlagi natančnega scenarija, v katerem smo vnaprej definirali prihode igralcev v kadre in odhode iz njih, določili začetne in končne pozicije intervalov.

Šivanje, retuširanje. Preden smo uvozili 360-° posnetek, je bilo treba vseh šest širokokotnih posnetkov združiti v en 360-° panoramski video. Šivanje videomateriala je potekalo v programu Insta360Sticher. Videoposnetki so bili s pomočjo prekrivanja združeni v enega in tako smo dosegli mehke in neopazne prehode med njimi. Z retuširanjem smo odstranili tripod in noge stojala.

Začetno točko 360-° posnetka smo določili z uporabo efekta, imenovanega VR Rotate Sphere. Ta nam je omogočil nadzor nad začetno usmerjenostjo pogleda gledalcev ob prvem ogledu videa.

Združitev prostorov in maskiranje. Poustvarjeni navidezni prostor in 360-° posnetek smo združili na interaktiven način, tako da v določeni točki pesmi vsi okvirji razen enega iz prvega prostora s pomočjo optičnih prehodov potemnijo in izginejo v črnino. Preostali okvir je bil zamaskiran v realni prostor, t. j. v galerijo (s programom After Effect), ki smo jo posneli s 360-° kamero, in je predstavljal povezavo med obema svetovoma. Tehniko maskiranja smo uporabili na scenah, kjer se posnetki videotehnologije mešajo z ozadjem računalniško generiranega virtualnega prostora.

Barvna korekcija, grafični elementi in izvoz. Na vsakem posnetku smo posamezno uredili kontrast, belino, ekspozicijo, nasičenost barv in ostale parametre vse do želenega videza. Video smo opremili še z naslovom, grafičnimi elementi in odjavno špico in tako je bil dokončno urejen. Za izvoz medija smo izbrali kodek HEVC (H.265), 8k ločljivost posnetka, frekvenco prikazovanja 25 fps in odkljukali polje prikaza videa kot VR.

3 Rezultati in razprava

V rezultatih in razpravi podajamo najpomembnejše ugotovitve opisa optimalnega delokroga produkcije, kjer se klasični, 360-° video, virtualna resničnost in objekti 3D računalniške grafike sestavljajo v enoten medij, t. j. glasbeni videospot. *VR prostor* uporabnikom omogoča raziskovanje povsem novega, računalniško ustvarjenega okolja. V našem izdelku to predstavlja virtualni svet okvirjev, v katere smo umestili 2D posnetke. Prostor *360-° videa* pa prikazuje realno, naravno okolje. Videoposnetek smo posneli v živo in v primeru 360-° videa nismo ustvarjali računalniško generiranih slik. 360-° produkcija in kombinacija z virtualnimi svetovi zahtevata *natančno igranje akterjev* (v smislu časovnega poteka in časovnice). V našem primeru je zgodba temeljila na premiku akterjev skozi okvirje, zato je bil ključen čas vstopa in izstopa iz kadra. Vse kadre je bilo treba snemati v celotni dolžini pesmi, saj se tudi v trenutku, ko igralci ne igrajo v določenem okvirju, slika premika. Ker smo nekaj kadrov snemali tudi na javnih površinah, je bilo zelo zahtevno ohranjati kader prazen skozi celotno dolžino pesmi (moteči avtomobili, mimoidoči ...). Vnaprej smo predvideli manjše napake v časovnih intervalih in posledično nastavili frekvenco prikazovanja na 100 fps, kar nam je omogočilo reševanje težav v fazi postprodukcije.

Režiranje 360-° snemanja je omejeno. Za snemanje 360-° posnetka smo uporabili 360-° kamero Insta360 Pro, ki nam je omogočila, da smo lahko s pomočjo aplikacije (Insta360 Pro Camera control app) »v živo« gledali posnetke in spremljali dogajanje. Problem je nastal, ker se posnetki ne predvajajo stoodstotno v živo, temveč z nekajsekundnim zamikom, kar je v primeru našega dela pomenilo, da je bil proces režije omejen. Nujno je bilo med vsako ponovitvijo režirati in deliti navodila igralcem, kako naj odigrajo naslednjo ponovitev, igralci pa so morali odigrati določen prizor z natančnostjo ene sekunde.

Osvetljevanje 360-° posnetkov je treba skrbno načrtovati že v predprodukciji. V našem primeru smo 360-° video snemali v galeriji, torej v interierju, kar je omogočalo konstantno kvaliteto osvetljevanja po celotnem prostoru ter prihranilo veliko časa pri sami produkciji, saj ni bilo potrebno dodatno osvetljevati določenih delov 360-° prostora.

Kvaliteta posnetkov 360-° kamere je ključna. Pri naši produkciji smo imeli možnost snemati z Insta360 Pro, ki je nudila snemanje visokokakovostnih videoposnetkov z ločljivostjo 7680x3840 (8K).

Kombinacija štirinajstih Full HD (100 fps) posnetkov z računalniško generirano sceno in posnetkom v 8k 360-° prostoru je zahtevala zelo zmogljivo računalniško opremo. V našem primeru je bil to računalnik z visoko zaslonsko resolucijo (5K), zmogljivim 10-jedrnim procesorjem Intel Xeon W, 2TB SSD prostora in grafično kartico Radeon Pro Vega 64. Po naših ugotovitvah so to tudi minimalne zahteve za postprodukcijo.

Osvetlitev VR prostora in objektov ter osvetlitev 360-° prostora se morata ujemati. To lahko dosežemo tako, da uporabimo za povezovanje enako sliko za VR in 360-° prostor. Slika tako odigra vlogo vezne slike in omogoča bolj realističen prehod. V našem primeru je bilo treba objekte VR prostora (okvirje) 3D zmodelirati in ne le uporabiti ukrivljene 2D slike v prostoru. Na ta način smo lahko ročno prilagajali oddaljenost objektov od kamere,

spremenili lečo kamere in upravljali z osvetlitvijo prostora za doseganje bolj verodostojnih rezultatov.
Proces maskiranja elementov pri kombinaciji tehnik z namenom ustvarjanja vizualne enotnosti prostorov je kompleksen. V naši produkciji smo s pomočjo maskiranja 2D video (maskiranje igralca) umestili v 360-° prostor in tako združili obe tehniki. Maskiranje je bilo zaradi kombiniranja kompleksno, zahtevalo je obilo časa in pojavljalo se je veliko tehničnih težav. Slika 3 prikazuje proces maskiranja.



Slika 3. Proces maskiranja.

Izzivi kombinacij elementov (posnetki, 2D grafika, 3D grafika) različnih prostorov. V našem primeru je problem predstavljala predvsem ukrivljenost elementov (okvirjev) v VR prostoru, ki je bila drugačna kot ukrivljenost elementov (okvirjev) v drugem 360-° posnetem prostoru. Na sliki 4 je prikazana končna postavitev kombinacije klasičnega videa, 360-° videa, 3D ustvarjenih objektov in virtualne resničnosti v skupen prostor medija, opazna pa je tudi različna ukrivljenost okvirjev.



Slika 4. Končna kompozicija 360-° videospota: (a) združitev 3D objektov in 2D videov, (b) izgled 360-° videa, (c) simulacija pogleda 1. prostora v 360-° formatu, (d) simulacija pogleda 2. prostora v 360-° formatu.

Omejeno delovanje orodij programske opreme za ustvarjanje učinkovitih kombinacij VR prostora in 360-° posnetkov. Premiere Pro (še) ne ponuja možnosti izbire leč VR kamere ter ročne postavitve kamere za VR v prostoru. Kamera je lahko samo statična in ima določeno lečo, kar pomeni, da sta ukrivljenost in oddaljenost 2D elementov že vnaprej določeni. Program tudi ne ponuja dovolj efektov, s katerimi bi se lahko naknadno ročno preoblikovalo 2D elemente v prostoru in s tem vplivalo na njihovo ukrivljenost. V našem primeru smo zato imeli težave pri prehodih iz VR prostora v 360-° posnetek preko enega okvirja, stena namreč ni bila enako oddaljena od pozicije kamere v obeh prostorih, kar je predstavljalo različno ukrivljenost objektov v prostoru in s tem nerealen izgled kombinacije elementov.

4 Zaključki

V prispevku predstavljamo eno od panog novih tehnologij, 360-° video, ki smo ga povezali z glasbo in tradicionalnim videom ter umestili v prostor virtualne resničnosti. Rezultat raziskovalnega dela je na slovenski glasbeni sceni novost, ki bo objavljena na omrežju Youtube in si jo bo možno ogledati s pametnim telefonom, računalnikom, tabličnim računalnikom ali s pomočjo preprostih očal za virtualno resničnost.
Rezultati in diskusija prispevka ponujajo praktična priporočila za ustvarjanje 360-stopinjskih avdio-vizualnih glasbenih spotov. Projektno delo je bilo tako tehnično kot umetniško izjemno zahtevno, pri čemer ocenjujemo, da je imela največji pomen sama ideja. Kot ustvarjalci smo izstopili iz linije ugodja predhodnih izkušenj in se soočili z izzivi ustvarjanja vizualne umetnosti s kombinacijo ploskih, tridimenzionalnih in 360-° tehnik. Na podlagi našega dela potrjujemo, da je vodenje gledalca oziroma pripovedovanje zgodbe v 360-° prostoru izjemnega pomena. Upoštevali smo vse strani sfere, ki v virtualnem prostoru obkrožajo gledalca, ter mu ponudili možnost izbire in posledično interakcije z medijem.

Literatura

[1] M. Almquist, V. Almquist: Analysis of 360° Video Viewing Behaviours. 2018, http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-144907
[2] E. Bonner, R. Lege: 360 Videos in the Classroom: A How-to Guide. 2018, let. 42, str. 29–31.
[3] M. Wohl: The 360° Video Handbook: A step-by-step guide to creating video for virtual reality (VR), https://www.goodreads.com/book/show/36383094-the-360-video-handbook
[4] G. C. Burdea, P. Coiffet: Virtual Reality Technology. John Wiley & Sons, 2003, 472 str.
[5] A. Sheikh, A. Brown, M. Evans: Directing attention in 360-degree video, IBC 2016 Conference
[6] A. Schmitz, A. MacQuarrie, S. Julier, N. Binetti, A. Steed: Directing versus Attracting Attention: Exploring the Effectiveness of Central and Peripheral Cues in Panoramic Videos, 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)
[7] L. Tong, S. Jung, R. Lindeman: Action Units: Directing User Attention in 360-degree Video based VR. 2019, str. 1–2.
[8] L. Zhang, D. Bowman, C. Jones: Virtual, Augmented and Mixed Reality. Applications and Case Studies, str. 410–425.
[9] Mamma Mia! Here We Go Again (in 360°) in Eight Takes Flat. https://blog.wevr.com/mamma-mia-here-we-go-again-in-360-1372

Urejanje in dostava video posnetkov z oblačno arhitekturo mikrostoritev

Uroš Zoretič[1], Grega Jakus[2]

[1]*Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, Ljubljana, Slovenija*
[2] *Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, Slovenija*
E-pošta: uros.zoretic123@gmail.com, grega.jakus@fe.uni-lj.si

Abstract. *Fast distribution of multimedia on the Web, including video content, is crucial to achieve the desired visibility of an organization. To accomplish this goal, several patents and commercial web video editing products have been developed. They are, however, hard to integrate into state-of-the-art content delivery systems based on cloud computing technologies.*

The building blocks of cloud-based applications are microservices. Cloud-based applications usually have complex topology and run in an unpredictable cloud environment. Therefore, each microservice must be virtualized for efficient use of resources and managed by orchestration platforms. To ensure reliable communication among microservices, service meshes which also provide service discovery, load balancing and request routing, are used.

To improve the efficiency of editing and delivery of video clips, we developed a cloud-based application with microservices as its building blocks. Each microservice manages an individual task, e.g. uploading, cutting, creating, streaming, and storing video clips and their metadata. The application was developed using latest popular technologies for communication among microservices, virtualization, orchestration and monitoring.

1 Uvod

Učinkovita in hitra distribucija video posnetkov na različna spletišča je ključnega pomena z vidika prepoznavnosti posameznika in podjetij. Video posnetke je običajno pred objavo potrebno urediti in jih pretvoriti v format, primeren za predvajanje na spletu. Za urejanje posnetkov lahko uporabimo klasična montažna orodja, kot so AVID, Adobe Premiere Pro in Final Cut Pro X. Če želimo zagotoviti predvajanje videa na spletu brez zatikanja, z minimalno zakasnitvijo in možnostjo prilagajanja njegove kvalitete razpoložljivi pasovni širini, je potrebno video posnetke predvajati po delčkih (ang. »chunks«), ki se med predvajanjem v ozadju ločeno prenašajo s strežnikov [1]. S prihodom standarda HTML5 (Hypertext Markup Language revision 5) se video delčki najpogosteje prenašajo s protokolom HTTP (Hypertext Transfer Protocol). Za opis, kje na strežnikih se posamezni video delčki nahajajo, se uporabljajo indeksne datoteke v formatu XML (Extensible Markup Language). Najpogosteje uporabljeni protokoli za pretočno predvajanje video posnetkov na spletu so MSS (Microsoft Smooth Streaming), MPEG-DASH (Moving Picture Experts Group – Dynamic Adaptive Streaming over Hypertext Transfer Protocol) in HLS (Hypertext Transfer Protocol Live Streaming) [1, 2].

Slabost klasičnih orodij in tehnik za urejanje in dostavo video posnetkov je njihova neučinkovitost v primeru, ko je potrebno na spletu objaviti posnetke, ki so bili že predhodno pripravljeni za objavo na platformah VOD (Video On Demand). Čeprav je takšne video posnetke za ponovno objavo navadno potrebno le malenkost spremeniti, je čas njihove obdelave in objave približno sorazmeren dolžini posnetka, kar je zelo neučinkovito.

Dejstvo, da so bili videi že pripravljeni za objavo na spletu, lahko izkoristimo s spletnim urejevalnikom, ki omogoča neposredno manipulacijo s posameznimi video delčki. Na tem področju obstajajo patenti [3, 4] in komercialne rešitve, kot je Typito. Problem obojih je, da so izdelani po klasičnem vzorcu odjemalec-strežnik in bi jih bilo težko razširiti, nameščati, vzdrževati in integrirati v obstoječe sisteme za dostavo vsebin.

Namen prispevka je predstaviti sistem za urejanje in dostavo video posnetkov, ki temelji na oblačni arhitekturi mikrostoritev in rešuje probleme obstoječih rešitev, saj skrajšuje procesa montaže in dostave za splet pripravljenih video posnetkov.

2 Oblačna arhitektura mikrostoritev

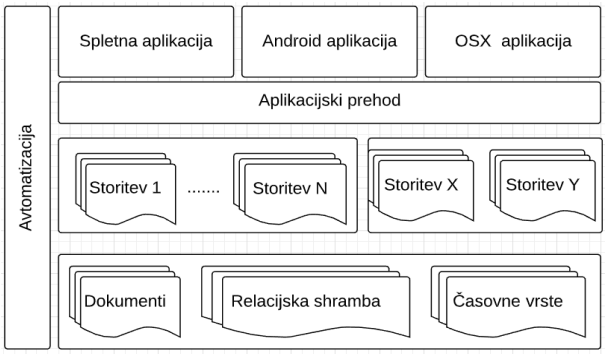
Aplikacije, ki se izvajajo v oblaku, so porazdeljene, prilagodljive, razširljive, odporne na napake, visoko razpoložljive in največkrat sestavljene iz samostojnih enot (mikrostoritev) [5]. Za doseganje omenjenih lastnosti je ključnega pomena ustrezno nadziranje in upravljanje z mikrostoritvami, ki te aplikacije sestavljajo. V nadaljevanju so opisani pomembnejši elementi oblačne arhitekture mikrostoritev.

2.1 Mikrostoritve

Mikrostoritve so gradniki aplikacije, specializirani za določeno funkcionalnost. Lahko so ustvarjene na zahtevo, horizontalno razširljive (več primerkov iste mikrostoritve si razdeli obremenitev), odporne na napake in šibko sklopljene [6]. To pomeni, da posamezna mikrostoritev ni odvisna od delovanja ostalih, temveč so zanjo pomembni le podatki, ki jih prejme in odda. Vsaka mikrostoritev je običajno upravljana neodvisno od ostalih. Arhitektura aplikacije, sestavljena iz mikrostoritev, je prikazana na Sliki 1.

Ker se obremenitev mikrostoritev nenehoma spreminja, se lahko ob velikem številu zahtevkov odzivni časi močno povečajo. Zato je potrebno

zagotoviti prilagajanje števila primerkov mikrostoritve količini zahtevkov ter mednje porazdeliti obremenitve (ang. »load balancing«), tako da vsak primerek opravi približno enako količino dela. Uporabniškim aplikacijam želimo običajno skriti kompleksnost zalednega sistema, za kar uporabimo aplikacijske prehode [6].



Slika 1. Arhitektura aplikacije, sestavljene iz mikrostoritev [6].

Da je aplikacija v oblačnih okoljih odporna na napake, prilagodljiva, horizontalno razširljiva in visoko razpoložljiva poskrbita virtualizacija in orkestracija.

2.2 Virtualizacija mikrostoritev

Virtualizacija omogoča neodvisno upravljanje mikrostoritev z izolacijo, več-najemniškim modelom (ang. »multitenancy«) ter razdeljevanjem in rezervacijo virov, kot so procesorske zmogljivosti ter delovni in trajni pomnilnik. Klasičen način virtualizacije predstavlja uporaba t.i. navideznih strojev (ang. »Virtual Machine«, VM), ki temeljijo na virtualizaciji strojne opreme. Za virtualizacijo mikrostoritev so navidezni stroji neučinkoviti, saj se prepočasi zaženejo in porabijo preveč sistemskih virov [7].

Primernejša je uporaba t.i. *lahke virtualizacije* (ang. »lightweight virtualization«) z vsebniki. Ta se zažene hitreje, saj učinkoviteje izrablja sistemske vire. Te si namreč vsebniki medsebojno delijo, medtem ko ima posamezen navidezni stroj svoje vire izolirane. Vsebniki omogočajo tudi boljšo prenosljivost in interoperabilnost, saj skrijejo kompleksnost in raznolikost programskih jezikov, ogrodij, arhitekturnih vzorcev, vmesnikov in operacijskih sistemov [7].

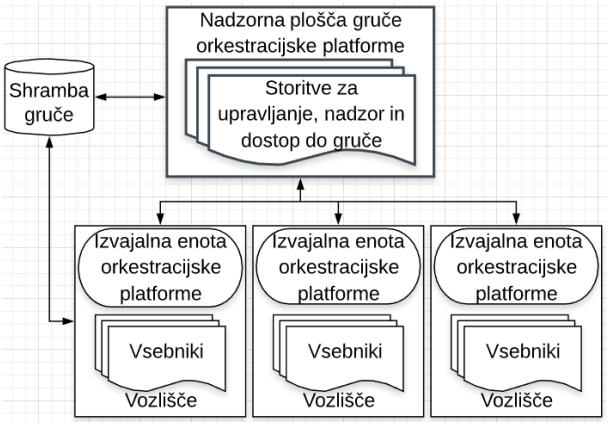
Za različne operacijske sisteme obstajajo različne lahke virtualizacijske tehnologije. Za operacijski sistem Windows je razširjena tehnologija Sandboxie, na Linuxu pa lahko uporabimo tehnologije Docker, LXC in OpenVZ. Najbolj razširjeni so vsebniki Docker, ki smo jih uporabili tudi v naši aplikaciji.

2.3 Orkestracija mikrostoritev

V produkcijskem okolju je vsebnikov pri lahki virtualizaciji zelo veliko. Ker so lahko zaradi različnih napak vsebniki v nekem trenutku neodzivni, jih je za zagotovitev nemotenega delovanja aplikacije potrebno učinkovito upravljati. Ker bi bilo ročno upravljanje zamudno, uporabimo t. i. *orkestracijske platforme*.

Orkestracija vsebnikov omogoča učinkovito upravljanje z viri, samodejno horizontalno razširljivost vsebnikov, visoko razpoložljivost, majhne režijske stroške, deklarativni model upravljanja z nastavitvenimi datotekami, odkrivanje storitev v gruči (ang. »cluster«), razdeljevanje obremenitve, preverjanje zdravja, samozdravljenje, posodobitve in nadgradnje [8].

Gruče orkestracijskih platform običajno sestavljajo nadzorna plošča, ki nadzoruje delovanje sistema, shramba podatkov za njeno konfiguracijo in posamezna vozlišča (ang. »node«). V vozliščih izvajalne enote različnih orkestracijskih platform skrbijo za pravilno delovanje vsebnikov, s katerimi so virtualizirane mikrostoritve [8]. Splošna arhitektura gruče orkestracijskih platform je prikazana na Sliki 2.



Slika 2. Arhitektura gruče orkestracijskih platform [8].

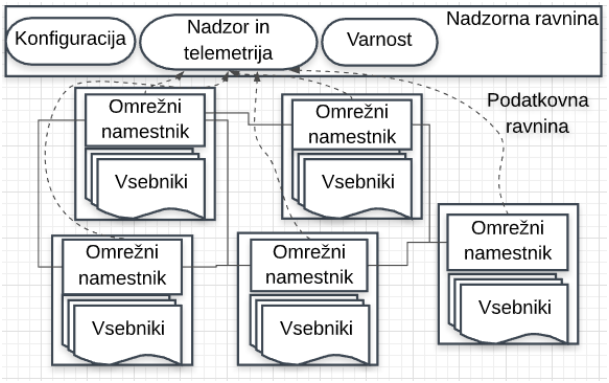
Odprtokodne rešitve za orkestracijo vsebnikov vključujejo Apache Mesos, CoreOS, OpenShift, Docker Swarm in Kubernetes. Slednjo smo uporabili tudi mi.

2.4 Storitvene mreže

Pri kompleksnejših oblačnih aplikacijah orkestracija vsebnikov ni dovolj, saj je potrebno zagotoviti zanesljivo komunikacijo skozi kompleksno topologijo mikrostoritev [9]. To lahko zagotovimo na več načinov, in sicer s programskimi knjižnicami v okviru vsake mikrostoritve, vozliščnimi agenti ali z inteligentnimi omrežnimi namestniki (ang. »network proxy«), ki se izvajajo poleg vsake namestitvene enote [9]. Najbolj uveljavljene storitvene mreže so Airbnb Synapse, AWS App Mesh, Linkerd2 in Istio.

Na splošno so najboljša možnost za namestitev storitvene mreže inteligentni omrežni namestniki, saj so neodvisni od različnih programskih jezikov, ne zahtevajo spreminjanja programske kode v mikrostoritvah in ne vnašajo dodatne zakasnitve pri komunikaciji [9]. Arhitektura takšne storitvene mreže je prikazana na Sliki 3.

Storitvena mreža je sestavljena iz podatkovne in nadzorne ravnine. Omrežni namestniki na podatkovni ravnini skrbijo za odkrivanje storitev, usmerjanje, porazdeljevanje obremenitve, overjanje, pooblastitve in spremljanje delovanja vsebnikov. Nadzorna ravnina upravlja s storitveno mrežo, zbira telemetrijske metrike in konfigurira omrežne namestnike za usmerjanje prometa [9].



Slika 3. Arhitektura storitvene mreže z namestitvijo omrežnih namestnikov [9].

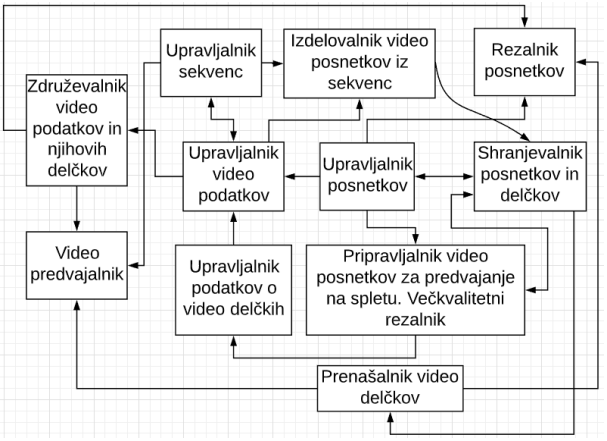
3 Oblačna aplikacija za urejanje in dostavo video posnetkov

Namen predstavljene aplikacije je učinkovito urejanje video posnetkov, pripravljenih za predvajanje na spletu z uporabo oblačnih arhitekturnih vzorcev, predstavljanih v prejšnjem poglavju. Aplikacija omogoča ogled knjižnice video posnetkov, njihovo nalaganje in razrez ter izbiro ali ustvarjanje projektov, v katerih se urejajo video posnetki na časovnicah ustvarjenih sekvenc. Na časovnico sekvence lahko poljubno dodajamo ali odstranjujemo posnetke. Po zaključku urejanja lahko video objavimo, tako da postane dostopen izven izbranega projekta ali pa ga prenesemo v lokalno shrambo. Vse opravljeno delo se sproti shranjuje, tako da lahko uporabnik z urejanjem nadaljuje kasneje.

3.1 Gradniki aplikacije

Izdelana aplikacija je sestavljena iz mikrostoritev. Te izvajajo nalaganje, rezanje in shranjevanje video posnetkov, prepoznavanje njihove vsebine, shranjevanje in pridobivanje metapodatkov o posnetkih in njihovih delčkih, pripravo datotek v zapisu m3u8 po protokolu HLS iz metapodatkov posnetkov, povezovanje video delčkov s pripadajočimi posnetki, izdelavo novih posnetkov iz obstoječih delčkov, pretvorbo posnetkov v različne kvalitete predvajanja ter objavo in predvajanje njihovih delčkov. Na Sliki 4 so prikazane ključne mikrostoritve in povezave med njimi.

Za komunikacijo med mikrostoritvami uporabljamo protokol gRPC (Google Remote Procedure Call), podatki pa so zapisani v formatu Protocol Buffers. gRPC za prenos podatkov uporablja protokol HTTP/2 [10]. S tem je gRPC učinkovitejši od klasičnega vzorca REST API (Representational State Transfer Application Programming Interface), ki ga sicer uporabljamo za uporabniške zahtevke, kot sta predvajanje video posnetka in prenašanje njegovih delčkov.



Slika 4. Mikrostoritvena arhitektura urejevalnika video posnetkov.

Mikrostoritve ki izvajajo izrazito asinhrone operacije, kot je priprava videov za predvajanje na spletu, vsebujejo čakalno vrsto za sporočila. V ta namen smo uporabili RabbitMQ, ki uporablja protokol AMQP (Advanced Message Queuing Protocol).

Pri analizi vsebine video posnetkov smo uporabili storitev Google Video AI. Zaradi velikega števila mikrostoritev in povezav med njimi smo za združevanje rezultatov poizvedb po različnih mikrostoritvah uporabili jezik GraphQL (Graph Query Language). S tem jezikom lahko zahtevamo že filtrirane podatke, s čimer optimiziramo njihovo količino pri prenosu med mikrostoritvami in uporabniškimi vmesniki.

3.2 Delovanje aplikacije

Pred začetkom urejanja v aplikacijo naložimo video posnetke. Za to je zadolžena storitev »Upravljalnik posnetkov«, ki s klicem storitve »Upravljalnik video podatkov« ustvari podatke o posnetku. S klicem storitve »Shranjevalnik posnetkov in delčkov« posnetek naloži v shrambo in pošlje zahtevek v sporočilno vrsto za njegovo pripravo na predvajanje na spletu. Posnetek se z orodjem FFmpeg razreže na pet-sekundne MPEG2-TS (MPEG2 Transport Stream) delčke različnih kvalitet.

Iz pripravljenih video posnetkov lahko izrežemo določen časovni izsek tako, da storitvi »Rezalnik posnetkov« sporočimo začetni in končni čas izseka. »Rezalnik posnetkov« iz izseka ustvari nov posnetek, tako da po časovni oznaki logično poveže obstoječe MPEG2-TS video delčke. Začetni in kočni video delček izbranega časovnega izseka videa se po potrebi razreže z orodjem FFmpeg in shrani s storitvijo »Shranjevalnik posnetkov in delčkov«. Ker je z rezalnikom video posnetkov potrebno razrezati samo začetni in končni pet-sekundni delček izbranega izseka posnetka, prihranimo na času obdelave. Izrezani video je nato brez dodatne obdelave pripravljen za predvajanje.

Za ustvarjanje novih posnetkov je najprej potrebno narediti sekvenco. Storitev »Upravljalnik sekvenc« pri dodajanju ali brisanju posnetkov iz sekvence pridobi od

storitve »Upravljalnik video podatkov« indeks posnetka in ga poveže z novo sekvenco.

Med urejanjem sekvence je mogoče pogledati nastajajoči posnetek s klicem storitve »Video predvajalnik« in podanim indeksnim parametrom sekvence. Storitev »Video predvajalnik« nato iz storitve »Upravljalnik sekvenc« pridobi indekse posnetkov v sekvenci. Nato s klicem storitve »Združevalnik video podatkov in njihovih delčkov« pridobi njihove metapodatke, ki se nahajajo v storitvah »Upravljalnik video podatkov« in »Upravljalnik podatkov o video delčkih«. Iz pridobljenih podatkov »Video predvajalnik« ustvari opisno datoteko m3u8 po protokolu HLS in jo posreduje odjemalcu. Ta nato video posnetek predvaja tako, da ga prenaša po delčkih iz storitve »Prenašalnik video delčkov«.

Sekvenco lahko po zaključku urejanja objavimo s pomočjo »Upravljalnika sekvenc«, ki zahtevo posreduje storitvi »Izdelovalnik video posnetkov iz sekvenc«. Objava posnetka je časovno učinkovita, saj storitev »Izdelovalnik video posnetkov iz sekvenc« pridobi metapodatke o sekvenci in vanjo vključenih posnetkih, jih indeksira v nov posnetek in to sporoči storitvi »Upravljalnik video podatkov«. Ustvarjeni video posnetek je za ogled na voljo s klicem storitve »Video predvajalnik« s podanim indeksnim parametrom videa.

3.3 Izvajanje mikrostoritev na oblačnih platformah

Ker mikrostoritve izvajamo na oblačnih platformah Microsoft Azure, Google Cloud in Amazon Web Service, smo uporabili rešitve za oblačne aplikacije.

Mikrostoritve smo zapakirali v vsebnike z lahko virtualizacijo Docker. Vsebnike upravlja orkestracijska platforma Kubernetes. Za zanesljivo komunikacijo, odkrivanje storitev, usmerjanje, porazdeljevanje obremenitve in opazovanje delovanja skrbi storitvena mreža Istio. Njeno stanje in konfiguracijo lahko spremljamo z orodjem Kiali. Metrike in dnevniške zapise delovanja vsebnikov zbira Prometheus. Vizualiziramo jih s storitvijo Grafana. Zahtevke spremljamo z orodjem Jaeger. Prometheus, Grafana in Jaeger so storitve na nadzorni ravnini storitvene mreže Istio. Postopek gradnje in namestitve smo avtomatizirali po principu DevOps (Development and Operations) z CI/CD (Continuous Integration and Continuous Delivery) [11] orodjem Travis CI. Ročna gradnja in namestitev mikrostoritev bi nam vzela preveč časa.

Za prikaz delovanja smo razvili spletno aplikacijo, ki se izvaja na brez-strežniški (ang. »serverless«) storitvi Google Cloud Run. Ker smo pred uporabnikom želeli skriti kompleksnost sistema, smo z enotnim vmesnikom aplikacijskega prehoda Istio Ingress izpostavili zahtevke, ki jih potrebuje spletna aplikacija.

4 Zaključek

Računalništvo v oblaku je postalo de facto standard za razvoj predvsem spletnih aplikacij, o čemer pričata tudi količina različnih odprtokodnih rešitev na portalu Cloud Native Computing Foundation in statistika uporabe oblačnih tehnologij [12].

Montaža in dostava video posnetkov z oblačno arhitekturo mikrostoritev pohitri proces objave za splet pripravljenih posnetkov na različnih spletiščih. Izdelano aplikacijo je mogoče enostavno vzdrževati in integrirati v obstoječe sisteme za dostavo vsebin. Aplikacijo bomo v bližnji prihodnosti prilagodili in uporabili za hitrejšo dostavo vsebin na različna spletišča podjetja Pro Plus, ki se ukvarja s produkcijo video vsebin.

Literatura

[1] Stockhammer, Thomas. “Dynamic adaptive streaming over HTTP --: standards and design principles.” MMSys (2011).

[2] Mueller, Christopher & Lederer, Stefan & Timmerer, Christian. (2012). An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments. 37-42. 10.1145/2151677.2151686.

[3] Ryan Jenee, Simon Rather, Julian Fruman, Wev-based system for video editing, United States Patent, 2. 9. 2014, https://patentimages.storage.googleapis.com/19/dd/75/1121a71203f3c8/US8826117.pdf

[4] Ian Lovejoy, Ken Wang, Kevin Wong, Eric Vossbrinck, Mark Moore, Henry Dall, Web based video editing, United States Patent, 15. 5. 2015, https://patentimages.storage.googleapis.com/af/dd/d6/c052d357586eae/US9032297.pdf

[5] Nane Kratzke, Peter-Christian Quint, Understanding cloud-native applications after 10 years of cloud computing - A systematic mapping study, Journal of Systems and Software, Volume 126, 2017, Pages 1-16

[6] Bob Familiar. 2015. Microservices, IoT, and Azure: Leveraging DevOps and Microservice Architecture to deliver SaaS Solutions (1st. ed.). Apress, USA.

[7] C. Pahl and B. Lee, "Containers and Clusters for Edge Cloud Architectures -- A Technology Review," 2015 3rd International Conference on Future Internet of Things and Cloud, Rome, 2015, pp. 379-386, doi: 10.1109/FiCloud.2015.35.

[8] Alex Williams, The State of the Kubernetes Ecosystem - The New Stack, https://lp.google-mkto.com/rs/248-TPC-286/images/TheNewStack_Book1_TheStateOfTheKubernetesEcosystem.pdf

[9] W. Li, Y. Lemieux, J. Gao, Z. Zhao and Y. Han, "Service Mesh: Challenges, State of the Art, and Future Research Opportunities," 2019 IEEE International Conference on Service-Oriented System Engineering (SOSE), San Francisco East Bay, CA, USA, 2019, pp. 122-1225, doi: 10.1109/SOSE.2019.00026.

[10] Vijay Pai, gRPC Design and Implementation, https://platformlab.stanford.edu/Seminar%20Talks/gRPC.pdf

[11] Martin Fowler, Continuous Integration, https://moodle2019-20.ua.es/moodle/pluginfile.php/2228/mod_resource/content/2/martin-fowler-continuous-integration.pdf

[12] CNCF Survey 2019, https://www.cncf.io/wp-content/uploads/2020/03/CNCF_Survey_Report.pdf

Primerjava HEVC in VP9

David Valič[1], Marko Meža[2], Urban Burnik[2]

[1] Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Fakulteta za elektrotehniko
[2]Univerza v Ljubljani, Fakulteta za elektrotehniko,

E-pošta: dv6510@student.uni-lj.si

Abstract: *The paper reviews and compares currently well-established video codecs HEVC and VP9 at several levels. The theoretical background and presentation of standard-specific compression stamps and procedures is followed by performance evaluation using characteristic video recordings with comments and conclusions.*

1 Uvod

Danes se video vsebine širijo bolj kot še nikoli prej, predvsem zaradi porasta uporabe socialnih omrežij in orodij za takojšnjo komunikacijo. Da bi omogočili čim učinkovitejši pretok vsebin, je potreben tudi napredek v kompresiji videa. Danes sta uveljavljena kodeka HEVC (*angl. High Efficency Video Codec*) in VP9. Cilj prispevka je primerjava obeh kodekov na različnih ravneh. Predstavljena je uporaba navedenih kodekov, sledi primerjava delovanja ter opis skupnih lastnosti in razlik. Prispevek se zaključi s praktično primerjavo obeh kodekov.

2 Ozadje kodekov HEVC in VP9 ter njuna uporaba

Kodek HEVC oz. H.265 sta leta 2013 standardizirali organizaciji ITU-T Video Coding Experts Group in ISO/IEC JTC1 Moving Picture Experts Group v okviru delovne skupine Joint Collaborative Team on Video Coding. Gre za naslednika kodeka AVC (*angl. Advanced Video Codec*) in prinaša prihranke do 50 % in več na pasovni širini pri enakovredni kvaliteti, sicer pa so prihranki odvisni tudi od vsebine posnetka.

VP9 je razvilo podjetje Google in je bil prvotno objavljen leta 2013 kot prosto-dostopni format za kompresijo videa. Tudi sama implementacija kodeka je odprtokodna.

Ker je za uporabo kodeka HEVC potrebna licenca, je najbolj razširjen v namenskih multimedijskih napravah, kot so kamere, pametni telefoni, televizije in računalniki. Za zasebno in spletno rabo pa je zaradi dostopnosti pogosteje v uporabi VP9. Največji uporabnik kodeka VP9 je skoraj zagotovo YouTube [5]. Tudi podjetje Netflix že od leta 2016 uporablja VP9 kot primarni kodek, od februarja 2020 pa ga za predvajanje na mobilnih napravah z operacijskim sistemom Android že nadomešča z naslednikom AV1 [9]. Slednji se danes uporablja tudi na Youtube-u.

3 Delovanje kodekov HEVC in VP9

Oba standarda za kompresijo temeljita na blokovnih operacijah v frekvenčnem prostoru. V osnovi je postopek dokaj podoben z nekaterimi razlikami v detajlih.

3.1 Osnovni princip delovanja kodekov

Preden se izvede katerokoli operacija, se posamična slika razdeli na ti. makrobloke, znotraj katerih se izvaja nadaljnje operacije. Makroblok je razdeljen na še manjše bloke, v okviru katerih se opravi predikcija. Obstajata dve različni predikciji:

- **Intra-predikcija** je predikcija znotraj istega okvirja videa in izkorišča podobnost med podatki v istem okvirju.
- **Inter-predikcija** je predikcija, ki upošteva zaporedne okvirje in s tem kompenzira gibanje.

Po dokončanem postopku predikcije se residualna slika kodira s transformacijskimi postopki. Po preslikavi v frekvenčni prostor se izvede utežena kvantizacija koeficientov, pri čemer se z utežmi določi stopnjo kompresije. Kvantizaciji sledi entropijsko kodiranje, ki končno velikost še dodatno zmanjša. Izhod iz kvantizatorja se dekvantizira in izvede se inverzna transformacija, ki da nazaj sliko, ki se prišteje rezultatu predikcije. Na celotnem okvirju se izvede post-procesiranje, ki omili blokovne artefakte.

3.2 Kodiranje HEVC

Makroblok v HEVC-u je predstavljen kot enota kodnega drevesa (*angl. Coding Tree Unit*) (CTU). Velikost CTU je lahko 16×16, 32×32 ali 64×64 pikslov. Večji CTU pripomore k učinkovitejšemu deljenju makroblokov pri videih visoke resolucije (4K in več). Večji bloki prinašajo deljenje CTU na 4 manjše bloke imenovane kodne enote *(angl. Coding Unit*) (CU). Vsak CU pa vsebuje še predikcijsko enoto (*angl. Prediction Unit*) (PU) ter transformacijsko enoto (*angl. Transform Unit*) (TU). CU se lahko nadaljnje deli na TU enote do najmanjše dovoljene vrednosti. V PU enotah so shranjeni podatki intra- in inter-predikcije (npr. vektorji premika). Deljenje CU na PU se lahko izvede na različne načine.

Intra-predikcija v HEVC-u se vedno izvaja na kvadratnih matrikah (blokih). Predikcija uporablja podatke sosednjih pikslov v naslednjih smereh: levo, levo-gor, levo-dol, desno-gor, gor (smeri nad antidiagonalo); definiranih s PU deljenjem. HEVC ponuja 35 različnih načinov intra-predikcije: način št. 0, poimenovan tudi planarni (*angl. planar*), način št. 1 imenovan tudi DC (*angl. Direct Currect*) ter še preostalih

33 kotnih načinov (načini od št. 2 do št. 34). DC način je najbolj preprost, saj so v njem vsi piksli v enem PU nastavljeni na povprečno vrednost sosednjih pikslov, medtem ko je planarni način najbolj procesorsko zahteven, saj se uporablja dvodimenzionalna linearna interpolacija pikslov. DC na robnih pikslih izvaja dodatno linearno filtriranje. Kotni načini pa so linearne interpolacije v različnih smereh glede na kot zasuka.

Za transformacijo se uporablja hitri algoritem za izračun celoštevilskih približkov diskretne kosinusne in sinusne transformacije (DCT tipa II in DST tipa VI).

Pri entropijskem kodiranju se uporablja aritmetični kodirnik imenovan sintaktični kontekstualno-adaptivni binarni aritmetični kodirnik (*angl. syntax-based context-adaptive binary arithmetic coder*) (SBAC), ki omogoča paralelno kodiranje in dekodiranje in se prilagaja znotraj samega okvirju.

Post-procesiranje pri HEVC-u se izvede v dveh korakih (oba koraka sta opcijska). Prvi korak DBF (*angl. Deblocking Filter*) odpravlja blokovne artefakte. Drugi korak SAO (*angl. Sample Adaptive Offset*) kompenzira splošno izgubo na dekodirani CTU enoti [1].

Za inter-predikcijo se uporabljata dva referenčna seznama L0 in L1, vsak seznam lahko vsebuje 16 referenčnih okvirjev, od katerih je lahko največ 8 slik unikatnih. To pomeni, da se za predikcijo lahko uporabi isto sliko dvakrat, vendar z različnimi utežmi. Za predikcijo vektorja premika se gradi seznam kandidatov, za vsako predikcijo pa se zapiše indeks kandidata. Za predikcijo vektorjev premika se lahko uporabi dva načina: združitev ali napredna predikcija vektorjev premika. Za vsak PU se določi kateri način predikcije je uporabljen. Vektorji premika so definirani za področja 16×16 pikslov, kar pomeni, da je potrebno pri dekodiranju imeti v medpomnilniku matriko 16×16 [2].

3.3 Kodiranje VP9

Pri VP9 so makrobloki imenovani superbloki, ti so velikosti 64×64 pikslov. Deljenje blokov je zelo podobno kot pri HEVC-u. Superblok se lahko razdeli do blokov velikosti 4×4 piksle. Deljenje poteka ravno tako na predikcijske in transformacijske bloke. Za razliko od HEVC-a pa dovoljuje tudi pravokotne bloke za intra-predikcijo, saj se dva pravokotna bloka lahko smatra kot en kvadratni blok. Tako imata sosednja pravokotna bloka isti način predikcije. Za predikcijsko deljenje je mogoče uporabiti deljenja 2N×2N, N×N, 2N×N in N×2N. Predikcija na pod-blokih velikosti 4×4 je determinirana z blokom 8×8 v katerega spadajo. Predikcija na večjih pod-blokih je determinirana posamično za vsak pod-blok.

Za intra-predikcijo se uporabljajo sosednji piksli v naslednjih smereh: levo, levo-gor, desno-gor in gor. VP9 za razliko od HEVC-a ponuja le 10 načinov intra-predikcije: DC, True Motion in osem kotnih načinov (horizontalni, vertikalni, koti 207º, 153º, 117º, 63º in 45º. Način DC je enak načinu kodeka HEVC, le brez dodatnega filtriranja robnih pikslov.

Pri kvantizaciji se uporablja celoštevilski približek DCT tipa II in DST tipa II za vse velikosti blokov (32×32, 16×16 in 8×8), razen za bloke 4×4 se uporablja DST tipa VI. Poleg DCT in DST pa se uporablja tudi asimetrična diskretna sinusna transformacija (ADST).

Za entropijsko kodiranje se uporablja 8-bitno aritmetično kodiranje imenovano *"bool-coder"*. Za razliko od HEVC-a se verjetnost kodiranja ne prilagaja znotraj okvirja, ampak so spremembe zapisane v glavi okvirja kot logične vrednosti (kontekst okvirja).

Post-procesiranje je sestavljeno iz enega samega koraka za odpravljanje artefaktov, ki filtrira na mejah od 8 do 2 piksla na vsaki strani. Pri filtriranju sodeluje tudi detektor, ki zaznava enakomernost območja pikslov in s tem prilagaja velikost filtra.

Inter-predikcija pri VP9 uporablja kompenzacijo gibanja na 8 pikslih. V osnovi se uporabi en vektor premika za en blok, opcijsko pa se lahko dovoli tudi predikcijo z dvema vektorjema premika, ki ju potem povprečimo. Za vsak blok ali okvir se lahko uporabi enega od treh filtrov:

- navadni – Lagrangianov interpolacijski filter,
- ostri – interpolacijski filter, ki bazira na DCT in se uporablja okrog ostrih robov ter
- mehki – ne-interpolacijski filter [1].

Vektor premika kaže na enega izmed treh možnih referenčnih okvirjev, poimenovanih *"Last"*, *"Golden"* in *"AltRef"*. Vektorji premika so napovedani iz urejenega seznama kandidatov referenčnih vektorjev premika. Kandidati so sestavljen iz do 8 sosednjih blokov iz iste referenčne slike. Sledijo še časovni prediktorji iz prejšnjega okvirja. Če se pri iskanju kandidatov seznam ne napolni, se izvede ponovno iskanje brez ujemanja z referenčno sliko. Če seznam še vedno ni poln, se napolni z vektorji brez premika (0, 0). V seznamu kandidatov je vedno 8 referenčnih slik, od teh so tri slike uporabljene za intra-predikcijo. Seznam kandidatov se posodablja sproti [2].

Tabela 1: Seznam videoposnetkov uporabljenih v praktični primerjavi ter njihove lastnosti

Ime videa	Ločljivost	Št. sličic na sekundo	Število okvirjev	Bitna hitrost	Format in barvno podvzorčenje	Velikost datoteke
Park Joy	1280×720	50	500	1106Mb/s	YUV 4:4:4	1,28GB
FourPeople	1280×720	60	601	664Mb/s	YUV 4:2:0	792Mb
Life	1920×1080	30	825	747Mb/s	YUV 4:2:0	2,38GB
Shields	1280×720	50	504	553Mb/s	YUV 4:2:0	664MB

4 Primerjava kodiranja videov s HEVC in VP9 kodekom

Za primerjavo kodeka HEVC in VP9 je bil izveden praktični eksperiment kodiranja različnih videov. Za primerjavo so bili iz arhiva nekomprimiranih posnetkov https://media.xiph.org/video/derf/ izbrani videoposnetki, ki jih podaja Tabela 1.

Opisi posnetkov:

- Na posnetku ”Park Joy” gre za kamero ki se pomika v desno smer, tako da se ozadje premika z gibanjem kamere.
- Na posnetku ”FourPeople” gre za statično kamero, ki je usmerjena na mizo, za katero sedijo štiri osebe, ki si podajajo dokumente.
- Posnetek ”Life” je edini računalniško generiran posnetek – gre za animacijo več statičnih scen, v katerih se kamera pomika v polkrožni smeri.
- Posnetek ”Shields” je precej podoben posnetku ”Park Joy”, le da se kamera giba v levo smer.

Za izvedbo kodiranja je bilo uporabljeno orodje FFmpeg verzije 20200403-52523b6, ki lahko kodira s kodekom HEVC s pomočjo knjižnice libx265 [3] in s kodekom VP9 s knjižnico libvpx-vp9.

Posamezni posnetek je bil kodiran s ciljnimi bitnimi hitrostmi 2 Mb/s, 5 Mb/s, 10 Mb/s, 25 Mb/s, 75 Mb/s in 150 Mb/s. FFmpeg za kodiranje z VP9 z determinirano bitno hitrostjo priporoča dvopasovno kodiranje [4], zato je bilo to izbrano tudi za kodiranje s kodekom HEVC. Vsa kodiranja so bila izvedena na prenosnem računalniku Asus VivoBook S510U s procesorjem Intel® Core™ i7-8550U 1.8GHz z integrirano grafično kartico Intel® UHD Graphics 620, 8GB RAM in operacijskim sistemom Windows 10 Home 64-bit.

Za primerjavo rezultatov kodiranja se je za vsak video in ciljno bitno hitrost primerjal čas trajanja kodiranja, in ocenjevala se je kvaliteta videa z metrikami *”Peak Signal to Noise Ratio”* (PSNR) in *”Structural Similarity Index”* (SSIM). PSNR je razmerje med vrhovi signala in šuma. Izračunan je na podlagi srednje kvadratne napake (*angl. Mean Square Error*). Ker PSNR ne daje najboljših rezultatov glede na človeško percepcijo je dobro upoštevati še SSIM. SSIM je metrika, ki upošteva lokalno luminanco, kontrast in strukturo za določanje kvalitete [5]. Oba parametra sta podana v dB.

Ker se za primerjavo rezultatov uporabljajo PSNR in SSIM, je potrebno pri kodiranju upoštevati nastavitev posebej za PSNR in posebej za SSIM. Rezultati so bili izračunani z orodjem FFmpeg. Ker ta izračuna PSNR in SSIM vrednosti za vsak okvir, je bilo pri rezultatih vzeto povprečje vseh okvirjev.

5 Rezultati

5.1 Rezultati in primerjava procesa kodiranja

Tabela 2 predstavlja povprečni čas, potreben za kodiranje videoposnetka za posamezne ciljne bitne hitrosti. Opazno je, da je kodiranje s HEVC-om časovno bolj učinkovito kot z VP9. Razlika v učinkovitosti je v razponu od 2x pa tudi do 5x in več. Razlog za tako veliko razliko je tudi procesor, ki je bil uporabljen za kodiranje. Intelovi procesorji osme generacije so bolj optimizirani za kodiranje s kodekom HEVC in uporabljajo več virov pri kodiranju. Kodiranje s kodekom HEVC zna izkoristiti od 2x do 3x več procesorske moči v primerjavi z VP9. Izvedena je bila tudi primerjava zahtevnosti dekodiranja. Kodek HEVC potrebuje približno 10 % več procesorske moči kot VP9 za dekodiranje istega posnetka pri isti bitni hitrosti. Zaključi se lahko, da je HEVC učinkovitejši pri kodiranju, a manj učinkovit pri dekodiranju.

5.2 Rezultati in primerjava objektivno merjene kvalitete videoposnetkov

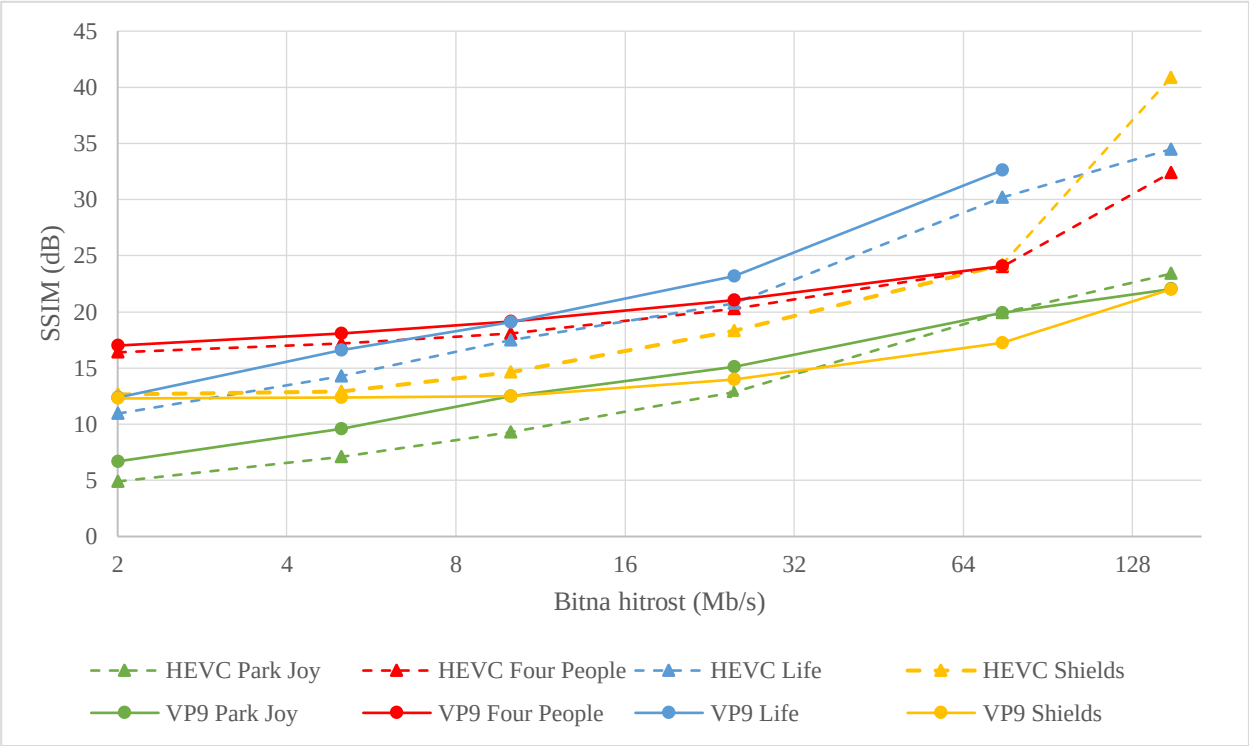
Sledi primerjava objektivno določene kvalitete videoposnetkov. Rezultati so bili pridobljeni tako za PSNR in SSIM, a ob podobnem poteku Slika 1 predstavlja le potek SSIM. Na splošno vsi rezultati kažejo, da sta si kodeka HEVC in VP9 pri isti bitni hitrosti dokaj primerljiva. Glede na posnetek ”Park Joy” velja, da daje SSIM pri nižjih bitnih hitrostih boljše rezultate za kodek VP9, medtem ko je pri višjih bitnih hitrostih malenkost boljši HEVC. Medtem ko je pri posnetkih ”Four People” in ”Life” konstanto boljši VP9. Pri posnetku ”Life” je to še bolj izrazito, zato se lahko sklepa, da je VP9 bolj optimiziran za računalniško generirane posnetke, kjer ni toliko šuma, kot pri resničnih posnetkih. Predvideva se lahko tudi, da se VP9 bolje obnese na posnetkih višje ločljivosti. Zanimiva je tudi razlika v rezultatih posnetka ”Shields”, kjer se HEVC izkaže za boljšega pri vseh bitnih hitrostih, iz tega primera lahko sklepamo, da je predikcija gibanja enega objekta pri HEVC-u precej učinkovitejša kot predikcija gibanja več objektov.

Če se dobljene rezultate primerja z ostalimi raziskavami [1, 5, 6, 7] in poljudnimi eksperimenti [8] je opaziti, da rezultati pridobljeni v letu 2013 [1] kažejo na to, da daje HEVC konstanto boljše rezultate – videoposnetki so pri isti bitni hitrosti višje kvalitete. Raziskavi iz leta 2016 [5, 6] sta že nekoliko drugačni od tiste iz 2013, tukaj so rezultati kodiranje s HEVC in VP9 zelo primerljivi. Novejši rezultati iz leta 2018 [8] in 2019 [7] pa so podobni rezultatom tega prispevka, kjer je VP9 boljši pri nizkih pasovnih širinah. Da se sklepati, da je VP9 v času od njegovega izida do danes prejel več izboljšav kot HEVC. Razlog je lahko tudi zaprtost kodeka HEVC, zaradi katerega je FFmpeg knjižnico za njegovo uporabo težje vzdrževati.

Tabela 2: Čas kodiranja s kodekoma HEVC in VP9 za različne ciljne bitne hitrosti – povprečje vseh videov

Ciljna bitna hitrost	**150 Mb/s**	**75 Mb/s**	**25 Mb/s**	**10 Mb/s**	**5 Mb/s**	**2 Mb/s**
Čas kodiranja HEVC	256 s	195 s	114 s	72 s	53 s	40 s
Čas kodiranja VP9	704 s	678 s	437 s	330 s	274 s	193s

Slika 1: Potek odvisnosti SSIM od bitne hitrosti za posnetke kodirane s kodekoma HEVC in VP9.



6 Zaključek

Glede na dobljene rezultate se zaključi, da pri isti bitni hitrosti kodek VP9 dosega boljše rezultate, vendar je pri uporabi potrebno upoštevati tudi zahtevnost kodiranja, saj traja kodiranje z VP9 precej dlje kot s HEVC-om. Odločitev za izbor je odvisna od uporabe, kje se bo video kodiral in kje predvajal.

Ker so bili v prispevku primerjani samo štirje različni videi, se težko sklepa, kateri kodek je boljši za katero vsebino ter kakšno ločljivost. Za tako primerjavo bi bilo potrebno primerjati večje število videoposnetkov, kjer bi več posnetkov imelo podobno vsebino. Rezultate posnetkov s podobno vsebino bi se združili v skupino in bi nato bile primerjane skupine med seboj.

Glede na to, da je od leta 2013, ko sta bila izdana kodeka minilo že kar nekaj let, bi bilo koristno v primerjavo vključiti tudi naslednika VP9, to je AV1. V prihodu pa je tudi VVC, naslednik HEVC-a. Oba novejša kodeka obljubljata še večje pridobitve pri videih visoke ločljivosti 4K in 8K.

Literatura

[1] M. P. Sharabayko, O. G. Ponomarev, R. I. Chernyak: Intra Compression Efficiency in VP9 and HEVC; Applied Mathematical Sciences, Vol. 7, 2013, no. 137, 6803 – _6824, HIKARI 2013.

[2] S. Akramullah: Digital Video Concepts, Methods and Metrics, Apres open 2014.

[3] H.265/HEVC Video Encoding Guide, FFmpeg, dostopno na https://trac.ffmpeg.org/wiki/Encode/H.265 [Poskus dostopa 18.4.2020].

[4] FFmpeg and VP9 encoding guide, FFmpeg, dostopno na https://trac.ffmpeg.org/wiki/Encode/VP9 [Poskus dostopa 18.4.2020]-

[5] J. D. Cock, A. Mavlankar, A. Moorthy, A. Aaron: A Large-Scale Video Codec Comparison of x264, x265 and libvpx for Practical VOD Applications, 2016, Netflix - Video Algorithms

[6] J. Kufa, L. Polak, T. Kratochvil: HEVC/H.265 vs. VP9 for Full HD and UHD Video: Is There Any Difference in QoE?; 58th International Symposium ELMAR-2016, Zadar Croatia, IEEE 2016.

[7] Z. Li, Z. Duanmu, W. Liu, Z. Wang: AVC, HEVC, VP9, AVS2 or AV1? — _A Comparative Study of State-of-the-Art Video Encoders on 4K Videos; Springer Nature Switzerland AG 2019, Springer 2019.

[8] AV1 vs VP9 vs AVC (h.264) vs HEVC (h.265) - 4 part comparison, Texpion, dostopno na https://www.texpion.com/2018/07/av1-vs-vp9-vs-avc-h264-vs-hevc-h265-1-lossless.html [Poskus dostopa 19.4.2020]

[9] Netflix Now Streaming AV1 on Android, Netflix Technology Blog, Medium 2020, dostopno na https://netflixtechblog.com/netflix-now-streaming-av1-on-android-d5264a515202 [Poskus dostopa 19.4.2020]

Sensitivity analysis for face detection

Romi Koželj, Žiga Emeršič, Luka Šajn

University of Ljubljana, Faculty of Computer and Information Science
E-mail: romi.kozelj@gmail.com, ziga.emersic@fri.uni-lj.si, luka.sajn@fri.uni-lj.si

***Abstract*—Face detection has become one of the most active research topics in computer vision for its interesting applications, such as face recognition, face tracking, facial expression analysis and video surveillance. A large number of the face related applications assume that the face region is perfectly localized or detected. In order to adopt or use a certain face detection algorithm, evaluation of its performance is needed. In this paper, we evaluate three face detection algorithms with six different methods, using three evaluation metrices and discussed common failure cases that are worth to be further investigated.**

I. Introduction

Face detection is a critical step to all facial analysis algorithms, including face alignment, face recognition, face verification, and face parsing. Given an arbitrary image, the goal of face detection is to determine whether or not there are faces in the image and obtain the location and the size of each face [1]. The challenges associated with face detection can be, in addition to variations in pose and facial expression, attributed also due to scale, occlusion, brightness conditions, noise, etc. In this paper we evaluate the accuracy of the three chosen face detectors, whose parameters are changed in different manners to show, how different conditions of images influence detection accuracy.

We compared performance of three face detectors, namely `opencv's` [2] *detectMultiScale* method, which uses Viola and Jones detection algorithm [3], `insightface` detector [4], which is minimalistic inference-focused repack of [5] and `mxnet-face` [6] detector. First detector was taken from Python *OpenCV* library and the code for the last two detectors was obtained from the corresponding authors.

For `opencv's` *detectMultiScale* method parameters *scale factor* and *min neighbors* need to be specified. Parameter *scale factor* is specifying how much the image size is reduced at each image scale. By rescaling the image, larger faces can be resized to a smaller ones, making them detectable by the algorithm. Small step for resizing means that the algorithm works slower since it is more thorough. Bigger step is used for faster detection, with the risk of missing some faces altogether. We set parameter value to 1.05, which means that we used step size 5% for rescaling the image. Parameter *min neighbors* is specifying how many neighbors each candidate rectangle should have to retain it. This parameter affects the quality of the detected faces. Higher value results in fewer detections but with higher quality. We set parameter to value 3. `Insightface` detector performs pixel-wise face localisation on different scales of faces by taking advantages of joint extra-supervised and self-supervised multi-task learning [7]. `mxnet-face` detector uses fast R-CNN network [8], [9].

We ran face detectors on images, whose quality we were changing with 6 different parameters. Then we compared predictions using 3 evaluation metrices and reported the results.

The remaining sections of our paper are organized as follows. We first do a short review of related work. Then, in the methods section, we explain how used methods and metrices work. In the results section, we show and compare the results of this methods. At last, we summarise and comment the results in the conclusion section.

II. Related work

Face detection has been studied for decades in the computer vision literature. Face detection algorithms can be categorized into four categories: cascade based methods [10], [11], part based methods [12], channel feature based methods [13], [14], and neural network based methods [15], [16]. The seminal work by Viola and Jones [11] introduces integral image to compute Haar-like features in constant time. These features are then used to teach AdaBoost classifier with cascade structure for face detection. Various later studies follow a similar pipeline. Among those variants, SURF cascade [10] achieves competitive performance. One of the well-known part based methods is deformable part models (DPM) [12]. Deformable part models define face as a collection of parts and model the connections of parts through Latent Support Vector Machine. The part based methods are more robust to occlusion compared with cascade-based methods. Aggregated channel feature (ACF) is first proposed by Dollar et al. [13] to solve pedestrian detection. Later on, Yang et al. [14] applied this idea on face detection. In particular, features such as gradient histogram, integral histogram, and color channels are combined and used to teach boosting classifier with cascade structure. Studies [15] and [16] show that face detection can be improved by using deep learning, leveraging the high capacity of deep convolutional networks.

III. Methodology

To measure performance of detectors, we were gradually changing quality of images in dataset and at each step evaluated detectors predictions. We manipulated images with changes of brightness, contrast, resolution, blurriness, noise and occlusion, separately. We were changing images in each of these manners by discrete steps and at each step ran detectors over all images in dataset and calculated three metrices for evaluation, namely *intersection over union*, *precision* and *recall*. For each metric we calculated its value for each image in dataset and averaged values over whole dataset.

With average *intersection over union* we calculated average proportion of area of correctly predicted face bounding boxes against area of both predicted and ground-truth bounding boxes on images. Metric is defined as

$$\text{intersection over union} = \frac{\sum_{i=1}^{m} \left(\frac{A^i_{overlap}}{A^i_{union}} \right)}{m},$$

where m denotes number of images, $A^i_{overlap}$ denotes overlapping area of ground-truth and predicted face bounding boxes for image i and A^i_{union} denotes area of union of ground-truth and predicted face bounding boxes for image i. With average *precision* we calculated average proportion of correctly predicted bounding boxes against all predicted bounding boxes and with average *recall* we calculated average proportion of correctly predicted bounding boxes against all gound-truth bounding boxes. We calculated average *precision* as

$$\text{precision} = \frac{\sum_{i=1}^{m} \left(\frac{tp^i}{tp^i + fp^i} \right)}{m}$$

and averege *recall* as

$$\text{recall} = \frac{\sum_{i=1}^{m} \left(\frac{tp^i}{tp^i + fn^i} \right)}{m},$$

where m again denotes number of images, tp^i denotes area of true positive detections, which is in this case the same as $A^i_{overlap}$. fp^i denotes area of false positive detections for image i – area on image i detected as face, but is not a face and fn^i denotes area of false negative detections for image i – area on image i where faces are present, but are not detected.

To calculate above three metrices, we represented each image as ground-truth matrix, representing image with ground-truth face bounding boxes and as prediction matrix, representing image with predicted face bounding boxes. Matrices have the same size as original image and are composed only of values 0 and 1. Ground-truth matrix has value 0 on indexes where faces are not present and value 1 on indexes where faces are present. Prediction matrix has value 0 on indexes not predicted as face and value 1 on indexes predicted as face. To get value for $A^i_{overlap}$ (and tp^i) we summed up these two matrices and counted the number of indexes of summed up matrix, where values where bigger than 1. Similarly, to get the value of A^i_{union} we counted the number of indexes of summed up matrix, where values where bigger than 0. To get value for $tp^i + fp^i$, which represents area on image i where faces are predicted, we counted the number of indexes of prediction matrix, where values were bigger than 0. To get value for $tp^i + fn^i$, which represents area on image i were faces are present, we counted the number of indexes of ground-truth matrix, where values were bigger than 0.

When testing the effect of brightness on detection quality, we were gradually changing brightness of images from 0% to 100% by 5%. In the same way we tested effect of contrast. When testing the effect of image sizes, we first reduced size of each image to 1% of its original size and then enlarged it by 5%, until we reached 101% of the original size. To incorporate effect of blurriness, we replaced each pixel value in the image by average value of the neighbouring pixels, called kernel. We changed kernel size by 1% from 0% to 13% of the original image, where we stopped, since accuracy of the predictions for all detectors fell to 0%. Similarly we incorporate effect of noise on images, where weighted randomize matrix of the same size as the image, with values from 0 to 255, was added to each image. We changed noise percentage (weight) by 4% intervals, from 0% to 80%, where accuracy of the predictions fell to 0%. At last, we added occlusions to images. Here we were randomly masking images with black boxes, whose sizes were gradually increased by 5% from 0% to 100% of the original image. The number of boxes for each image was determined by the number of ground-truth faces on the image.

IV. Results

We found dataset with ground-truth face bounding boxes (i.e., the hand labeled bounding boxes for images that specify the faces locations) on `Dataturk` website [17]. Dataset includes 409 labeled images. Since the dataset contains a smaller amount of images, we split the images in 5 folds and calculate mentioned 3 metrices for each fold separately and average the results. To show how spread out the results are, we also calculate standard deviation over 5 folds.

Figures 1– 6 show performance of chosen three detectors evaluated with described metrices. In terms of resolution and occlusion nice hierarchy of these detectors can be seen along the whole scale of changes for all three metrices. `mxnet-face` detector performs the best, next is `insightface` detector and the worst is `opencv` detector. The same hierarchy can be observed from other quality changes, but holds only to some extent. When quality of images is changed to a grater extreme, `opencv` detector out-performs one or both other detectors. This is probably because `insightface` and `mxnet-face` detector were both trained on WIDER FACE dataset, which contains images with occlusions as stated in [18]. Also different sizes of images are probably included in this dataset. For this reason detectors trained on this dataset perform well on images with different resolutions and added occlusions. Other types of quality changes like extreme lightning conditions, high contrast, blurriness or noise, are probably not included in images in the WIDER FACE dataset, so this two detectors perform poorly when used on images whose quality is reduced in such way.

`mxnet-face` and `insightface` detectors remain in the same hierarchy order for all types of quality changes, except in terms of blurriness, where for images blurred for more than aprox. 4.5% (for all three metrices), `mxnet-face` detector shows the worst results.

From all types of quality changes, `opencv` detector shows better results for average *recall* metric in comparison to the other two metrices, since line jumps up to the other two lines. This can be due to the lower value for *min neighbors* parameter, which we specified to value 3. If parameter would be higher, detections would have higher quality and less incorrect predictions of the faces would be made. For that reason the area of images where faces are predicted would be lower, which would improve values for average *intersection of union* and average *precision*. Although as observed from figures showing the results for average *recall*, described hierarchy would still hold for all types of quality changes, except maybe for contrast and noise, where `opencv` detector out-performs `insightface` detector on almost whole scale.

Figures show low values for metrices even in cases when detectors perform best. The reason for such low values is that hand labeled bounding boxes are much wider and higher than detectors predict, even though, in this cases, detectors detect faces perfectly.

V. Conclusion

We tested performance of three detectors on images whose quality was changed in terms of brightness, contrast, resolution, blurriness, noise and occlusion. In terms of resolution and occlusion nice hierarchy of these detectors is observed, `mxnet-face` detector performing the best and `opencv` detector performing the worst. In terms of other types of quality changes this hierarchy holds only to some point, where then `opencv` detector out-performs one or both other detectors. `mxnet-face` detector, which overall shows the best results, performs poorly on images blurred for more than aprox. 4.5%.

For detectors that need training on images, quality of these images need to be changed in various ways, to improve detectors performance. Since `opencv` detector shows better results for average *recall* metric than for the other two metrices, future work should be done on testing `opencv` detector performance with different parameter values and evaluating results. To increase values of metrices for all three detectors and show more realistic results, future work should be done on excluding outer margins for hand labeled bounding boxes, as they are to big, or try to evaluate detectors on some other dataset as well.

References

[1] Ming-Hsuan Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, Jan 2002.

[2] "https://docs.opencv.org/2.4/modules/objdetect/doc/cascade_classification.html."

[3] M. Hassaballah, K. Murakami, and S. Ido, "Face detection evaluation: A new approach based on the golden ratio," *Signal Image and Video Processing*, vol. 7, 03 2013.

[4] "https://github.com/kiselev1189/insightface-just-works."

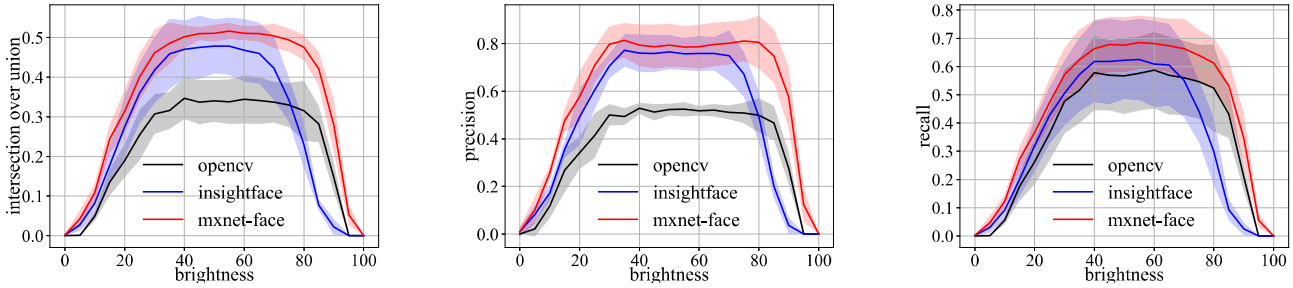


Figure 1: Plots of average intersection over union (left), average precision (middle) and average recall (right) for 5 folds of overall 409 images along the brightness of these images for `opencv`, `insightface` and `mxnet-face` detector. Lines show mean values and shaded areas show standard deviations over folds for each detector. Values on x axis represent brightness percentage of the images.

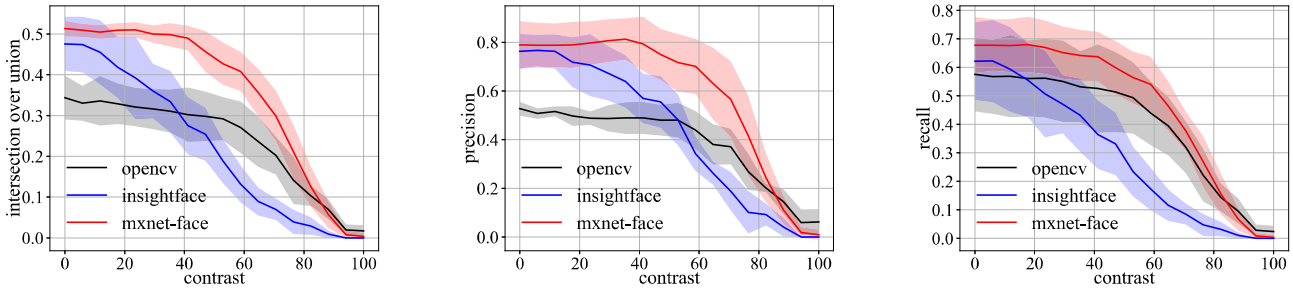


Figure 2: Plots of average intersection over union (left), average precision (middle) and average recall (right) for 5 folds of overall 409 images along the contrast of these images for `opencv`, `insightface` and `mxnet-face` detector. Lines show mean values and shaded areas show standard deviations over folds for each detector. Values on x axis represent percentage of contrast enforced to the images.

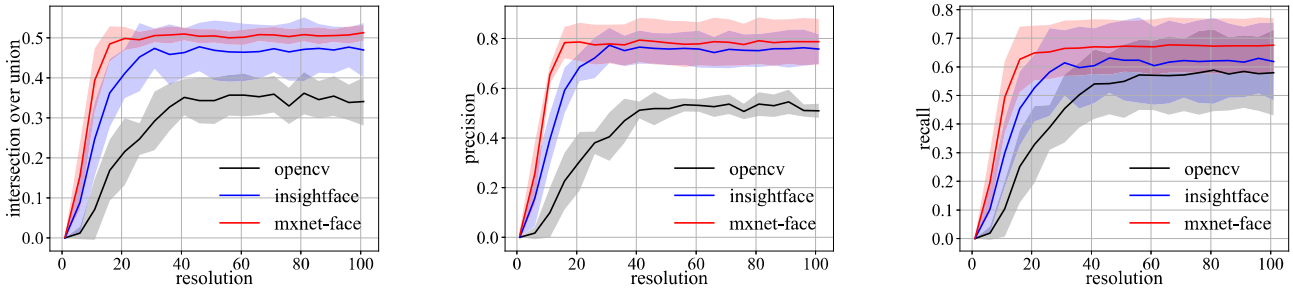


Figure 3: Plots of average intersection over union (left), average precision (middle) and average recall (right) for 5 folds of overall 409 images along the resolution of these images for `opencv`, `insightface` and `mxnet-face` detector. Lines show mean values and shaded areas show standard deviations over folds for each detector. Values on x axis represent size of images in percentage of original size of the images.

[5] "https://github.com/deepinsight/insightface."

[6] "https://github.com/tornadomeet/mxnet-face."

[7] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *ArXiv*, vol. abs/1905.00641, 2019.

[8] R. Girshick, "Fast r-cnn," 04 2015.

[9] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," pp. 650–657, May 2017.

[10] J. Li and Y. Zhang, "Learning surf cascade for fast and accurate object detection," pp. 3468–3475, June 2013.

[11] P. Viola and M. Jones, "Robust real-time object detection," 2001.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[13] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," pp. 91.1–91.11, 2009.

[14] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," *IEEE International Joint*

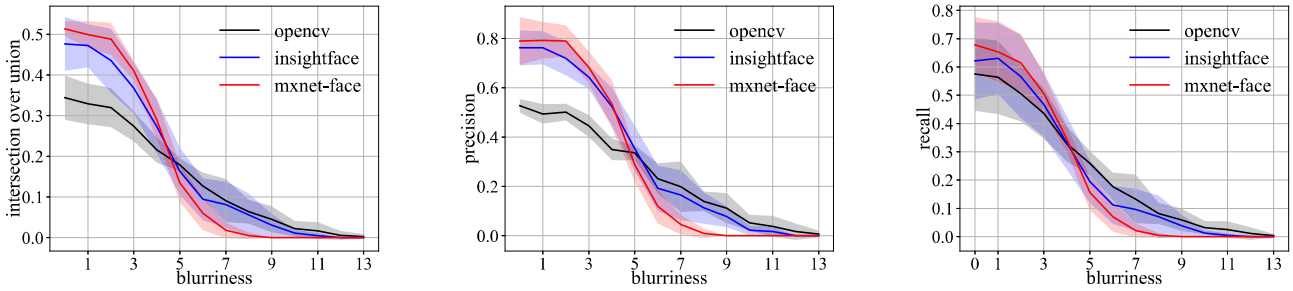


Figure 4: Plots of average intersection over union (left), average precision (middle) and average recall (right) for 5 folds of overall 409 images along the blurriness of these images for `opencv`, `insightface` and `mxnet-face` detector. Lines show mean values and shaded areas show standard deviations over folds for each detector. Values on x axis represent kernel size in percentage of original size of the images.

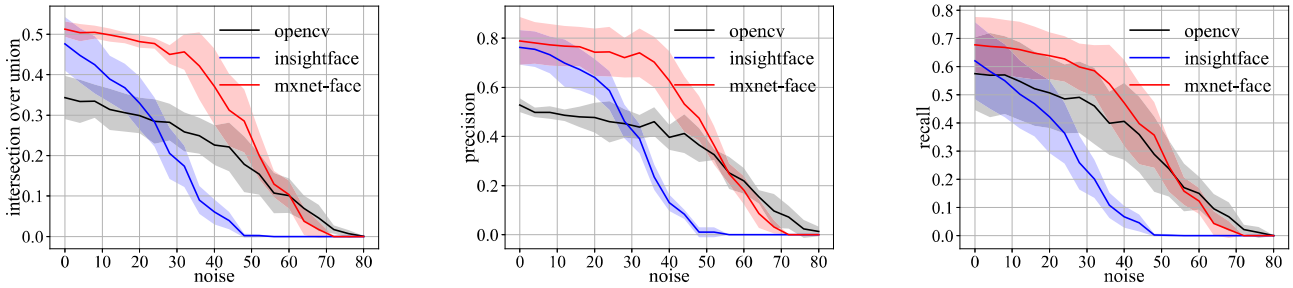


Figure 5: Plots of average intersection over union (left), average precision (middle) and average recall (right) for 5 folds of overall 409 images along the noise of these images for `opencv`, `insightface` and `mxnet-face` detector. Lines show mean values and shaded areas show standard deviations over folds for each detector. Values on x axis represent percentage of noise added to the images.

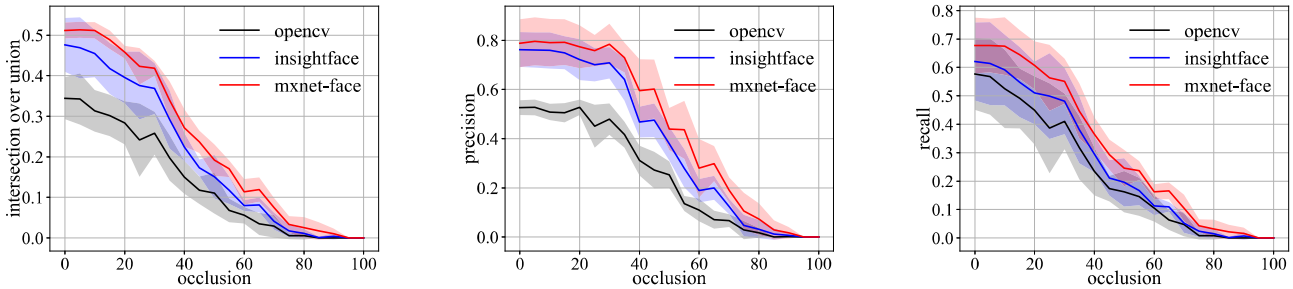


Figure 6: Plots of average intersection over union (left), average precision (middle) and average recall (right) for 5 folds of overall 409 images along the occlusion of these images for `opencv`, `insightface` and `mxnet-face` detector. Lines show mean values and shaded areas show standard deviations over folds for each detector. Values on x axis represent black box sizes inforced to the images in percentage of original size of the images.

Conference on Biometrics, pp. 1–8, 2014.

[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," pp. 5325–5334, June 2015.

[16] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," pp. 3676–3684, 12 2015.

[17] "https://dataturks.com/projects/devika.mishra/face_detection."

[18] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," pp. 5525–5533, June 2016.

Razvoj filtra za obogatenje resničnosti

Miha Malenšek, Blaž Pridgar, izr. prof. dr. Narvika Bovcon

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
E-pošta:
mmalen98@gmail.com
bp4769@student.uni-lj.si
narvika.bovcon@fri.uni-lj.si

Development of an Augmented Reality (AR) Experience

Abstract. *This article describes the development of an Augmented Reality filter, or "lens", designed for use on the social network app Snapchat. We started the project at the beginning of the COVID-19 pandemic, and since we did not find any relevant filters online that would warn of the dangers of spreading the virus, we opted for a two-part thematic filter that would remind users in a fun way to use a protective mask and keep the safety distance in public. On the front phone camera with the help of face tracking algorithms the lens projects a 3D protective mask over the user's nose and mouth, which can also be removed and then raised again by raising the eyebrows. On the rear phone camera our mascot named "Safey" appears in front of the user and keeps the safety distance from the user, even if one approaches him. We created all the 3D models in Blender and developed the lens in Lens Studio, the dedicated platform for creating lenses for Snapchat.*

1 Uvod

Članek opisuje razvoj filtra za obogatenje resničnosti, ki je namenjen uporabi na družbenem omrežju Snapchat. V članku predstavimo izhodišča projekta, razvoj ideje, uporabljene tehnologije, implementacijo filtra in načrtovanje oblikovalske rešitve.

Projekt je bil zasnovan ob začetku pandemije novega korona virusa: ker na spletu ob začetku projekta nismo zasledili nobenih aktualnih filtrov, ki bi opozarjali na nevarnosti širjenja virusa, smo se odločili, da izdelamo tematski filter, sestavljen iz dveh delov, ki bo uporabnike na zabaven način opomnil na upoštevanje varnostnih ukrepov. Prvi del filtra opozarja na uporabo zaščitne maske, drugi pa na upoštevanje varnostne razdalje v javnosti. Dvodelni filtri so tudi sicer zelo popularni, saj se lahko en del uporablja na sprednji telefonski kameri, drugi pa na zadnji [1]. Izhajali smo iz ideje o AR izkušnji, ki bi simulirala nameščanje varnostne maske na obraz, kar se odlično poda uporabi sprednje (selfie) telefonske kamere, želeli pa smo ustvariti tudi zabavno maskoto, ki bi filter naredila še bolj všečen. Ob razvijanju filtra smo se oprli predvsem na naše dosedanje izkušnje [2] ter na primere filtrov, ki ljudi osveščajo o boleznih, npr. o virusu HIV [3].

Končni cilj projekta je bil javna objava filtra in s tem osveščanje ljudi glede varnosti med pandemijo ter seveda pridobivanje izkušenj za prihodnje projekte.

2 Razvoj AR

Obogatena resničnost obstaja že kar nekaj let. Enega izmed prvih uporabnih AR sistemov je razvil Louis Rosenberg leta 1992 za uporabo pri Vojnem letalstvu Združenih držav Amerike (USAF). Komercialna uporaba tehnologije je sledila nekaj let kasneje, s prvo AR predstavo *Dancing in Cyberspace* Julie Martin leta 1994. Z razvojem interneta in z internetom povezanih tehnologij se je dalje razvijala tudi tehnologija obogatene resničnosti. Leta 2009 se je za uporabo na spletnih brskalnikih pojavilo orodje ARToolkit, ki je izključno namenjeno razvijanju AR izkušenj. [4] Septembra 2015 je Snapchat kot prvo družabno omrežje z obogateno resničnostjo razvil AR izkušnjo v realnem času z uporabo telefonske kamere. Razvoj tehnologije se je tako samo pospešil, kmalu sta se trgu pridružila tudi Facebook in Instagram. Decembra 2017 je Snapchat razvil prosto dostopen program Lens Studio, ki je namenjen izdelavi Snapchat AR filtrov (AR filter je v angleškem izvirniku poimenovan "lens", torej leča). Temu je leto kasneje sledil Facebook s programom Spark AR. [5] Področju pa se je pridružil tudi Google s splošnim orodjem WebAR za razvoj AR aplikacij. V okviru svojega širokega raziskovalnega projekta, ki snuje nove modele komunikacije skozi povezavo novih tehnologij z različnimi umetniškimi mediji, je Google eno izmed zbirk inovativnih eksperimentov posvetil eksperimentom z obogateno resničnostjo. [6]

3 Lens Studio

Lens Studio [7] je zastonjski program, sprva razvit kot interno orodje za izdelavo filtrov na socialnem omrežju Snapchat, ki pa je postal javno dostopen. Uporabniku omogoča, da sam razvije filter za Snapchat in ga tudi javno objavi. Uporabniško narejeni filtri so dostopni vsem.

Program ima redne posodobitve z novimi funkcionalnostmi in je dobro vzdrževan z zelo aktivno skupnostjo, ki se s tem ukvarja.[1] Prav tako ima

1 Tudi avtor projekta Miha Malenšek je član programa Official Lens Creator pri podjetju Snap, zaradi česar ima dostop do določenih novih funkcionalnosti aplikacije Snapchat in programa za ustvarjanje filtrov Lens Studio. Poleg testiranja in razvoja izboljšav sodeluje v različnih tematskih projektih, ki jih izvaja Snapchat, npr. pri izdelavi lokacijske AR izkušnje na Brandenburških vratih v Berlinu.

podroben in dobro razdelan API (application program interface), ki je dostopen na spletni strani.

Delo v programu temelji na slojih (ang. layer) in virtualnih kamerah, kar omogoča uporabniku, da ima aktivnih več 3D ali 2D elementov brez prekrivanj, ki pa jih morajo uporabniki ustvariti v drugih, zunanjih aplikacijah. Obnašanje elementov filtra lahko uporabniki animirajo z že obstoječimi funkcijami ali pa sami napišejo skripto v programskem jeziku JavaScript. Za preprostejše filtre je dovolj, da je uporabnik splošno računalniško pismen, za kompleksnejše interakcije, kot so zaznavanje dotika, potega ali drugih gest s prsti, spremljanje uporabnikove lokacije v virtualnem 3D prostoru, pa je potrebno znanje programiranja.

Programiranje v Lens Studiu temelji na dogodkih (ang. event-based programming), čemur se odlično poda JavaScript. V programu se definirajo vhodi, ki so elementi filtra. To so lahko 3D modeli, 2D slike v 2D prostoru (ang. screen-space) ali v 3D prostoru (ang. world-space), teksture, virtualne kamere, itd. Delovanje teh elementov je ločeno glede na prostor, v katerem se nahajajo. 3D elementov ne moremo naslikati na 2D ekransko ploskev, ta se poslužuje 2D elementov – slik. Tem lahko dinamično spreminjamo velikost in položaj. 2D elemente pa lahko uporabimo tudi v prostoru, kjer se obnašajo enako kot 3D elementi. V API-ju za Lens Studio so definirane funkcije in dogodki, ki so uporabniku na voljo v tem okolju in jih lahko uporabi za kompleksnejše delovanje svojega filtra. Ti dogodki so lahko odpiranje ust, oči, ali, za kar smo se odločili tudi mi, dvig obrvi. Ko se dogodek zgodi, lahko znotraj skripte definiramo primeren odziv.

Program, ki vodi delovanje filtra, lahko uporabniki pišejo kar v Lens Studiu, saj od verzije 3.0 naprej vključuje interaktiven urejevalnik, ki ima dostop do knjižnjice tako osnovnih JavaScript funkcij kot tudi funkcij, po meri napisanih za Lens Studio.

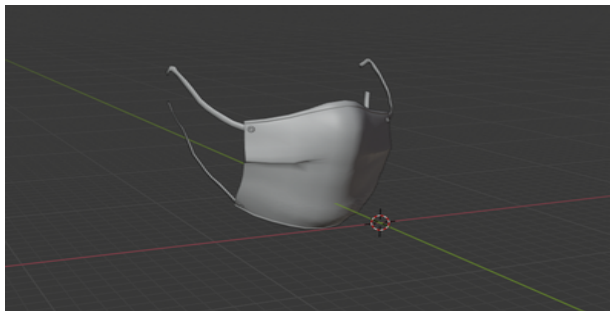
4 Oblikovanje in izdelava 3D modelov

Za izdelavo virtualnih modelov smo izbrali že dobro znano orodje Blender; v članku opozorimo le na posebnosti, na katere moramo biti pozorni pri ustvarjanju modelov za AR filtre, saj predstavljajo nekaj omejitev, ki modeliranje nekoliko otežijo.

Zaradi prostorske omejitve izvoženega filtra (4 MB) moramo biti pazljivi pri velikosti vseh modelov, ki jih pri filtru uporabljamo. Lens Studio je preventivno postavil tudi zgornjo mejo za število trikotnikov na vsakem modelu (65,536), kar prepreči prezahtevnost prikazovanja modelov ob uporabi filtra. Zaradi omenjenih omejitev morajo biti modeli sestavljeni iz enostavnejših oblik z manj detajli, kar včasih pomeni manj realistične prikaze modelov predmetov, za katere bi sicer želeli, da bi bili videti karseda realistični. Postavljena je tudi zgornja meja za ločljivost tekstur, ki jih nanašamo na modele (2048 x 2048), kar pa ne predstavlja večjih težav, še posebej, če teksture ustvarimo sami.

Za doseganje lepših, bolj realističnih, vendar še vedno ne preveč kompleksnih modelov ima Blender na voljo nekaj funkcij, ki ustvarjanje precej olajšajo. Najpomembnejša je funkcija *Smooth Shade,* ki enostavnim modelom z malo ploskvami pogladi robove, kar nam omogoči, da tudi z malim številom ploskev dobimo lepe in gladke oblike.

Za naš filter je bilo potrebno ustvariti dva modela in sicer zaščitno masko (Slika 1) in maskoto »Safey« (Slika 2).



Slika 1: Model zaščitne maske (brez teksture)



Slika 2: »Safey« v akciji

V nasprotju z modeliranjem maskote, ki je fiktivni lik, sama izdelava maske ni zahtevala veliko testiranja, saj smo poustvarili predmet iz realnosti, na katerega videz smo se lahko naslonili in ga skušali čim bolj vestno ponoviti. Vendar pa je bilo samo ustvarjanje, zaradi posebne oblike in tekstur maske, nekoliko zahtevnejše. Poiskati smo morali pravo razmerje med gladkostjo modela in realnostjo gub na maski, ki se pojavijo ob uporabi dejanske maske. Nasproten pristop smo morali ubrati pri modeliranju maskote, saj smo želeli čim bolj enostavno in prikupno maskoto, ki bi bila všeč uporabnikom, hkrati pa bi že sama po sebi spominjala ne ukrepe za zaščito pred korona virusom.

Zato smo se odločili za okroglo obliko (to je namreč oblika mila in milnih mehurčkov), ki lebdi na več milnih mehurčkih – ti naj bi uporabnika spomnili na pomembnost pravilne higiene in umivanja rok. Dodatka klobuka in nekaj preprostih obraznih potez sta maskoto ravno dovolj počlovečila in ji dodala osebnost, da je uporabnika pritegnila in ustrezno nagovorila tudi na čustveni ravni. »Safey-ja« smo izdelali v več različicah z različnimi razmerji velikosti trupa, mehurčkov ter klobuka, z različnimi teksturami, pozicijo klobuka in celo dodatkom nog. Za pomoč pri izbiri modela smo se obrnili na nekaj sošolcev in prijateljev, ki so podali konstruktivno kritiko vseh modelov z vidika bodočih uporabnikov, na koncu smo se odločili za obliko, ki jo prikazuje slika 2.

5 Implementacija filtra v programu Lens Studio

Filter sestavljata dva dela, ki sta povezana prvi s sprednjo kamero in drugi z zadnjo kamero. Na sprednji, t. i. "selfie" kameri se pojavi 3D model kirurške maske, projiciran na uporabnikov obraz. Ob dvigu obrvi se 3D model maske skrči po osi Y, ki je v Lens Studiu globalna vertikalna (ang. world up) os. Projiciranje maske in njeno premikanje v 3D prostoru glede na uporabnikov obraz je avtomatsko implementirano v programu Lens Studio s funkcionalnostjo *head-tracking*. Dodatne interakcije in obnašanje ter videz maske smo implementirali sami.

Na zadnji kameri se pojavi 3D maskota Safey, ki uporabnika opazuje na primerni varnostni razdalji. Če se mu uporabnik preveč približa, Safey odskoči in se umika tako dolgo, dokler ni zopet na primerni varnostni razdalji.

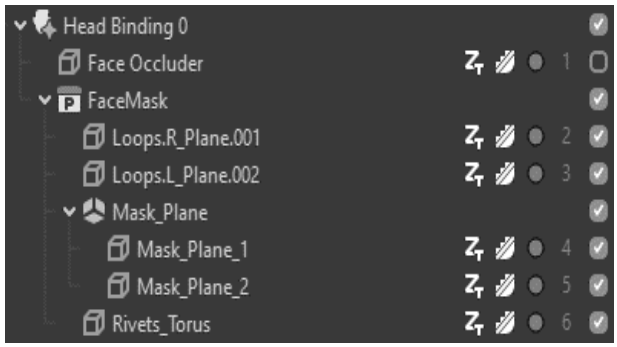
Prikazovanje elementov je vezano na kamere telefona – to ločimo z uporabo dogodka iz API knjižnice Lens Studia, ki zazna, katera kamera je aktivna in kdaj pride do menjave. Temu primerno, glede na aktivno kamero, omogočimo ali onemogočimo željene elemente.

5.1 Implementacija maske

Kot je razvidno na sliki 3 je 3D model maske, imenovan FaceMask, sin objekta Head Binding 0, ta ga projicira uporabniku na obraz in vodi rotacijo in lokacijo 3D modela v prostoru. Sam 3D model maske je razdeljen na več delov – Loops.R_Plane.001, Loops.L_Plane.002, Mask_Plane, in Rivets_Torus. Razloga za to delitev sta: 1) množica različnih materialov, ki so uporabljeni za teksturo 3D modela, 2) prikazovanje maske s pomočjo segmentacije.

Nekateri deli 3D modela maske, kot so trakovi, ki masko držijo na obrazu, namreč ne smejo biti vidni iz vseh zornih kotov, saj jih ponekod prekriva obraz. To smo dosegli s spremljanjem rotacije obraza. Ko se uporabnikov obraz obrne po osi Y za določen kot, se za enega od trakov zamenja njegov sloj (ang. layer). Ker prikazovanje elementov v Lens Studiu temelji na slojih, smo tako dosegli učinek segmentacije obraza, saj menjava sloja 3D elementa omogoča, da ga prikazuje maskirana virtualna kamera. Ta kamera uporablja že v naprej obstoječo segmentacijo obraza uporabnika, torej prikazuje samo to, kar je vidno zunaj obraza.

Odstranjevanje in ponovno pojavitev zaščitne maske smo animirali z uporabo knjižnice funkcij *Tween* v Lens Studiu. Ta funkcionalnost temelji na dogodku dviga obrvi. S preprostim binarnim števcem v programu spremljamo dvig obrvi uporabnika, da se 3D model maske pravilno odstrani in ponovno pojavi. Animacija je preprosta, vendar zaradi hitrosti in preprostosti giba tega ne izda. Model se skrči po osi Y in zavrti po osi X. Transformacije se dogajajo v lokalnem prostoru modela, saj bi v globalnem lahko prišlo do napak, zaradi gibanja obraza uporabnika. Hkrati s transformacijami pa se tudi poveča transparentnost maske, dokler ne postane popolnoma nevidna.



Slika 3: Prikaz slojev za 3D model maske

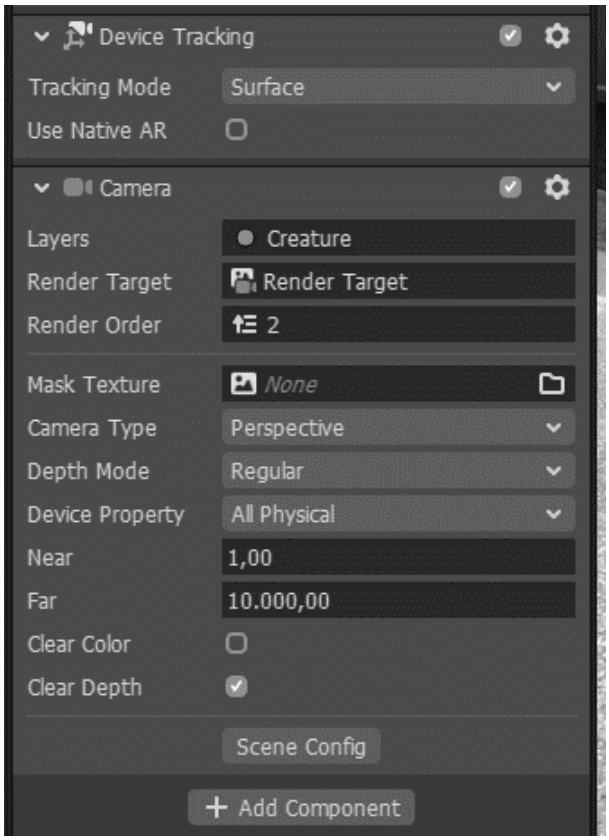


Slika 4: Maska v uporabi

5.2 Implementacija 3D maskote Safey

Implementacija 3D maskote Safey temelji na uporabi funkcionalnosti *Surface Tracking* v Lens Studiu. Ta zahteva svojo virtualno kamero in sceno, kar pa pomeni

tudi svoj sloj. Iz slike 5 je razvidno, da je 3D model maskote na sloju "Creature", virtualna kamera pa ima komponento "Device Tracking" nastavljeno na "surface" oz. tla. Z uporabo te komponente lahko med delovanjem spremljamo uporabnikovo lokacijo v 3D prostoru, katerega izhodišče je tam, kjer je uporabnik začel z uporabo filtra.



Slika 5: Prikaz komponente *Surface Tracking* za maskoto

```
//GIBANJE
var updateEvent = script.createEvent("UpdateEvent");    //Omogoča klic funkcije ob začetku novega framea
updateEvent.bind(function(eventData){
    var sTransform = script.safey.getTransform();   //Transform 3D lika Safey
    var laTransform = script.lookAt.getTransform(); //Transform komponente "Look At" - vedno usmerjena proti kameri
    var laRot = laTransform.getWorldRotation(); //rotacijski vektor "Look At"
    var sRot = sTransform.getWorldRotation(); //rotacijski vektor 3D lika Safey
    var laPos = laTransform.getWorldPosition(); //pozicijski vektor "Look At"
    var sPos = sTransform.getWorldPosition(); //pozicijski vektor 3D lika Safey
    var camPos = script.creatureCamera.getTransform().getWorldPosition(); //pozicijski vektor virtualne kamere

    //Nastavi rotacijo 3D lika Safey tako, da je orientiran proti kameri
    sTransform.setWorldRotation(new quat(sRot.w, sRot.x, laRot.y, sRot.z));
    //Razdalja med likom in virtualno kamero
    var distance = calculateDistance(sPos.x, camPos.x, sPos.z, camPos.z);

    //Razdalja v Z in X osi
    var toCamZ = sPos.z - camPos.z;
    var toCamX = sPos.x - camPos.x;
    //normirana razdalja Z in X osi, glede na celotno razdaljo
    toCamZ = toCamZ/distance;
    toCamX = toCamX/distance;

    //ce je razdalja manjsa od 250 enot, se izracuna nov polozaj lika, ki se uporabi za premik s Tweenom
    if(distance < 250) {
        var newPos = new vec3(sPos.x + toCamX*200, sPos.y, sPos.z + toCamZ*200);
        global.tweenManager.setEndValue(script.creatureTween, "move", newPos);
        global.tweenManager.startTween(script.creatureTween, "move");
    }
});
```

Slika 6: Koda, ki upravlja razdaljo maskote od uporabnika

Sledenje lokaciji omogoča preprosto izračunavanje aktualne razdalje med uporabnikom in projekcijo 3D modela v prostor, kar posledično omogoča tudi spreminjanje te razdalje. Tako smo lahko implementirali varnostno razdaljo, ki jo vzdržuje 3D model maskote do uporabnika.

6 Sklep

V prispevku predstavimo postopek razvoja tematskega filtra za bogatenje resničnosti v programu Lens Studio. Zadovoljni smo z izdelavo 3D modelov zaščitne maske in maskote Safey, saj smo kljub omejitvam ustvarili prepoznavne modele z malo ploskvami, a kljub temu dovolj detajli. Večino dela je predstavljal razvoj filtra, kjer smo implementirali projiciranje maske na uporabnikov obraz in odmikanje maskote od uporabnika. Za to smo uporabili že v Lens Studiu implementirano detekcijo obraza in vezanje filtra na lokacijo, vseeno pa je bilo potrebno še veliko dela za popravljanje malenkosti, kot so razdalja, na katero se maskota umakne, boljše prileganje maske na obraz, pospravljanje in ponovna uporaba maske ... Filter smo objavili na Snapchatu za širšo uporabo in dobili precej pozitivnih odzivov. Menimo, da smo dosegli zastavljeni cilj širjenja ozaveščenosti glede nevarnosti virusa in upoštevanja varnostnih ukrepov.

Od začetka pisanja članka smo zasledili nekaj podobnih filtrov drugih razvijalcev, omeniti velja poziv Združenih narodov digitalnim ustvarjalcem k razvoju AR izkušenj za osveščanje populacije o pandemiji, katerega rezultat je bil mdr. filter »Snap Safe«, ki uporabnike opozarja na vzdrževanje varnostne razdalje. [8]

Do našega filtra lahko dostopate na povezavi: https://www.snapchat.com/unlock/?type=SNAPCODE&uuid=ef1f64fcb331497c942edf0e6bb33453&metadata=01

Literatura

[1] Space miner filter, https://www.reddit.com/r/SnapLenses/comments/ihx96d/space_miner/

[2] Snap filtri Mihe Malenška, https://lensstudio.snapchat.com/creator/w75o4pLLh2EUQAYljbB94A

[3] (RED), https://www.red.org/reditorial/2015/11/30/24-hours-only-get-snapchat-red-filters-for-world-aids-raise-3-million

[4] "The Lengthy History of Augmented Reality", Huffington Post, http://images.huffingtonpost.com/2016-05-13-1463155843-8474094-AR_history_timeline.jpg

[5] The brief history of social media AR filters, https://www.indestry.com/blog/2019/10/11/the-brief-history-of-social-media-ar-filters#:~:text=Just%20a%20couple%20of%20hours,expanded%20on%20Instagram%20as%20well

[6] Experiments with Google, AR Experiments, https://experiments.withgoogle.com/collection/ar

[7] Lens Studio, https://lensstudio.snapchat.com/

[8] Snap Safe, https://wearesocial.com/blog/2020/04/snap-safe-our-answer-to-the-uns-call-to-creatives

Influence of interaction designs of in-vehicle infotainment systems on drivers' preferences

Tomaž Čegovnik[1], Kristina Stojmenova[1], Sašo Tomažič [1] in Jaka Sodnik[1]

E-pošta: tomaz.cegovnik@fe.uni-lj.si, kristina.stojmenova@fe.uni-lj.si, saso.tomazic@fe.uni-lj.si, jaka.sodnik@fe.uni-lj.si

Abstract. *In this paper, we report on a user study where we compared three specific interaction designs: a multifunctional button on the steering wheel, touchscreen and freehand interaction. The button represents a traditional interaction design that proved useful in production vehicles; whereas touchscreen interfaces are very common in new vehicles and freehand interaction is considered as a promising interaction design for future use. In this study, we explored the influence on user experience, usability and user preference of three interface designs in different driving conditions, and for two different task difficulties. The results showed that in most cases the users preferred using the button interface for all of the driving conditions and task difficulties. The usability rating were in-line with the usability findings.*

1 Interaction

Interaction with in-vehicle information and multimedia systems is a very specific type of human-machine interaction (HMI). When designing HMI in vehicles, it is important to take in consideration that driving is the task that has the highest priority while operating the vehicle. The driver should never lose control of the road vehicle, even when interacting with other devices [1]. This applies to manual driving and to highly automated driving, where the driver still has complete responsibility for the driving task. Therefore, when designing user interfaces, we need to consider the drivers 'working' space and its limited ability to perform multiple concurrent tasks, which contributes to the usability of such systems. Furthermore, it is also very important to take into consideration the user preference and consequently their acceptance of the selected user interfaces and interaction designs.

2 Interaction designs

The most commonly used interaction designs in vehicles are tactile buttons, rotary buttons and levers, both on the steering wheel and on the dashboard. However, this type of interaction is suitable for a limited set of functionalities. With the increasing complexity of systems over time, the addition of new functionalities and digital displays, new concepts of HMI are emerging. Vehicles nowadays offer alternative modes of interaction, such as touch screens, multi-function rotary buttons, speech recognition, gesture recognition, hands-free interaction or touchpad [2][3].

When assessing the usability and user experience, it has already been found that performing tasks using freehand interaction can cause less errors, take shorter times to complete the task, and require shorter and fewer eye glances of the road compared to using tactile or rotary buttons [4][5]. Users rated freehand interaction as more enjoyable and found it less annoying compared to using a touch screen [5][6]. The most increasingly used interfaces in new vehicles are touchscreens. However, touchscreens have been associated with problems such as visual distractions and longer eyes-off-road periods. This has already been found in other studies, e.g. that the visual complexity of the managed touch screen increases with the reduction of buttons [7].

In this study, we were interested in the usability and user preference of using buttons, touch screen and freehand interaction for Slovenian drivers. Previous studies have shown that for example, British drivers found the touchscreen as a direct input device more usable than an indirect input device [8]. Another study found that US drivers drive safer using freehand interaction systems, but it causes longer completion times and a higher cognitive load [6].

3 Driving simulator study

The study was performed in a Nervtech [9] driving simulator. It had two factors – driving difficulty and task difficulty. The driving difficulties were easy, where only following a lead vehicle was required and difficult, where the driving required overtaking on the highway. In addition to driving in two different driving conditions, participants had to complete a list of tasks. The tasks were divided into two groups – easy and difficult tasks.

Each participant completed four sets of tasks – two different driving conditions and two different task difficulties. The order of driving modes and task difficulty were counterbalanced to eliminate the doubt of a learning effect affecting the results.

4 Methodology

4.1 Participants

28 participants (19 male and 9 female) aged 19 to 37 years (M = 24.3 years, SD = 3.9 years) participated in the study. All participants had a valid driving license with an average of 6 years of driving experience.

4.2 Interface designs

The interaction system in the simulator consists of two output channels and three input channels. The first output channel is visual and was displayed as a screen on the right side of the driver as in a real vehicle in the middle of the dashboard. The second output channel was audio feedback. Input channels were a touchscreen that was also the visual output channel, a button on the steering wheel and a freehand gesture recognition system.

The system allowed interaction through all three modes (hands-free interaction, button on the steering wheel or via touch screen) at all times. In the first part, participants were allowed to freely choose their favourite mode of interaction, in the second, they had to perform three tasks with a specific way of interaction determined by the experimenter, and in the last part were allowed to again freely choose the interface they prefer most.

4.3 Driving environment

During the experiment, participants had to follow a red vehicle in front of them all the time. The red vehicle periodically changed its speed between 110 km/h and 140 km/h so that the driver was actively involved in driving with dynamic vehicle tracking at all times.

The experiment was conducted on two driving scenarios with lower and higher difficulty. Both driving difficulties took place on a two-lane two-way highway separated by a fence. The difference in the two levels of difficulty was in the traffic density. In lower difficulty, there were no other vehicles in the direction of travel and infrequent traffic in the opposite direction. In difficult driving, there was also some traffic in the direction of travel - a few trucks at a distance of about 500 m, so that the red vehicle also overtook the other participants, and a few fast vehicles in the overtaking lane. This way of driving therefore also required monitoring other traffic in addition to the vehicle in front and was consequently more difficult.

4.4 Tasks

Participants' primary task was safe driving. They were instructed to follow traffic rules and try to drive as they would in real life. They were additionally asked to follow a specific vehicle with appropriate safety distance and to never overtake.

In addition to the primary task of driving, the participants also performed a list of tasks using the in-vehicle information system with all three interaction designs. The tasks were divided into two groups - easy and difficult. Easy tasks required less steps - a maximum of three to complete, while difficult at least five. An example of an easy task was turning on a specific radio station »`Radio -> Rock Radio`«. An example of a difficult task was turning on a specific song »`Media -> Music -> Nina Pušlar -> To mi je všeč`«.

Each participant completed four sets of tasks - two in different driving modes and two with different task difficulties. Each set consisted of 5 tasks: 1 with their choice of interaction, 3 with predetermined interaction design, and at the end one more with their choice of interaction. Half of the participants first drove on the driving scenario with lower difficulty followed by the one with higher difficulty, and the other half started with the one with higher difficulty, followed by driving on the scenario with lower difficulty. Each of these groups were further divided into two groups, with half of them starting the trials with easy tasks followed by difficult tasks and vice-versa.

4.5 Variables

We observed two dependent variables: user preference and usability. *User preference* was the frequency of interface design selection. In each of the four completed sets, participants had the possibility to freely choose from the three interface designs for the given task.

Usability was evaluated with task completion time and drivers cognitive load. Task completion time was defined as the time interval between the start of the task indicated by pressing the "Start" button when instructed, and the moment the correct function was selected in the IVIS. Cognitive load was assessed by measuring changes in the driver's pupil size during task performance and the time intervals in between tasks.

Additionally, we observed *driving safety* with standard deviation of lane position (SDLP) and eyes-off-road time. SDPL was defined as the deviation from the centre of the lane. Eyes-off-road was assessed with the eye-tracker.

5 Results

5.1 User preference

When participants had a free choice of choosing the interaction design, they preferred the button on the steering wheel (56.3%), followed by the touch screen (27.7%), and at last the freehand interaction design (16.1%). See Figure 1.

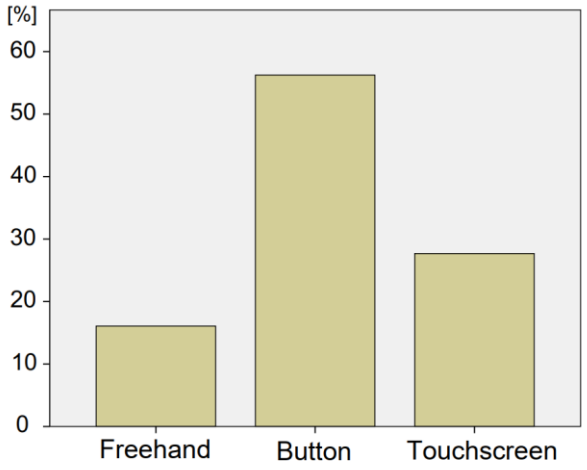


Figure 1. This figure shows how many user choose a specific interaction design.

5.2 Usability

As expected, we found significant effect of task difficulty on task completion time F (1,87), $p < 0,05$, where difficult tasks took more time to complete (M = 16,289 ms, SD = 10,431 ms) compared to easy tasks (M = 10,463 ms, SD = 4,732 ms).

We also found statistically significant effects of interaction design on completion time F (2,87), $p < 0,05$. Post hoc Bonferroni test revealed statistically significant difference in completion times with longest for the freehand interaction (21,787 ms), followed by button on the steering wheel (11,819 ms) and the shortest when performing the tasks using the touchpad (9,720 ms), figure 2.

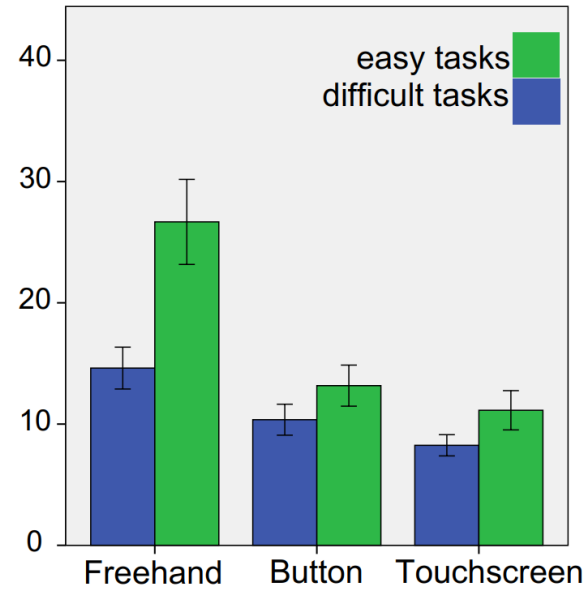


Figure 2. This figure shows the task completion time for easy and difficult tasks.

The results from the pupil dilation also revealed that cognitive load was higher when performing tasks in the driving scenario with higher difficulty (M = 0,06105, SD = 0,04674) compared to lower difficulty (M = 0,07567, SD = 0,04772), F (1,90), $p < 0,05$.
The mode of interaction also affects cognitive load F (2,90), $p < 0,05$. Post hoc tests revealed statistically significantly lower cognitive load while using the button on the steering wheel (M = 0,05327, SD = 0,03898) compared to using touchpad (M = 0,06814, SD = 0,04770) or freehand interaction (M = 0,077990, SD = 0,04438).

5.3 Driving safety

We found significant effects on SDLP only from the task difficultyF (1,36), $p < 0,1$. Difficult tasks (M = 0,1171 m, SD = 0,1989 m) caused higher SDLP compared to easy tasks (M = 0,8325 m, SD = 0,1569 m).

The other driving safety measure, eyes-off-road time, showed significant effects caused driving difficulty F (1,94), $p < 0,05$, where the time drivers looked on the road increased from the lower driving difficulty scenario (M = 0,67205, SD = 0,26028) compared to the more difficult one (M = 0,76276, SD = 0,19209).
The interaction design also had an effect on eyes-off-road time F (2,94), $p < 0,05$. Shortest times were found for touchscreen (M = 0,64369, SD = 0,25979) compared to button on the steering wheel (M = 0,72950, SD = 0,22035) and freehand (M = 0,78352, SD = 0,19058).

6 Discussion

When observed together, the results reveal that most people have chosen the button on the steering wheel as their preferred interaction design. The reason behind this can be discovered in the usability results, where this interface also proves to be the best. The results revealed the least preferred interface to be the freehand interaction, which is also the least useful, as it has high task completion times and results in high cognitive load. This results are in line with a study of US drivers, which also reported higher cognitive load and longer completion times for freehand interaction [6]. The second preferred device, the touchscreen, showed the highest usability results, which is in line with the results from other studies from for example Britain, where they found that a touchscreen showed a higher usability as a direct input device [8].
However, we find that safety related measures are not directly related to preference and usability, as users don't feel a direct impact on safety. The result of eyes-off-road time is expected, but it is not related to usability, as usability measures do not take into account the primary task of driving. We found the shortest eyes-off-road times in freehand and button interaction, and at longest via the touch screen.
Considering user preference, impact on driving safety and thus usability of the interaction designs, based on the results of this study, we can conclude the traditional button on the wheel to be most favourable for Slovenian drivers.

Acknowledgment

This work has been supported by the Slovenian Research Agency within the bilateral agreement BI-US/18-19-086, and the research program ICT4QoL - Information and Communications Technologies for Quality of Life, grant number P2-0246.
The authors thank Nervtech for providing the driving simulator software.

Literature

[1] B. Pfleging, S. Schneegass, D. Kern, and A. Schmidt. 2014. Vom Transportmittel zum rollenden Computer – Interaktion im Auto. *Informatik-Spektrum* 37, 5: 418–422.

[2] Christian Müller and Garrett Weinberg. 2011. Multimodal Input in the Car, Today and Tomorrow. IEEE Multimedia 18, 1: 98–103.

[3] Andreas Riener and Michael Rossbory. 2011. Natural and intuitive hand gestures: a substitute for traditional vehicle control? Proceedings of the AutomotiveUI 11: 2.

[4] Kenneth Majlund Bach, Mads Gregers Jæger, Mikael B. Skov, and Nils Gram Thomassen. 2008. You can touch, but you can't look: interacting with in-vehicle systems. Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08, ACM Press, 1139.

[5] Lisa Graichen, Matthias Graichen, and Josef F. Krems. 2019. Evaluation of Gesture-Based In-Vehicle Interaction: User Experience and the Potential to Reduce Driver Distraction. Human Factors: The Journal of the Human Factors and Ergonomics Society: 001872081882425.

[6] Keenan R May, Thomas M Gable, and Bruce N Walker. 2014. A Multimodal Air Gesture Interface for In Vehicle Menu Navigation. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14*, ACM Press, 1–6.

[7] Ayse Leyla Eren, Gary Burnett, and David R. Large. 2015. Can in-vehicle touchscreens be operated with zero visual demand? An exploratory driving simulator study. The 4th International Conference on Driver Distraction and Inattention.

[8] Catherine Harvey, Neville A. Stanton, Carl A. Pickering, Mike McDonald, and Pengjun Zheng. 2011. To twist or poke? A method for identifying usability issues with the rotary controller and touch screen for control of in-vehicle information systems. *Ergonomics* 54, 7: 609–625.

[9] nervtech | simuation technologies. *nervtech*. Retrieved March 25, 2019 from https://www.nervtech.com.

Avtomatika in robotika

Automatic Control and Robotics

Vodenje rotacijskega invertiranega nihala

Aljaž Blažič, Gorazd Karer

Fakulteta za elektrotehniko, Univerza v Ljubljani
E-pošta: aljaz96@gmail.com, gorazd.karer@fe.uni-lj.si

Control of a rotary inverted pendulum

Abstract. *The rotary inverted pendulum represents a compact platform of a nonlinear mechanical system with fast dynamics. The system is found in many laboratories, as it represents an impressive platform for the demonstration and development of control of distinctly nonlinear systems.*

The paper includes theoretical modelling, simulation of the dynamic system and development of a closed-loop control system for the rotary inverted pendulum. On a linearized model of the system, a linear quadratic regulator was designed in order to stabilize the system in the upper unstable equilibrium position. The states of the process were estimated with an extended Kalman filter. A swing-up regulator was implemented, which brought the pendulum from the lower equilibrium position close to the upper unstable position. These approaches were implemented and verified in simulation and then tested on the pilot plant [1].

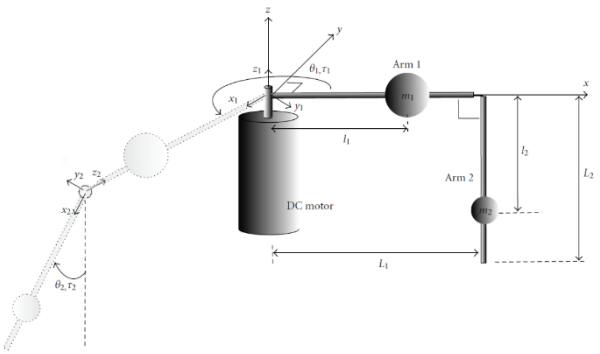
1 Uvod

Rotacijsko invertirano nihalo predstavlja kompaktno platformo nelinearnega mehanskega sistema s hitro dinamiko. Sistem najdemo v številnih laboratorijih, saj predstavlja impresivno platformo za demonstracijo in razvoj vodenja izrazito nelinearnih sistemov.

Prispevek obsega teoretično modeliranje, simulacijo dinamičnega sistema in razvoj sistema zaprtozančnega vodenja rotacijskega invertiranega nihala. Na linearizirnem modelu sistema je bil načrtan optimalni regulator stanj z namenom stabilizacije sistema v zgornji ravnovesni legi. Stanja procesa smo ocenjevali z razširjenim Kalmanovim filtrom. Implementiran je bil regulator za dvig nihala, ki je nihalo pripeljal iz spodnje ravnovesne lege v bližino zgornje nestabilne ravnovesne lege. Omenjene pristope smo načrtali in preverili simulacijsko, nato pa smo jih preizkusili na pilotni napravi [1].

2 Opis naprave

Nihalo Furuta nihalo je bilo prvič razvito leta 1992 na Tokijskem inštitutu za tehnologijo. Naprava je sestavljena iz ročice, ki rotira v horizontalni ravnini okrog navpične osi. Na rotirajočo ročico je pritrjeno nihalo, ki prosto niha okrog vzdolžne osi ročice (slika 1). Enosmerni motor je mehansko povezan z ročico, na katero deluje z navorom τ. Ročica ima dolžino L_1 in maso m_1, nihalo pa dolžino L_2 ter maso m_2. Dolžini l_1 in l_2 sta razdalji od osi vrtenja do težišča. Vztrajnostna momenta okoli težišča ročice in nihala sta J_1 in J_2. Koeficienta dušenja sta predstavljena z b_1 in b_2, kjer je b_1 dušenje v ležajih motorja, b_2 pa dušenje ležajev, ki povezujejo ročico in nihalo. Na rotacijsko invertirano nihalo lahko gledamo kot na sistem z enim vhodom in dvema izhodoma. Vhod u predstavlja napetost, s katero vzbujamo enosmerni motor. Izhoda ϑ_1 in ϑ_2 predstavljata zasuk ročice in kot nihala.



Slika 1: Shematski prikaz rotacijskega invertiranega nihala [2]

3 Nelinearni matematični model

Za razvoj nelinearnega modela smo uporabili teoretično modeliranje. Celoten postopek izpeljave s pomočjo Lagrangove enačbe je prikazan v [2]. Ravnotežni enačbi iz članka smo modificirali, da se ujemata s koordinatnim sistemom modela na (sliki 1). Enačbi (1) in (2) prikazujeta ravnotežni enačbi nelinearnega modela:

$$\alpha\ddot{\vartheta}_1 + \beta\dot{\vartheta}_1\dot{\vartheta}_2 - \gamma\ddot{\vartheta}_2 + \delta\dot{\vartheta}_2^2 + b_1\dot{\vartheta}_1 = \tau, \qquad (1)$$

$$\gamma\ddot{\vartheta}_1 - [J_2 + m_2 l_2^2]\ddot{\vartheta}_2 + \frac{1}{2}\beta\dot{\vartheta}_1^2 - g m_2 l_2 \sin(\vartheta_2) - b_2\dot{\vartheta}_2 = 0, \qquad (2)$$

kjer so:

$$\alpha = J_1 + m_1 l_1^2 + m_1 L_1^2 + [J_2 + m_2 l_2^2]\sin^2(\vartheta_2),$$
$$\beta = [J_2 + m_2 l_2^2]\sin(2\vartheta_2),$$
$$\gamma = m_2 L_1 l_2 \cos(\vartheta_2),$$
$$\delta = m_2 L_1 l_2 \sin(\vartheta_2).$$

Za opisani nelinearni model veljajo naslednje predpostavke [2], [3]:

- Gred motorja in ročica sta togo spojeni.
- Ročica in nihalo sta povsem toga.

- Uporabili smo le koeficiente viskoznega trenja. Zanemarili smo dušenje, ki nastane zaradi zračnega upora, statičnega in Coulombovega trenja.
- Vztrajnostni moment smo upoštevali le glede na glavno os gibanja.
- Vztrajnostni moment motorja je zanemarljiv.

Določili smo matematični model, ki povezuje vzbujalno napetost motorja u in navor τ. Upoštevali smo predpostavko, da je dinamika zaradi induktivnosti dovolj hitra, da jo lahko zanemarimo. Iz Kirchoffovega zakona izrazimo povezavo med napetostjo u in tokom i na motorju oblike:

$$R_m i + K_m \dot{\vartheta}_1 = u, \tag{3}$$

pri čimer R_m predstavlja električno upornost motorja, K_m pa inducirano protinapetostno konstanto. Enačba (4) prikazuje linearno odvisnost med električnim tokom i na motorju in navorom motorja τ:

$$\tau = K_t i, \tag{4}$$

pri čimer K_t predstavlja navorno konstanto motorja [3].

Enačbi (3) in (4) vstavimo v ravnotežno enačbo (1). V ravnotežni enačbi vpeljemo nekaj dodatnih spremenljivk, katere nimajo direktnega fizikalnega pomena, a naredijo zapis kompaktnejši:

$$\alpha\ddot{\vartheta}_1 + \beta\dot{\vartheta}_1\dot{\vartheta}_2 - \gamma\ddot{\vartheta}_2 + \delta\dot{\vartheta}_2^2 + \hat{b}_1\dot{\vartheta}_1 = K_u u, \tag{5}$$

$$\gamma\ddot{\vartheta}_1 - \hat{J}_2\ddot{\vartheta}_2 + \frac{1}{2}\beta\dot{\vartheta}_1^2 - g m_2 l_2 \sin(\vartheta_2) - b_2\dot{\vartheta}_2 = 0, \tag{6}$$

kjer so:

$$\begin{aligned}
\alpha &= J_1 + m_1 l_1^2 + m_1 L_1^2 + \hat{J}_2 \sin^2(\vartheta_2), \\
\beta &= \hat{J}_2 \sin(2\vartheta_2), \\
\gamma &= m_2 L_1 l_2 \cos(\vartheta_2), \\
\delta &= m_2 L_1 l_2 \sin(\vartheta_2), \\
\hat{J}_2 &= J_2 + m_2 l_2^2, \\
\hat{b}_1 &= b_1 + K_t K_m R_m^{-1}, \\
K_u &= K_t R_m^{-1}.
\end{aligned}$$

Nelinearni matematični model (5-6) pretvorimo v zapis v prostoru stanj:

$$\begin{aligned}
\dot{x}_z &= f_z(x_z, u_z), \\
y_z &= C x_z,
\end{aligned} \tag{7}$$

pri čimer $x_z(t) = [\vartheta_1 \quad \vartheta_2 \quad \dot{\vartheta}_1 \quad \dot{\vartheta}_2]^T$ predstavlja vektor stanj sistema, $u_z(t)$ vzbujanje sistema, $y_z(t)$ vektor izhoda sistema, C izhodno matriko, $f_z(x_z, u_z)$ pa vektorsko nelinearno preslikavo, ki jo lahko izrazimo:

$$f_z(x_z, u_z) = \frac{1}{\Delta}\begin{bmatrix} \Delta\dot{\vartheta}_1 \\ \Delta\dot{\vartheta}_2 \\ b_2\gamma\dot{\vartheta}_2 + \hat{J}_2\mu + \gamma\eta \\ b_2\alpha\dot{\vartheta}_2 + \gamma\mu + \alpha\eta \end{bmatrix} + \frac{1}{\Delta}\begin{bmatrix} 0 \\ 0 \\ -\hat{J}_2 K_u \\ -\gamma K_u \end{bmatrix} u_z, \tag{8}$$

kjer so:

$$\begin{aligned}
\Delta &= \gamma^2 - \hat{J}_2\alpha, \\
\mu &= \hat{b}_1\dot{\vartheta}_1 + \beta\dot{\vartheta}_1\dot{\vartheta}_2 + \delta\dot{\vartheta}_2^2, \\
\eta &= \left[g m_2 l_2 - \hat{J}_2 \cos(\vartheta_2)\,\dot{\vartheta}_1^2\right] \sin(\vartheta_2).
\end{aligned}$$

Izhod sistema predstavljata prvi dve stanji, zato lahko izhodno matriko zapišemo kot $C = [I_{2\times2}\ 0_{2\times2}]$.

3.1 Določitev parametrov nelinearnega matematičnega modela

Najprej smo določili parametre elementov sistema, ki jih je bilo mogoče enostavno izmeriti. Rezultati meritev so prikazani v tabeli 1 [3].

Tabela 1. Izmerjeni parametri modela

Oznaka	Vrednost
L_1	95 mm
L_2	72 mm
m_2	3,30 g
l_2	52,5 mm

Ostale parametre smo ocenili z izvajanjem eksperimentov na sistemu. Uporabili smo optimizacijo z metodo Nelder-Mead, kjer smo iskanji minimum izbrane kriterijske funkcije. Določanje parametrov modela smo razdelili na dva dela. V prvem delu smo fiksirali ročico. S tem smo zmanjšali kompleksnost modela in lažje določili en del parametrov modela. V drugem delu smo izvajali eksperimente na celotnem modelu in tako določili še vse preostale parametre.

4 Linearni matematični model

Linearizacijo nelinearnega matematičnega modela smo izvedli s pomočjo aproksimacije s Taylorjevo vrsto v okolici zgornje ravnovesne lege $x_0 = [0, \pi, 0, 0]^T$ in $u_0 = 0$. Definicija spremenljivk v delovni točki (x_0, u_0): $\tilde{x}_z = x_z - x_0$, $\tilde{u}_z = u_z - u_0$ in $\tilde{y}_z = y_z - C x_0$.

Linearni matematični model v delovni točki opišemo z naslednjima enačbama:

$$\begin{aligned}
\dot{\tilde{x}}_z &= A_z\tilde{x}_z + B_z\tilde{u}_z, \\
\tilde{y}_z &= C\tilde{x}_z,
\end{aligned} \tag{9}$$

kjer sta A_z in B_z definirani kot Jacobijevi matriki:

$$A_z = \left.\frac{\partial f_z(x_z,u_z)}{\partial x_z}\right|_{(x_0,u_0)} \text{ in } B_z = \left.\frac{\partial f_z(x_z,u_z)}{\partial u_z}\right|_{(x_0,u_0)}. \tag{10}$$

Zapis zveznega linearnega sistema v prostoru stanj (9) pretvorimo v diskretnega z metodo stopnične invariance [4]. Za čas vzorčenja t_s smo vzeli 5 ms, ki je tudi uporabljen na pilotni napravi. Enačbi (11) prikazujeta zapis diskretnega linearnega sistema v prostoru stanj:

$$\begin{aligned}
x_k &= A_d x_{k-1} + B_d u_{k-1}, \\
y_k &= C x_k,
\end{aligned} \tag{11}$$

kjer ocenimo matriki A_d in B_d z naslednjima enačbama:

$$A_d \doteq \sum_{i=0}^{n} \frac{t_s^i}{i!} A_z^i \text{ in } B_d \doteq \left[\sum_{i=0}^{n} \frac{T_s^{i+1}}{(i+1)!} A_z^i\right] B_z. \quad (12)$$

V praksi se izkaže, da dobimo zadovoljive rezultate že pri majhnih vrednostih n.

5 Regulator za dvig nihala

V poglavju predstavimo metodo, s katero pripeljemo nihalo iz spodnje ravnovesne lege v okolico zgornje. Regulator za dvig nihala temelji na potenciranju pozicije nihala. Sistem močneje vzbujamo, ko smo blizu spodnje ravnovesne lege in vse manj, ko se bližamo zgornji ravnovesni legi. Regulacijski zakon predstavimo z enačbo:

$$u_{su} = k_v |\tilde{x}_z(2)|^n sign\big(\tilde{x}_z(4)\cos\big(\tilde{x}_z(2)\big)\big), \quad (13)$$

pri čimer k_v in n predstavljata konstanti s katerima vplivamo na dinamiko nihanja sistema [5]. Funkcija $sign$ vrne predznak podanega argumenta. Spremenljivka $\tilde{x}_z(2)$ predstavlja odklon nihala od zgornje ravnovesne lege, medtem ko $\tilde{x}_z(4)$ predstavlja kotno hitrost nihala. Konstanti smo določili s poizkušanjem na pilotni napravi in znašata $k_v = -0.03$ in $n = 3$.

6 Optimalni regulator stanj

Optimalni regulator stanj načrtamo z minimizacijo kriterijske funkcije:

$$J = \int_0^\infty (\tilde{x}_z{}^T Q \tilde{x}_z + r\tilde{u}_z^2)\, dt, \quad (14)$$

pri čimer Q predstavlja pozitivno semidefinitno in simetrično matriko, r pa pozitiven skalar [6], [7]. Optimalni regulator stanj smo implementirali s pomočjo Matlabove funkcije *lqr*. Za vhode funkcija sprejme matriki A_z in B_z ter matriko Q in skalar r. Kot izhod nam funkcija poda vektor k. Regulirno veličino izračunamo z enačbo:

$$\tilde{u}_z = -k^T \tilde{x}_z. \quad (15)$$

Enačbo (15) modificiramo, da zagotovimo sledenje referenčnemu signalu zasuka ročice.

$$\tilde{u}_z = k^T(\tilde{r} - \tilde{x}_z) \quad (16)$$

pri čimer je referenčni signal $\tilde{r} = [r_{\vartheta_1} \quad 0 \quad 0 \quad 0]^T$. S prvo komponento spremenljivke $\tilde{r}$ spreminjamo referenco položaja ročice.

Optimalni regulator stanj deluje dobro le v okolici delovne točke, okoli katere smo ga načrtali. Enačba (17) prikazuje preklapljane med optimalnim regulatorjem stanj in regulatorjem za dvig nihala:

$$u_z = \begin{cases} \tilde{u}_z, & |\tilde{x}_z(2)| < \dfrac{\pi}{6} \\ u_{su}, & sicer \end{cases} \quad (17)$$

Pomanjkljivost optimalnega regulatorja stanj je v tem, da morajo biti vsa stanja sistema merljiva. Dostopnost vseh stanj pa je mnogokrat omejena zaradi raznih tehnoloških in ekonomskih omejitev. Problemu merljivosti stanj se je mogoče izogniti z njihovim ocenjevanjem iz ostalih merljivih spremenljivk.

7 Ocenjevanje stanj procesa

V poglavju predstavimo pristop k ocenjevanju stanj procesa. Problem nedostopnosti stanj rešimo z uporabo metode razširjenega Kalmanovega filtra.

7.1 Razširjeni Kalmanov filter

Razširjeni Kalmanov filter (v nadaljevanju EKF) je nelinearna izpeljanka Kalmanovega filtra, ki linearizira model v okolici trenutne ocene. Uporaba EKF omogoča združevanje podatkov tako, da je rezultat zanesljivejši od vsake posamezne meritve s senzorja, saj upošteva lastnosti nelinearnega modela ter šuma na vhodu kot tudi izhodu procesa. Algoritem ERK poteka v dveh korakih. Najprej določimo predikcijsko oceno (*a priori*) naslednjega stanja. Ocena je odvisna od prejšnjega stanja kot tudi od vhoda v sistem. Natančnost rezultata je odvisna od natančnosti matematičnega modela sistema. V drugem delu pa izvedemo korekcijsko oceno (*a posteriori*) trenutnega stanja. Ocena je odvisna od predikcijske ocene stanja in trenutne meritve [8].

Za vhodni in izhodni šum mora veljati, da njuni kovariančni matriki znašata:

$$\begin{aligned} Q_k &= E[w_k w_k^T] \\ R_k &= E[v_k v_k^T] \end{aligned} \quad (18)$$

pri čimer predstavlja E operator matematičnega upanja, w_k vrednost vhodnega šuma in v_k vektor izhodnega šuma. S poskušanjem smo nastavili kovariančne matrike šuma na diagonalne matrike konstantnih vrednosti.

Enačbi (19) opisujeta predikcijski korak [9]:

$$\begin{aligned} \hat{x}_{k|k-1} &\doteq \hat{x}_{k-1|k-1} + t_s f_z\big(\hat{x}_{k-1|k-1}, u_{k-1}\big) \\ P_{k|k-1} &= A_d^* P_{k-1|k-1} A_d^{*T} + B_d^* Q_{k-1} B_d^{*T} \end{aligned} \quad (19)$$

pri čimer $\hat{x}$ predstavlja oceno stanj, P pa oceno kovariance. Z indeksiranjem $k|k-1$ povemo, da gre za oceno spremenljivke v trenutku k iz podatkov v trenutku $k-1$. Matriki A_d^* in B_d^* določimo z enačbama (12). Uporabljene matrike A_z in B_z določimo z enačbama (10), le da namesto delovne točke (x_0, u_0) uporabimo oceno stanja $\big(\hat{x}_{k-1|k-1}, u_{k-1}\big)$.

Enačbe (20) opisujejo korekcijski korak [9]:

$$\begin{aligned} K_k &= P_{k|k-1} C^T \big(C P_{k|k-1} C^T + R_k\big)^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k\big(y_k - C\hat{x}_{k|k-1}\big) \\ P_{k|k} &= P_{k|k-1} - K_k C P_{k|k-1} \end{aligned} \quad (20)$$

Pred začetkom algoritma EKF je potrebno določiti začetno oceno stanja $\hat{x}_0$ in začetno oceno kovarinace P_0. Ob zagonu se nihalo nahaja v spodnji ravnovesni legi,

zato začetno stanje $\hat{x}_0$ nastavimo na nič. Začetno vrednost kovariance P_0 pa nastavimo na dovolj veliko vrednost.

8 Preizkus na pilotni napravi

Model, regulator in EKF smo najprej implementirali v simulaciji in preizkusili njihovo delovanje. Ko smo bili zadovoljni z obnašanjem sistema, smo delovanje preizkusili še na pilotni napravi. Pojavile so se težave, saj model ni popolnoma odražal vseh lastnosti realne naprave. Pri majhnih vzbujalnih napetostih je lepenje in trenje v jermenskem prenosu preveliko, da bi se motor zavrtel. Neželeno nelinearnost smo poskušali odpraviti z zamikom napetosti, kot je predlagano v [1]. S poizkušanjem smo določili matriko Q in skalar r, tako da je optimalnemu regulatorju stanj uspelo sistem stabilizirati v zgornji ravnovesni legi ter hkrati slediti referenci ročice.

Z EKF ocenjujemo tudi vrednosti, ki jih merimo s senzorji zasuka. Izkazalo se je, da dobimo boljše rezultate, ko upoštevamo meritve direktno iz senzorjev. Tako s EKF ocenjujemo le hitrosti ročice in nihala, ki nista direktno merljivi na napravi. Razlog za nenatančno napoved EKF je v slabi kakovosti modela.

Slika 2 prikazuje primer eksperimenta, kjer primerjamo odzive sistema v simulaciji in na napravi. Nihalo se najprej nahaja v spodnji ravnovesni legi. Z regulatorjem za dvig nihala sistem pripeljemo v okolico zgornje ravnovesne lege. V simulaciji potrebuje nihalo en nihaj, medtem ko na napravi potrebuje dva nihaja, da se približa zgornji ravnovesni legi. To nakazuje na odstopanje med modelom in napravo. Ko se približamo zgornji ravnovesni legi, se zgodi preklop na optimalni regulator stanj, ki sistem stabilizira v zgornji ravnovesni legi, hkrati pa še sledi referenci ročice. Na napravi opazimo, da se ročici ne uspe ustaliti ob referenci. Nihanje okrog nje je najverjetneje posledica zračnosti prenosnega mehanizma.

9 Zaključek

Za inverzno rotacijsko nihalo je značilno, da izkazuje izrazite nelinearne lastnosti. Nelinearnost je otežila določitev parametrov modela, kar je zmanjšalo tudi kakovost vodenja. Problem nelinearnosti bi lahko zmanjšali z nadgradnjo strukture modela, kjer bi upoštevali še statično in Coulombovo trenje. Na pilotni napravi nam je uspelo nihalo dvigniti v okolico zgornje ravnovesne lege in ga tam tudi stabilizirati. V nadaljevanju bi se lahko posvetili izboljšavi modela in poskušali odpraviti nihanje ročice okoli reference.

Literatura

[1] M. Bošnak, „Majhno invertirano nihalo“, Zbornik 25. mednarodne Elektrotehniške in računalniške konference ERK, str. 143–146, 2016.

[2] B. Cazzolato in Z. Prime, „On the dynamics of the Furuta pendulum“, Journal of Control Science and Engineering, 2011.

[3] M. Pernuš in M. Peterka, „Modeliranje rotacijskega nihala“, Zbornik 10. konference Avtomatizacija v industriji in gospodarstvu, 2017.

[4] S. Blažič, „Digitalno vodenje, Prosojnice predavanj“, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2018.

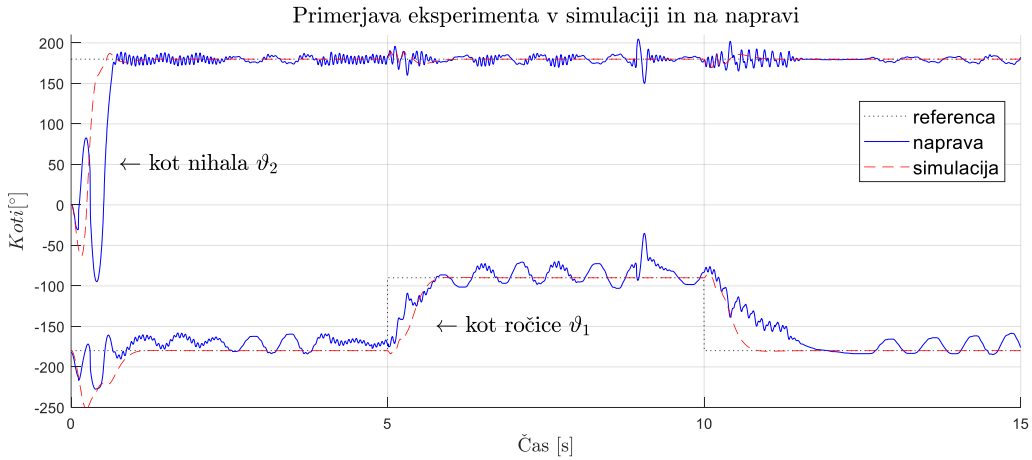
[5] P. Seman, B. Rohal Ilkiv, M. Juhas, in M. Salaj, „Swinging up the Furuta pendulum and its stabilization via model predictive control“, Journal of Electrical Engineering, vol.64,str.152–158, 2013.

[6] E. Guechi, S. Bouzoualegh, Z. Youcef in S. Blažič, „MPC control and LQ optimal control of a two-link robot arm: A comparative study“, Machines, vol. 6, str. 1–14, avg. 2018.

[7] R. Mohammadi Asl, A. Mahdoudi, E. Pourabdollah in G. Klančar, „Combined PID and LQR controller using optimized fuzzy rules“, Soft Comput, vol. 23, str. 5143–5155, jul. 2019.

[8] R. Juhant, „Sistemi za navigacijo v okoljih s senzorskimi in zunanjimi omejitvami“, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2015.

[9] G. Klančar, „Autonomous mobile systems: Nondeterministic events in mobile systems, Prosojnice predavanj“, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2019.



Slika 2: Odziv sistema v simulaciji in na pilotni napravi. Prikaz spreminjanja kota nihala in odmika ročice.

Avtomatizacija namiznega nogometa

Matevž Bošnak, Andrej Zdešar, Simon Tomažič, Goran Andonovski, Gregor Klančar, Sašo Blažič, Igor Škrjanc

Fakulteta za elektrotehniko, Univerza v Ljubljani
E-pošta: matevz.bosnak@fe.uni-lj.si

Table football automation

We present a project for automation of a table football game that enables competitions with mixed (computer or human) players on each side. The system relies on a pair of overhead cameras to capture the image of the playing field and decode the position of all playing rods and the position of the ball. Our approach skips the slow standard camera image rectification methods and for the purpose of playing rods tracking extracts the expected pixel positions from the camera directly. Moreover, we present an approach to improving the resolution of an optical Gray code encoder by using the phase data captured in the camera image. Actuation of the rods is implemented using modular drive units consisting of linear-rotary axes, designed to be attached to any of the rods. The goal of the project is a development of an eye-catching system for promotion of studies at the Faculty of Electrical engineering, as well as a learning platform that is used during the laboratory exercises covered by the Laboratory of Control Systems and Cybernetics.

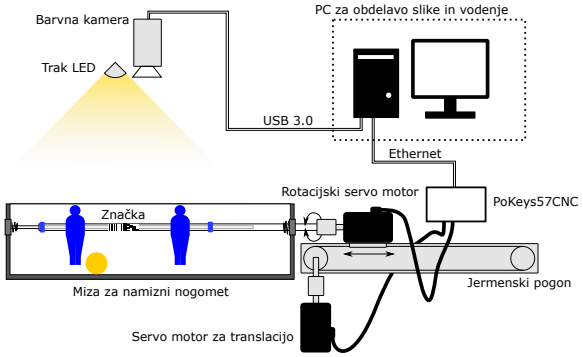
1 Uvod

V sklopu internih projektov FE za izboljšanje kakovosti pedagoškega procesa in promocije IPP&P 2019 smo v Laboratoriju za avtomatiko in kibernetiko predlagali projekt avtomatizacije namiznega nogometa. Projekt predstavlja izdelavo laboratorijske naprave, ki bo lahko uporabljena tako za promocijo študija elektrotehnike, kot tudi za pedagoške namene (npr. izvedbe laboratorijskih vaj [1], izvedbe študentskih tekmovanj v sklopu izven kurikularnih dejavnosti ipd.).

Namen projekta je nadgradnja standardne mize za namizni nogomet z vsemi potrebnimi komponentami za izvedbo različnih načinov igre. V nasprotju s podobnimi projekti [2],[3],[4] želimo ponuditi tudi možnost igre z mešanimi ekipami, kjer bodo člani ekipe morali pripraviti algoritme za računalniškega tekmovalca ter hkrati izuriti enega izmed svojih človeških članov. Osnovo pri naši implementaciji predstavlja sistem kamer nad igriščem, ki omogoča hkratno sledenje žogi in položajev igralcev, ter modularni pogonski sistem, ki ga bomo lahko namestili na katero koli igralno palico.

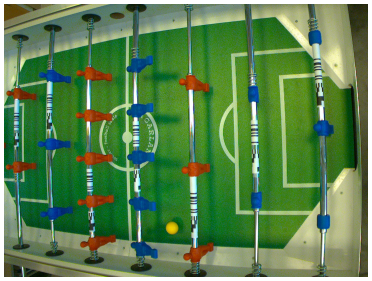
2 Zasnova sistema

Avtomatizirana miza za namizni nogomet je sestavljena iz senzorskega, pogonskega in procesorskega sistema (slika 1). Za uspešno delovanje je potrebno razpoznati položaj in smer premikanja žoge, položaje ter rotacije vseh igralnih palic ter po ustreznem algoritmu voditi krmiljene igralne palice. Osnovna ideja je bila razvoj prilagodljivega sistema, ki nam bo omogočal avtomatizirano vodenje do 4 poljubnih igralnih palic. Predstavljena rešitev je zato zasnovana modularno – senzorski in procesorski sistem je nameščen fiksno nad in pod igralno mizo, pogonski sistem pa je mogoče namestiti na poljubno izbrana mesta. Tovrsten pristop omogoča izvedbo kibernetske igre namiznega nogometa človek proti računalniku ter igre mešanih ekip, torej ekip človeka in računalnika proti drugi takšni mešani ekipi.



Slika 1: Zasnova sistema

Sistem za strojni vid predstavlja osnovni senzorski sistem za zajem podatkov o položaju ter orientaciji igralnih palic in položaju žoge [5]. Nad igriščem sta nameščeni dve kameri proizvajalca Basler z ločljivostjo 1,6 milijona točk. Širok kot (53,2°) uporabljenih leč omogoča pokritje celotnega igrišča že z eno kamero, vendar sta dve kameri uporabljeni za izboljšanje delovanja zaradi pogostega prekrivanja žoge s strani igralnih palic ali igralcev (slika ene od kamer je prikazana na sliki 2). Obdelava slike se izvaja na osebnem računalniku, celoten sistem pa omogoča zajem slik s frekvenco do 200 slik na sekundo. Neodvisnost delovanja sistema za zajem slike od ambientalne svetlobe je zagotovljena z umetno osvetlitvijo v obliki traku LED nad igriščem.

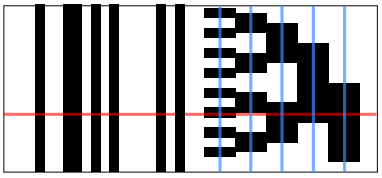


Slika 2: Slika zajeta z eno od kamer

Pogonski sistem je sestavljen iz krmilnika PoKeys57CNC, linearnega ter rotacijskega pogona. Linearni pogon je izveden s pomočjo jermensko gnane linearno pomične osi, rotacijski pogon pa uporablja direkten sklop osi servo motorja na igralno palico preko potisnega ležaja in adapterja za ročaj palice.

2.1 Detekcija položajev igralcev

Za uspešno izvajanje naprednejših strategih igre pri namiznem nogometu je potrebno poznati položaje vseh posameznih figur igralcev – poznati je potrebno linearne položaje ter rotacije vseh 8 igralnih palic. Namesto fizičnih senzorjev smo se odločili za uporabo barvne kamere, kot predlagajo avtorji v [6]. V ta namen smo vse palice opremili z značkami, ki omogočajo sledenje po obeh prostostnih stopnjah. Značka (slika 3) je sestavljena iz dveh delov, črtne kode na levi ter zapis s 5-bitno Grayevo kodo na desni. Pri zajemu slike igrišča s kamero pride zaradi upo-

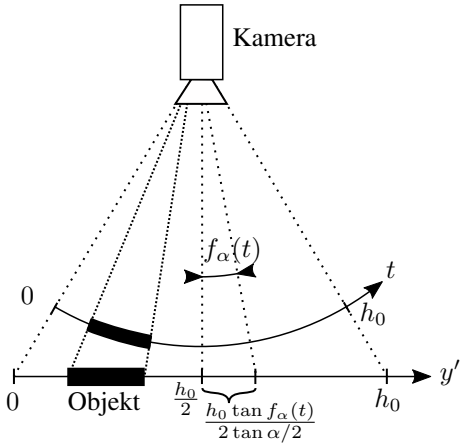


Slika 3: Grafika značke z označenim nizom točk L_m (rdeča vodoravna črta) in področji za vzorčenje Grayeve kode (presečišče modrih in rdečih črt)

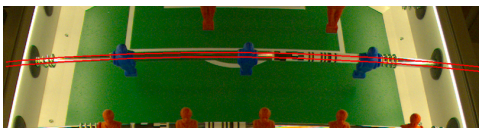
rabe širokokotne leče na kameri do sodčkastih popačenj v sliki[7] – ravne črte so tako preslikane v krivulje na sliki. Da se izognemo časovno potratni operaciji odpravljanja popačenj na celotni zajeti sliki, iz slike izluščimo le točke, ki pripadajo igralni palici. Za vsako od palic z indeksom palice m uporabimo preslikavo

$$g_m\left(\left\lfloor \frac{h_0}{2}\left(1+\frac{\tan f_\alpha(t)}{\tan \alpha/2}\right)\right\rfloor\right) \leftarrow I_o\left[x_m(t), t\right], \quad (1)$$

kjer z $x_m(t) = a_m \cdot t^2 + b_m \cdot t + c_m$ opišemo krivuljo, ki ustreza obliki popačene slike ravne palice. Parameter t uporabimo za glavno spremenljivko preslikave, ki označuje točko na loku v popačeni sliki). h_0 ustreza vertikalni ločljivosti zajete slike, $\alpha = 53{,}2°$ vidnemu kotu kamere, s funkcijo $f_\alpha(t) = \frac{t-h_o/2}{h_o} \cdot \alpha$ pa preslikamo spremenljivko t v kotni prostor kamere (slika 4). Parametre a_m, b_m in c_m določimo za vsako palico posebej v postopku kalibracije sistema pred uporabo.



Slika 4: Compensation for the distance variation between tracked object and the camera



Slika 5: Območje palice

S preslikavo podano z enačbo (1) za vsako igralno palico pridemo do niza točk L_m, ki ga nato najprej filtriramo s pasovno prepustnim filtrom ter uporabimo za iskanje linearnega položaja palice. Položaj je določen s položajem maksimuma križno-korelacijske funkcije med binarno kodo značke (zaporedje 1001101010000101) ter niza zajetih točk L_m.

S pomočjo poznanega linearnega položaja palice se nato osredotočimo na merjenje rotacije igralne palice. Na podlagi vrednosti točk v nizu točk L_m na mestih, ki ustrezajo posameznim bitom Grayeve kode (na sliki 3 označenimi z vertikalnimi modrimi črtami), določimo kodiran zapis kota. Prednost kodiranega zapisa v Grayevi kodi je v Hammingovi razdalji 1 med zaporednimi zapisi kotov – šum pri določanju vrednosti posameznih bitov v kodnem zapisu (te so lahko slabo določene na meji med svetlim in temnim poljem) je tako omejen le na najmanjši bit v končnem zapisu kota.

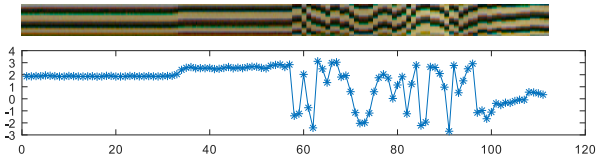


Slika 6: Območje za izračun faze

Z uporabo zapisa kota s 5-bitno Grayevo kodo lahko dosežemo ločljivost po kotu približno 11°, kar ne omogoča dovolj natančne igre. Ločljivost zato izboljšamo z uporabo interpolacije nad podatki zadnjega (najmanjšega) bita kodnega zapisa. Na zajeti sliki izberemo dodatne vzorce D_m v prečni smeri (vzdolž obsega značke na igralni palici, kot je prikazano na sliki 6) ter izračunamo fazo vzorca ϕ_m kot

$$\phi_m = \arctan \frac{\sum_{i=-N_p}^{N_p} D_m(i) \cdot \sin 2\pi i/\hat{T}}{\sum_{i=-N_p}^{N_p} D_m(i) \cdot \cos 2\pi i/\hat{T}}, \quad (2)$$

kjer je N_p število uporabljenih vzorcev, $\hat{T}$ pa pričakovana perioda kodnega zapisa zadnjega bita kode. Dobljena vrednost faze glede na zasuk palice je ilustrirana na sliki 7. Ker perioda kodnega zapisa zadnjega bita ustreza raz-



Slika 7: Časovni potek vzorca v sliki (**zgoraj**) z izračunano fazo (**spodaj**).

dalji 4 v zapisu rotacije oz. kotu $45°$, lahko zadnja dva bita v dekodiranem zapisu rotacije α_m zamenjamo z ustrezno transformirano vrednostjo faze ϕ_m.

1. ϕ_m zamaknemo za ϕ_{offset} glede na izhodišče Grayeve kode ter preskaliramo, da dosežemo periodo 4: $\phi'_m = \frac{4}{2\pi}\left[\phi_m - \phi_{offset}\right]$,
2. odstranimo 2 bita iz dekodiranega zapisa kota α_m s pomočjo binarne IN operacije z masko $b11100$), $\alpha'_m = \alpha_m$ IN $b11100$,
3. vrednost $j \in \mathbb{Z}$ določimo tako, da se vrednost izraza $\phi'_m + \alpha'_m + 4j - \alpha_m$ nahaja na intervalu $[-2, 2]$,
4. končno vrednost kota $\theta_m \in [0, 32) \cap \mathbb{R}$ izračunamo z $\theta_m = \phi'_m + \alpha'_m + 4j$.

Končna ločljivost, ki jo lahko dosežemo s predlagano metodo, je vezana na nivo šuma v zajetem signalu, ta pa je ocenjen s standardno deviacijo zajetih podatkov. Pri dobri osvetlitvi z nameščenim trakom LED smo tako dosegli natančnost $0{,}7°$ po kotu ter $0{,}6 - 0{,}8$ mm po translaciji palici (odvisno od položaja palice v sliki). Za isto ločljivost z uporabo Grayeve kode bi potrebovali kodni zapis z 9 biti. Ker trenutna postavitev kamere omogoča zanesljivo zaznavanje in dekodiranje le 6 bitov, smo izboljšali ločljivost meritve po kotu za faktor 8.

2.2 Detekcija žoge

Detekcija žoge je eden pomembnejših postopkov za zanesljivo delovanje avtomatiziranega sistema namiznega nogometa. Čeprav je v literaturi zaslediti različne postavitve sistema za zajem slike, večina avtorjev uporablja pristop sledenja žoge od spodaj. To zahteva namestitev steklene igralne površine, ki pa žal vpliva na samo igro (moteči odboji za igralce, drugačen koeficient trenja žogice ipd.). Naš pristop za sledenje uporablja par kamer, nameščenih nad igrišče, s čimer se zagotovi redundanca meritve in zmanjša težava s prekrivanjem žoge s strani igralcev. Zaradi hitrih premikov in pogostih prekrivanj ne uporabljamo algoritmov sledenja, temveč le optimizirano detekcijo na podlagi filtriranja po barvi v prostoru HSV (uporabljena je bila rumena igralna žoga), postopkov erozije in dilatacije (za odstranitev morebitnega šuma), nato pa iskanje in filtriranje po konturah. Položaj žoge je na koncu določen na podlagi slikovnih momentov izbrane konture.

Izmerjen položaj žoge se nato filtrira z namenom ugotavljanja vektorja hitrosti žoge (ocenjevanje smeri in hitrosti). Sistem za detekcijo žoge je še v testno-razvojni fazi, a že deluje z eno od kamer ter omogoča sledenje položaja žoge s frekvenco 100 Hz.

3 Pogonski sistem

Kot je omenjeno v uvodu, je pogonski sistem zasnovan modularno, posamezna pogonska enota je namenjena premikanju igralne palice po linearni in rotacijski osi. Pogonske enote so pritrjene na aluminijasto ogrodje, ki je bilo nemoteče nameščeno na igralno mizo, sama igralna palica pa je na pogonsko enoto pritrjena preko adapterja, ki smo ga natisnili s 3-D tiskalnikom (slika 8). Adapter objame ročaj palice ter omogoča hitro pritrditev ali snemanje pogonske enote.



Slika 8: Pogonska enota in kamere z osvetlitvijo

3.1 Mehanska izvedba

Mehanska konstrukcija je sestavljena iz standardnih 45 mm industrijskih aluminijastih profilov ter plastičnih povezovalnih elementov, ki so bili natisnjeni na 3D tiskalniku. Linearen pomik je dosežen z uporabo igus jermenskega pogona, ki je nameščen navpično. Ker smo lahko pogonski servo motor namestili v navpični smeri, je izvedba pogonske enote kompaktna ter dovoljuje namestitev dveh pogonskih enot tudi na igralne palice, ki so narazen le 150 mm.

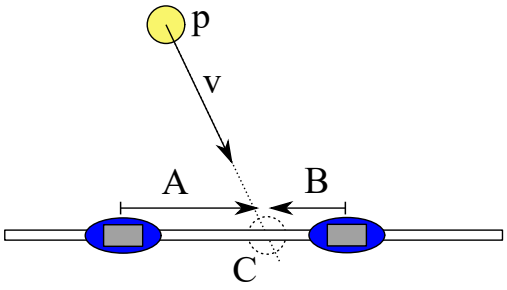
Pogonski servo motor za rotacijsko os je preko parkljaste sklopke in potisnega ležaja direktno sklopljen z adapterjem igralne palice ter omogoča prenos sile v aksialni smeri ter navora.

3.2 Sistem vodenja

Ker je treba v realnem času voditi 8 motorjev v položajnem načinu (4 pogonske enote, vsaka z linearno in rotacijsko osjo), smo izbrali krmilnik PoKeys57CNC. PoKeys57CNC omogoča krmiljenje motorjev s signali *korak/smer* (angl. *step/direction*) po trapeznem hitrostnem profilu, ki ga lahko sproti prilagajamo (spreminjanje položajne reference in hitrosti). Preko vodila USB ali Ethernet je krmilnik povezan na osebni računalnik, kjer se izvajajo analiza slike in algoritmi vodenja.

3.3 Algoritem branilca

Za namene testiranja in razvoja smo pogonsko enoto namestili na drugo igralno palico (obrambni igralec) ter implementirali enostaven algoritem branilca. Iz podatkov o trenutnem položaju in vektorju hitrosti žoge najprej ugotovimo presečišče med potjo žoge in linijo obrambnih figur (slika 9), ki predstavlja želeni položaj obrambne figure (označen s C). Nato moramo na podlagi trenutnih položajev in omejitev pri premikanju izbrati s katero obrambno figuro bomo igrali (premika označena z A in B). Pri izbiri najprej upoštevamo omejitve ter izločimo figure, s katerimi ciljnega položaja ne moremo doseči, nato pa izberemo figuro, ki je že trenutno najbližje želenemu položaju.



Slika 9: Pokrivanje predvidenega položaja žoge

Algoritem ima nato dve dodatni stanji, t.j. umik in napad. Če žogo zaznamo za linijo obrambnih figur, lete umaknemo za bolj učinkovit napad iz zadnjih vrst. V primeru, da žogo z obrambno figuro ustavimo in ju poravnamo, izvedemo napad, ki je sestavljen iz zamaha in udarca. Za dosego želene hitrosti figure ob udarcu ob žogo, moramo figure najprej umakniti, nato pa jo pospeševati po trapeznem profilu proti žogi. S kompaktnimi servo motorji za pogon rotacijske osi tako dosežemo hitrosti žoge, ki tudi presegajo hitrosti človeških igralcev (izmerjene hitrosti žoge okoli 10 m/s). V prihodnjih nadgradnjah bomo algoritemski pristop poskusili nadomestiti z nevronskimi mrežami[8].

4 Varnostni sistem

Zaradi kopice računalniško vodenih premičnih delov v področju igralne površine smo se za zagotavljanje varnosti pri uporabi naprave odločili za namestitev varnostne zavese Sick miniTwin4 nad igralno površino. V primeru poseganja v igralno področje ali preko pritiska na gumb za zaustavitev sistema preko kontaktorjev za enosmerno napetost izklopimo napajanje servo motorjev. Logika varnostnega krmiljenja je implementirana v programirljivem Sick Flexisoft varnostnem releju.

Dodatne varnostne ukrepe smo upoštevali tudi pri razvoju pogonskih enot, ki bodo zaprte s prozornim pokrovom, ter pri izbiri same mize za namizni nogomet. Izbrana miza ima teleskopske igralne palice, ki ne posegajo v prostor nasprotnikov.

5 Zaključek

Čeprav je predstavljen projekt še v razvojni fazi, prejemamo predvsem pozitivne odzive testnih uporabnikov. Zavedamo pa se, da je sama izvedba senzorskega in pogonskega sistema le majhen del na poti do naprave, ki bo omogočala zanimivo igro tudi bolj veščim igralcem namiznega nogometa. Pri izbiri samih komponent smo stremeli k izenačevanju zmogljivosti človeškega igralca, vendar bo na končni rezultat vplivalo predvsem pravilno delovanje implementiranih algoritmov.

6 Zahvala

Zahvaljujemo se Fakulteti za elektrotehniko, Univerze v Ljubljani, za financiranje projekta v sklopu internih projektov FE za izboljšanje kakovosti pedagoškega procesa in promocije – IPP&P 2019. Zahvaljujemo se tudi industrijskim 'partnerjem', ki so podprli izvedbo projekta:
Inea RBT d.o.o. (alu konstrukcija)
igus (Hennlich d.o.o.) (jermenski pogon)
PoLabs d.o.o. (krmilnik PoKeys57CNC)
IKU d.o.o. (DC kontaktor)
SICK d.o.o. (varnostna zavesa, varnostni rele)
FOBMARK d.o.o. (LED osvetlitev)

Literatura

[1] A. Zdešar, S. Blažic, and G. Klančar, "Engineering Education in Wheeled Mobile Robotics," *IFAC-PapersOnLine*, vol. 50, pp. 12173–12178, jul 2017.

[2] E. Guenat, C. Picard, R. Serigado, and B. Ulrich, "Automatic Foosball - Concurrent engineering project," tech. rep., 2012.

[3] R. Janssen, J. De Best, and R. Van De Molengraft, "Real-time ball tracking in a semi-automated foosball table," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5949 LNAI, pp. 128–139, Springer, Berlin, Heidelberg, 2010.

[4] P. Riden, "It's only a game: Robots defeat humans on foosball playing field - New Atlas - Robotics, https://newatlas.com/epfl-robot-table-soccer-foosball/44863/ [Accessed 2020-06-15]," 2016.

[5] M. Bošnak and G. Klančar, "Fast and Reliable Alternative to Encoder-Based Measurements of Multiple 2-DOF Rotary-Linear Transformable Objects Using a Network of Image Sensors with Application to Table Football," *Sensors*, vol. 20, p. 3552, jun 2020.

[6] A. Zdešar, I. Škrjanc, and G. Klančar, "Visual Trajectory-Tracking Model-Based Control for Mobile Robots," *International Journal of Advanced Robotic Systems*, vol. 10, p. 323, sep 2013.

[7] G. Klančar, M. Kristan, and R. Karba, "Wide-angle camera distortions and non-uniform illumination in mobile robot tracking," *Robotics and Autonomous Systems*, vol. 46, pp. 125–133, feb 2004.

[8] G. Cerne and I. Skrjanc, "Initial study on evolving state space neural networks (eSSNN)," in *2018 IEEE International Conference on Evolving and Adaptive Intelligent Systems, EAIS 2018*, pp. 1–8, Institute of Electrical and Electronics Engineers Inc., jun 2018.

Izdelava laboratorijskega demonstracijskega sistema (Mitsubishi)

Matej Milavec[1], **Goran Andonovski**[2]

[2]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: [1]*mm5723@student.uni-lj.si,* [2]*goran.andonovski@fe.uni-lj.si*

Production of a laboratory demonstration system (Mitsubishi)

The laboratory demonstration system is designed for presenting the basic operation of programmable logic controllers (PLCs) and other elements of automation in a transparent and practical way. The purpose of this paper is to present the individual components used and the entire process of building the system and moreover, to present a practical usage in a pedagogical/educational manner. We addressed the problem by first finding the necessary components to connected them to each other. Finally, we tested the operation by preparing a test programs on the PLC, on the touch screen display and we tested the functionality of the frequency converter. In conclusion, we have summarized the findings, which mostly relate to the possibility of further work on the system.

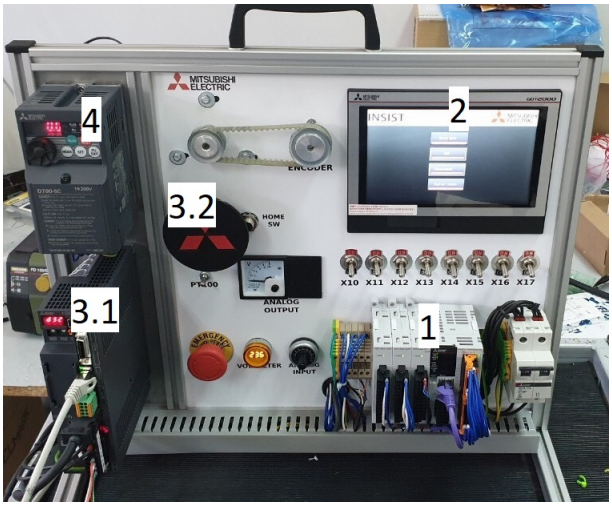
1 Uvod

Osnovna ideja tega članka je predstavitev izdelave laboratorijskega demonstracijskega sistema, ki temelji na opremi proizvajalca Mitsubishi. Poleg osnovnega PLK-ja (Programirljivi Logični Krmilnik) in pripadajočih modulov so v sistem vključeni tudi motorji, frekvenčni pretvornik, stikala, zaslon na dotik, itn. Z uporabo tega sistema imamo možnost predstavitve delovanja PLK-jev in ostalih elementov avtomatizacije na praktičen način.

Izdelava sistema je predstavljena kronološko, po posameznih sklopih. Najprej je bilo potrebno pregledati elektro načrt in naročiti vse potrebne elemente. Sledila je razporeditev teh elementov po osnovni plošči, izdelava izvrtin in montaža elementov na ploščo. Predzadnji korak je bila vezava elementov med seboj, kot je to bilo prikazano na električni shemi. Končni, časovno najbolj zahteven korak je bilo testiranje delovanja ter razhroščevanje, kar je vključevalo tudi spoznavanje z zbirko programske opreme Melsoft, proizvajalca Mitsubishi. Nadaljni koraki pa obsegjo predstavitev nekaj tipičnih nalog avtomatizacije za študente, ki bodo del učnega procesa.

2 Opis osnovnih gradnikov sistema

V tem poglavju bomo predstavili posamezne komponente in gradniki, ki so del laboratorijskega demonstracijskega sitema. Vsi gradniki so tipično del vsakega avtomatiziranega procesa. Izdelava našega sistema temelji na obstoječem sistemu, ki so ga izdelali na univerzi v Novem Sadu. Njihov sistem je predstavljen na sliki 1. Okviren izgled sistema smo ohranili, določene elemente smo nadomestili s primernejšimi ter nekaj jih odstranili.



Slika 1: Sistem, izdelan na univerzi v Novem Sadu.

V demonstracijskemu sistemu so uporabljeni osnovni elementi avtomatizacije, ki bodo predstavljeni v nadaljevanju[1]:

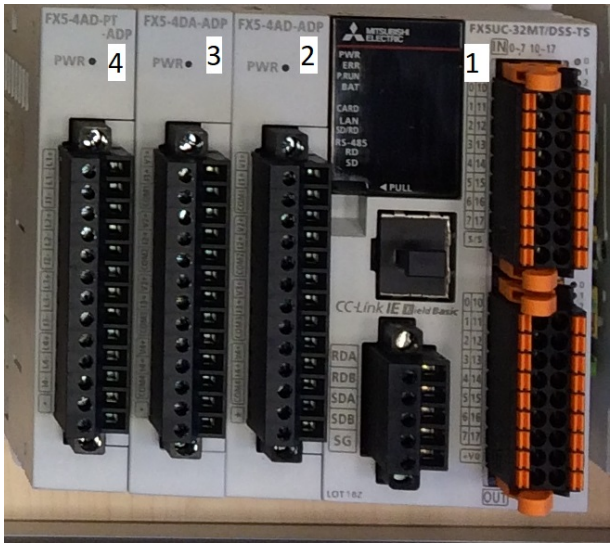
- PLK s pripadajočimi moduli (slika 1, številka 1);
- HMI[2] zaslon na dotik (slika 1, številka 2);
- servo ojačevalnik (slika 1, številka 3.1);
- servo motor (slika 1, številka 3.2); in
- frekvenčni pretvornik (slika 1, številka 4).

2.1 PLK

Programirljivi logični krmilnik (PLK) predstavlja osredni in povezovalni del vsakega sistema avtomatizacije. V našem primeru (slika 2) je sestavljen iz centralne procesne enote (ang. *CPU-central processing unit*) z digitalnimi vhodi in izhodi ter treh analognih modulov.

[1]Na tem mestu je potrebno poudariti, da smo te komponente dobili, kot delno sponzorka sredstva našega industrijskega partnerja Inea Rbt d.o.o. Več v zahvali.

[2]ang. Human Machine Interface



Slika 2: Prikaz uporabljenega PLK-ja.

Centralna procesna enota [1], na sliki 2 označena s številko 1, ima 16 tranzistorskih digitalnih vhodov in 16 tranzistorskih digitalnih izhodov. Uporabili smo CPU s serijsko oznako FX5UC-32MT/DSS-TS.

Prvi analogni modul, na sliki 2 označen s številko 2, je vhodni modul [2]. Namenjen je priklopu največ štirih temperaturnih sond, ki so lahko tipa Pt100 ali Ni100. Uporabili smo analogni modul s serijsko oznako FX5-4AD-PT-ADP.

Drugi analogni modul, na sliki 2 označen s številko 3, je izhodni modul, s štirimi kanali [2]. Vsak kanal omogoča štiri standardna območja napetostnih izhodov in dve standardni območji tokovnih izhodov. Uporabili smo analogni modul s serijsko oznako FX5-4DA-ADP.

Tretji analogni modul, na sliki 2 označen s številko 4, je vhodni modul, s štirimi kanali [2]. Vsak kanal omogoča štiri standardna območja napetostnih vhodov in dve standardni območji tokovnih vhodov. Uporabili smo analogni modul s serijsko oznako FX5-4AD-ADP.

2.2 Zaslon na dotik - HMI

V avtomatizaciji se zasloni na dotik uporablja kot vmesnik med strojem oziroma med napravo in človekom. Preko zaslona operater upravlja z napravo te služi za prikaz trenutne vrednosti vhodov/izhodov PLK-ja (npr. temperaturo okolice, ki jo merimo s Pt100 sondo priklopljeno na analogno-vhodni modul). V našem primeru ima zaslon premer 7" in resolucijo 800x460 točk [3]. Na zaslonu prav tako lahko nastavljamo posamezne vrednosti izhodov PLK-ja. Uporabili smo zaslon na dotik s serijsko oznako GT2107-W.

2.3 Servo ojačevalnik s servo motorjem

Servomotorji se uporabljajo v veliko industrijskih aplikacijah. Največja prednost napram ostalim vrstam motorjev je velik pospešek, zaradi manjšega vztrajnostnega momenta. Servo motor ima vgrajen enkoder, ki omogoča merjenje trenutnega zasuka, kar v praktični uporabi pomeni možnost natančnega pozicioniranja. Servo motor krmilimo s servo ojačevalnikom. Glede na želene odzive je mogoče uporabiti več načinov krmiljenja:

- ohranjanje željene hitrosti,
- ohranjanje željenega navora,
- zasuk za željen kot,
- itn.

V našem primeru na servo ojačevalnik je možno priklopiti servomotor z maksimalno močjo 100W. Uporabili smo servo ojačevalnik s serijsko oznako MR-JE-1C-RJ ter servomotor s serijsko oznako HG-KN13J.

2.4 Frekvenčni pretvornik

Motorji, ki delujejo na izmenično napetost (angl. *alternating current – AC*) so najbolj razširjena vrsta motorjev. V industriji se uporabljajo v veliko različnih aplikacijah, na primer za pogon: mešal, črpalk, tekočih trakov itd. Z uporabo frekvenčnih pretvornikov se uporabnost AC motorjev še poveča, saj ti omogočajo krmiljenje hitrosti vrtenja motorja. S frekvenčnim pretvornikom krmilimo hitrost vrtenja motorja. Na frekvenčni pretvornik je možno priklopiti motorje z maksimalno močjo 200W. Uporabili smo frekvenčni pretvornik s serijsko oznako FR-D720S-014 SC-ES.

3 Izdelava

Pred začetkom izdelave je bilo potrebno naročiti vse komponente. Na podlagi načrta električnih povezav in slik sistema iz Novega Sada (slika 1) smo izdelali seznam potrebnih oziroma manjkajočih komponent. Za vsako komponento smo vpisali oznako iz električne sheme, serijsko oznako, količino in kratek opis (pri elementih kjer je to bilo potrebno). Potrebne elemente smo poiskali na naslednjih petih spletnih straneh:

- www.meanwell.si,
- www.si.farnell.com,
- www.eu.mouser.com,
- www.at.rs-online.com, in
- www.ic-elect.si.

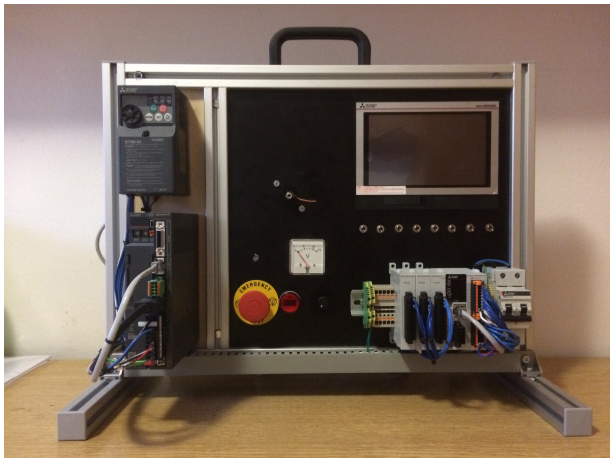
V seznam komponent smo dodali spletne povezave do izbranih elementov in njihove cene. Dodali smo tudi seštevek povprečnih cen v različnih trgovinah, s čimer smo ugotovili predvideno skupno ceno komponent. Po naročilu in prejemu vseh komponent je sledilo sestavljanje sistema.

3.1 Sestava sistema

Komponente smo najprej razporedili po površini, kot je to prikazano na sliki 3. Pri tem smo bili pozorni na to, da imajo komponente, s katerimi upravljamo z rokami dovolj prostora v svoji okolici. Razporeditev je olajšala naslednji korak, ki je bil določitev mest za izvrtine in izreze.



Slika 3: Razporeditev komponent.



Slika 4: Sestavljen sistem - spredaj.

Nato smo naredili izvrtine z vrtalnim strojem ter izreze z vbodno žago. Za lepši izgled končnega izdelka smo okolice odprtin pobrusili z brusnim papirjem. Ploščo smo vstavili v sestavljiv okvir iz aluminijastih profilov. Sledila je pritrditev komponent. Večina manjših komponent se na ploščo pritrdi z matico na zadnji strani plošče. Zaslon se vstavi v odprtino in privijači iz zadnje strani. Večje komponente (frekvenčni pretvornik, servo ojačevalnik in servo motor) so pritrjene z vijaki skozi ploščo.

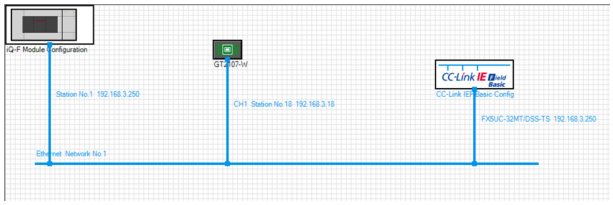
Na obe strani plošče smo dodali DIN letvi, na kateri smo pritrdili še ostale elemente, kot so vrstične sponke, PLK, napajalnik in inštalacijski odklopnik. Na spodnji del okvirja smo dodali kabelski kanal, v katerem bodo skrite žice.

3.2 Vezava

Vezave elementov smo se lotili tako, da smo najprej povezali visoko-napetostni del (260 V izmenična napetost) in nato nizko-napetostni del (24 V enosmerna napetost). Za visoko-napetostni del smo uporabili žice preseka $1,5$ mm^2 (črno za fazni vodnik, modro za ničelni vodnik). Za nizko-napetostni del pa smo uporabili žico preseka $0,75$ mm^2 (modro za vodnik 24 V in belo za vodnik 0 V). Ker smo uporabili mnogo žične vodnike, smo konce žic zaščitili z votlicami. To smo storili povsod, kjer je bila žica na element pritrjena z vijakom. Na nekaj elementov (stikala, V-meter, ...) smo žice prilotali. Tako povezan sistem je prikazan na sliki 4.

4 Preizkus delovanja

Za potrebe testiranja smo v programskem okolju MELSOFT Navigator sestavili konfiguracijo omrežja, prikazano na sliki 5. V omrežje so vključeni: PLK, HMI in servo ojačevalnik. Preko USB kabla smo na zaslon HMI naložili program, za osnovno delovanje zaslona, vključno z delom programa za ethernet komunikacijo. Po prvem prenosu ne potrebujemo več USB kabla, saj se ostali podatki prenašajo preko ethernet kabla. Na PLK-ju in servo ojačevalniku so osnovni programi že naloženi, zaradi česar je mogoča takojšnja komunikacija med komponentami preko ethernet omrežja.



Slika 5: Konfiguracija omrežja.

S programom MELSOFT GX Works 3 smo programirali PLK. Najprej smo nastavili globalne oznake tipa bit (digitalni vhodi/izhodi). Definirali smo imena 16-ih digitalnih izhodov Y0 do Y7 in Y10 do Y17 ter 16-ih digitalnih vhodov X0 do X7 in X10 do X17.

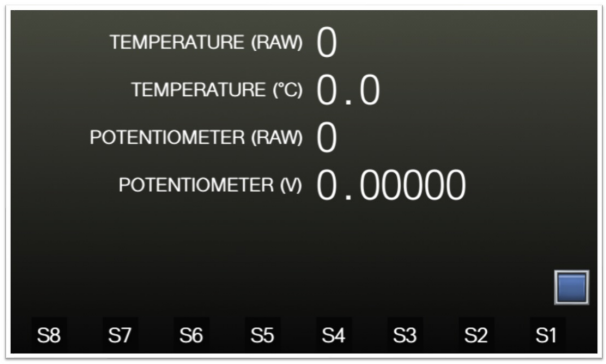
Nastavili smo še globalne oznake tipa double word (analogni vhodi/izhodi), pri tem smo si pomagali z uporabniškimi navodili[4]. Številka vhoda/izhoda je odvisna od tega kateri po vrsti je zaporedni modul. V našem primeru smo nastavili imena in oznake opisane v tabeli 1.

Tabela 1: Osnovni slogi pri oblikovanju prispevka.

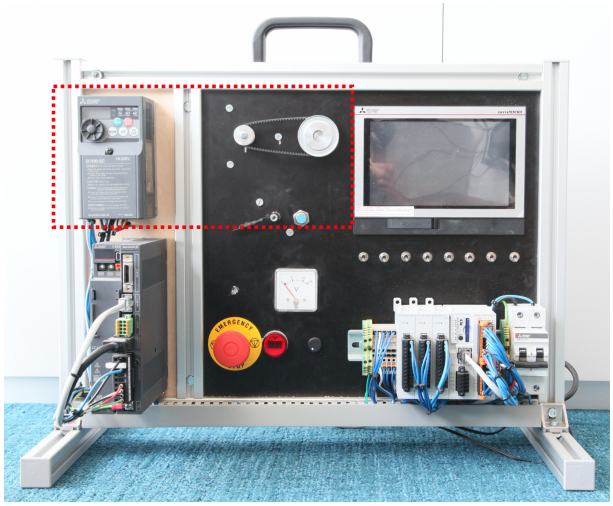
Ime	**Oznaka**
Analogni vhod za Pt100 sondo	SD6300
Analogni izhod za V-meter	SD6660
Analogni izhod za frekvenčni pretvornik	SD6700
Analogni vhod za potenciometer	SD7020

Za preizkus delovanja smo v programu MELSOFT GT Designer 3 izdelali testni prikaz (slika 6). Prve dve vrednosti na zaslonu (od zgoraj navzdol) se nanašata na analogni vhod za Pt100 sondo (SD6300). Prvi prikaz je vrednost prepisana neposredno z naslova SD6300. Druga vrednost je prejšnja deljena z 10, s čimer dobimo dejansko vrednost temperature okolice v stopinjah celzija. Drugi dve vrednosti se nanašata na analogni vhod za potenciometer (SD7020). Prvi prikaz je vrednost prepisana neposredno z naslova SD7020. Druga vrednost je prejšnja deljena z 10^8, s čimer dobimo dejansko vrednost napetosti na priklopu potenciometra.

V programu smo vrednost z analognega vhoda za po-



Slika 6: Testni prikaz zaslona na dotik HMI.



Slika 7: Končna verzija sistema.

tenciometer prepisali na analogni izhod za V-meter. Tako je V-meter kazal napetost, ki je bila prikazana na zaslonu in je odražala dejansko vrednost napetosti na priklopu potenciometra.

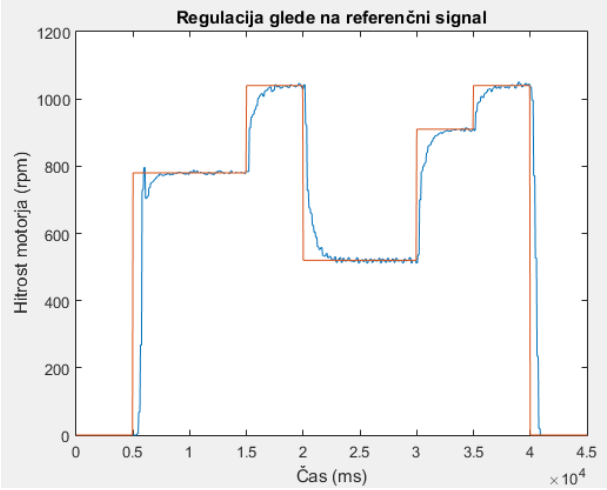
Na dnu zaslona smo dodali osem prikazov od *S1* do *S8*. Kvadrati okoli napisov (*S1* do *S8*) so črne barve, dokler so fizična stikala pod zaslonom izklopljena in se obarvajo zeleno, ko stikala aktiviramo.

5 Vodenje *AC*-motorja

V tem poglavju bomo predstavili vodenje *AC*-motorja s frekvenčnim pretvornikom. Frekvenčni pretvornik krmilimo z ukazi preko PLK-ja, dokler hitrost vrtenja motorja dobimo preko enkoderja, ki je povezan na motor preko zobatega jermena (slika 7).

Za namene vodenja smo najprej nastavili parametre frekvenčnega pretvornika ter smo pripravili program na PLK-ju za realizacijo PID-regulatorja. Poleg tega smo pripravili tudi novo okno na HMI zaslonu. Preko novega okna na zaslonu lahko nastavljamo parametre PID-regulatorja ter izvajamo eksperimente.

Najprej smo izvedli pomerili statično karakteristiko sistema tako, da smo stopničasto povečevali frekvenco na izhodu frekvenčnega pretvornika (vhodna veličina) in opazujemo hitrost motorja (izhodna veličina). Izkazalo



Slika 8: Regulacija hitrosti vrtenja *AC*-motorja.

se je da je sistem precej linearen in smo nato določili še model v delovni točki. Parametre regulatorja smo najprej izračunali s pomočjo metode Zeigler-Nichols in potem smo jih še ročno spreminjali. Dobljen odziv okoli izbrane reference je prikazan na sliki 8.

6 Zaključek

Predstavljen sistem zaradi svojo kompaktnost in prenosljivost nam lahko služi kot platforma za demonstracijo osnovnih gradnikov, ki se uporabljajo na področju avtomatizacije. Poleg tega na enostaven način lahko predstavimo tudi delovanje regulacijskih algoritmov, kot na primer delovanje in lastnosti PID-regualtorja.

Nadaljnji koraki so izdelava učnega gradiva z različnimi nalogami za spoznavanje z vsako komponento posebej, kot tudi izvajanje eksperimentov na celotnem sistemu.

Zahvala

Zahvaljujemo se industrijskemu partnerju INEA RBT d.o.o. za podporo pri izdelavi projekta.

Literatura

[1] „CPU module MELSEC iQ-F series — MITSUBISHI ELECTRIC FA“. [Na spletu]. Dostopno na: https://www.mitsubishielectric.com/fa/products/cnt/plcf/pmerit/cpu/index.html. [Pridobljeno: 20-feb-2020].

[2] „Analog Control module MELSEC iQ-F series — MITSUBISHI ELECTRIC FA“. [Na spletu]. Dostopno na: https://www.mitsubishielectric.com/fa/products/cnt/plcf/pmerit/analog/index.html. [Pridobljeno: 20-feb-2020].

[3] „GT21 Wide model GOT2000 Series Human-Machine Interfaces(HMIs)-GOT— MITSUBISHI ELECTRIC FA“. [Na spletu]. Dostopno na: https://www.mitsubishielectric.com/fa/products/hmi/got/pmerit/got2000/wide/gt21wide.html. [Pridobljeno: 20-feb-2020].

[4] „MELSEC iQ-F FX5 User's Manual (Analog Control - CPU module built-in, Expansion adapter)“[Pridobljeno: 21-feb-2020].

Trajectory Planning By Applying Optimal Velocity Profile Algorithm on Bernstein-Bézier Motion Primitives

Martina Loknar, Gregor Klančar, Sašo Blažič

Faculty of Electrical Engineering, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
E-mail: martina.loknar@fe.uni-lj.si

Abstract

This paper deals with minimal-time smooth trajectory planning for autonomous mobile systems. The resulting path is defined piecewise by multiple Bernstein-Bézier motion primitives that enable continuous velocity and curvature at the junctions. The solution is found by calculating travel time along Bernstein-Bézier segments with some free parameters by applying the optimal velocity profile algorithm that considers velocity, acceleration and jerk constraints. The proposed optimization approach was validated in a simulation environment.

1 Introduction

Path planning algorithms generate a geometric path from an initial to a final point through pre-defined via-points, either in the joint space or in the operating space of the robot. On the other hand, trajectory planning algorithms assign a time law to the geometric path. There is an increasing demand for robots and automatic machines to operate while some function is optimized and quite often the research is focused in time optimization. Path planning and trajectory planning algorithms are therefore attracting considerable interest and remain the crucial issues in the field of automation and robotics [1].

Defining the time of via-points progression influences the kinematic and the dynamic properties of the motion: the inertial forces and torques depend on the accelerations along the trajectory, while the vibrations of its mechanical structure are mostly determined by the values of the jerk. In order to satisfy kinematic feasibility of the vehicle the resulting path should be smooth; it should be realizable at high speed and at the same time be harmless for the robot in terms of avoiding excessive accelerations of the actuators and vibrations of its mechanical system [1]. Path smoothing is often not integrated in path planning but is usually done after the optimal path is found, which requires additional collision checks and can influence the path optimality. One promising approach would therefore be finding a smooth path to combine motion primitives in a path planning phase [2]. The smoothing techniques, not only limited to sample based planners, rely on using curves to interpolate or fit the given waypoints.

The first studies to obtain the shortest curvature constrained smooth paths consisting of straight lines and circular arcs were performed by [3], but the resulting tracks had a discontinuous curvature. This is why the authors in [4] used Dubins paths and smoothed them with clothoid arcs. The main advantage of clothoids is the linear change of their curvature. Unfortunately clothoids are defined in terms of Fresnel integrals; transcendental functions that cannot be solved analytically. This fact makes clothoids difficult to use in real-time applications so that for real-time motion planning many authors resort to curves with nonlinear curvature that are easier to compute, for example Bézier curves [2]. Other smoothing methods also include application of cubic, quintic or higher order polynomials [5] and B-splines [6].

This paper presents the results of a study where the optimal velocity algorithm was validated on fifth-order Bernstein-Bézier curve segments with continuous velocity and curvature transitions. The optimal velocity algorithm considers restrictions of speed, radial and tangential acceleration and radial and tangential jerk. The purpose of the analysis was to determine the travel time along Bernstein-Bézier motion primitives in order to find a time optimal corridor-restricted path with the minimum travel time and to provide a better insight into possible trajectory planning strategies.

2 Problem statement

Consider the problem of a ground vehicle with a mission defined by corridor and dynamical constraints in a two-dimensional free space. Our goal is to develop and implement an algorithm for generating a trajectory that satisfies these restrictions.

Let the motion of a particle along a three times continuously differentiable plane curve $\mathcal{C}$ be described as a function of time $t \in [0, t_f]$ by the position vector $\mathbf{r}(t)$ measured from a given fixed origin. Velocity $\mathbf{v}(t)$, acceleration $\mathbf{a}(t)$ and jerk $\mathbf{j}(t)$ vectors can be expressed in the tangential-normal form as:

$$\mathbf{v}(t) = v(t) \cdot \hat{\mathbf{T}} \tag{1a}$$

$$\mathbf{a}(t) = a_{\mathrm{T}}(t) \cdot \hat{\mathbf{T}} + a_{\mathrm{R}}(t) \cdot \hat{\mathbf{N}} \tag{1b}$$

$$\mathbf{j}(t) = j_{\mathrm{T}}(t) \cdot \hat{\mathbf{T}} + j_{\mathrm{R}}(t) \cdot \hat{\mathbf{N}}, \tag{1c}$$

where $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ are the unit tangent vector and the unit normal vector, respectively:

$$\hat{\mathbf{T}}(t) = \frac{\mathbf{v}(t)}{\|\mathbf{v}(t)\|}, \quad \hat{\mathbf{N}}(t) = \frac{\dot{\hat{\mathbf{T}}}(t)}{\|\dot{\hat{\mathbf{T}}}(t)\|}. \tag{2}$$

Given a feasible segment of the path, which in our case is a Bernstein-Bézier curve, the optimization problem is to find the velocity profile $v(t)$ that reaches the end of the curve in minimum time in a way that none of the velocity, acceleration or jerk constraints from Eqs. $(3a, 3b, 3c)$ are violated:

$$0 \leq \|\mathbf{v}(t)\| \leq v_{\mathrm{MAX}}; \quad \forall t \in [0, t_f], \tag{3a}$$

$$\frac{a_T^2(t)}{a_{T_{\mathrm{MAX}}}^2} + \frac{a_R^2(t)}{a_{R_{\mathrm{MAX}}}^2} \leq 1; \quad \forall t \in [0, t_f], \tag{3b}$$

$$\frac{j_T^2(t)}{j_{T_{\mathrm{MAX}}}^2} + \frac{j_R^2(t)}{j_{R_{\mathrm{MAX}}}^2} \leq 1; \quad \forall t \in [0, t_f]. \tag{3c}$$

The acceleration and jerk constraints are defined in a similar way as in [7].

3 Bernstein-Bézier motion primitives

An N-dimensional, n-th order Bernstein polynomial $\mathbf{r}_n(\lambda) : [0,1] \to \mathbb{R}^N$ can be defined as:

$$\mathbf{r}_n(\lambda) = \sum_{i=0}^{n} \mathbf{P}_{i,n} B_{i,n}(\lambda), \qquad \lambda \in [0,1], \tag{4}$$

where λ is normalized time ($0 \leq \lambda \leq 1$), $\mathbf{P}_{i,n} \in \mathbb{R}^N$ is the i-th control point and $B_{i,n}(\lambda)$ is the Bernstein polynomial basis defined as:

$$B_{i,n}(\lambda) = \binom{n}{i} \lambda^i (1-\lambda)^{n-i}, \tag{5}$$

for all $i \in \{0, \ldots, n\}$. Binomial coefficient is defined as:

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}. \tag{6}$$

Letting $\mathbf{P}_n = [\mathbf{P}_{0,n}, \ldots, \mathbf{P}_{n,n}] \in \mathbb{R}^{N \times (n+1)}$ be the vector of control points of $\mathbf{r}_n(\lambda)$, the Bernstein polynomial in Eq. (4) can be rewritten in matrix form as:

$$\mathbf{r}_n(\lambda) = \mathbf{P}_n \begin{bmatrix} B_{0,n}(\lambda) \\ B_{1,n}(\lambda) \\ \cdots \\ B_{n,n}(\lambda) \end{bmatrix}. \tag{7}$$

In the case when two or three-dimensional Bernstein polynomials are used to describe planar and spatial curves, Bernstein polynomials are often referred to as Bézier curves. These curves have useful path planning properties. The first and the last points of the Bernstein polynomial introduced in Eq. (4) are its endpoints:

$$\mathbf{r}_n(0) = \mathbf{P}_{0,n} \quad \text{and} \quad \mathbf{r}_n(1) = \mathbf{P}_{n,n}. \tag{8}$$

The N-dimensional, n-th order Bernstein polynomial also lies within the convex hull defined by its control points. Furthermore, the start and the end of the curve is tangent to the first and the last section of the convex polygon, respectively (Fig. 1).

$$\left.\frac{\mathrm{d}\mathbf{r}_n}{\mathrm{d}\lambda}\right|_{\lambda=0} = n(\mathbf{P}_{1,n} - \mathbf{P}_{0,n}), \tag{9}$$

$$\left.\frac{\mathrm{d}\mathbf{r}_n}{\mathrm{d}\lambda}\right|_{\lambda=1} = n(\mathbf{P}_{n,n} - \mathbf{P}_{n-1,n}). \tag{10}$$

Other properties of Bernstein polynomials (derivatives, calculating definite integrals, the de Casteljau's algorithm, degree elevation etc.) do not fall within the scope of this article; more details on this topic can be found in [8].

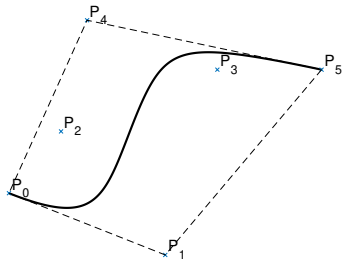


Figure 1: Fifth order Bernstein-Bézier curve with its convex hull (dotted lines). The curve is tangent to the sides of the convex hull, line segments $\overline{P_0P_1}$ and $\overline{P_4P_5}$.

3.1 Merging of motion primitives

For the sake of generality, we consider a merging of two Bézier curves of different orders, $\mathbf{r}_{n_j}^j(\lambda)$ of order n_j and $\mathbf{r}_{n_{j+1}}^{j+1}(\lambda)$ of order n_{j+1}, into a spline. The three conditions for continuous first and second derivatives (C^2 continuity) in the junction are:

$$\lim_{\lambda \to 1} \mathbf{r}_{n_j}^j(\lambda) = \lim_{\lambda \to 0} \mathbf{r}_{n_{j+1}}^{j+1}(\lambda), \tag{11}$$

$$\lim_{\lambda \to 1} \frac{d\mathbf{r}_{n_j}^j(\lambda)}{d\lambda} = \lim_{\lambda \to 0} \frac{d\mathbf{r}_{n_{j+1}}^{j+1}(\lambda)}{d\lambda}, \tag{12}$$

$$\lim_{\lambda \to 1} \frac{d^2\mathbf{r}_{n_j}^j(\lambda)}{d\lambda^2} = \lim_{\lambda \to 0} \frac{d^2\mathbf{r}_{n_{j+1}}^{j+1}(\lambda)}{d\lambda^2}, \tag{13}$$

This yields the following relations between the control points of both Bézier curves:

$$\mathbf{P}_{0,n_{j+1}}^{j+1} = \mathbf{P}_{n_j,n_j}^j, \tag{14}$$

$$\mathbf{P}_{1,n_{j+1}}^{j+1} = (1 + \frac{n_j}{n_{j+1}})\mathbf{P}_{n_j,n_j}^j - \frac{n_j}{n_{j+1}}\mathbf{P}_{n_j-1,n_j}^j, \tag{15}$$

$$\begin{aligned} \mathbf{P}_{2,n_{j+1}}^{j+1} = &\left(1 + \frac{n_j}{n_{j+1}}\left(2 + \frac{1+n_j}{1+n_{j+1}}\right)\right)\mathbf{P}_{n_j,n_j}^j \\ &- 2\left(1 + \frac{n_j(1+n_j)}{n_{j+1}(1+n_{j+1})}\right)\mathbf{P}_{n_j-1,n_j}^j \\ &+ \frac{n_j(1+n_j)}{n_{j+1}(1+n_{j+1})}\mathbf{P}_{n_j-2,n_j}^j. \end{aligned} \tag{16}$$

3.2 Path generation

Bézier curves constructed by large numbers of control points are numerically unstable. For this reason, in path planning, it is desirable to construct a smooth way by joining together low degree Bézier curves.

We employed the curves of the fifth order, as this is the least degree of Bézier curves that can satisfy the requirement for curvature (C^2) continuity. The fifth-order Bernstein -Bézier curve $\mathbf{r}_5(\lambda) = [x(\lambda), y(\lambda)]^{\mathrm{T}}$ is defined by six control points $\mathbf{P}_{i,5} = [x_i, y_i]$, $i \in \{0, 1, \ldots 5\}$.

$$\begin{aligned} \mathbf{r}_5(\lambda) = &(1-\lambda)^5\mathbf{P}_{0,5} + 5\lambda(1-\lambda)^4\mathbf{P}_{1,5} \\ &+ 10\lambda^2(1-\lambda)^3\mathbf{P}_{2,5} + 10\lambda^3(1-\lambda)^2\mathbf{P}_{3,5} \\ &+ 5\lambda^4(1-\lambda)\mathbf{P}_{4,5} + \lambda^5\mathbf{P}_{5,5}. \end{aligned} \tag{17}$$

The complete path through the corridor consists of several Bernstein-Bézier curve sections. According to Eqs. (14-16) for $n_j = 5$ and $n_j = n_{j+1}$ the first three control points of the $(j+1)$-th Bernstein-Bézier curve $\mathbf{r}_5^{j+1}$ are:

$$\mathbf{P}_{0,5}^{j+1} = \mathbf{P}_{5,5}^{j}, \tag{18}$$

$$\mathbf{P}_{1,5}^{j+1} = 2\mathbf{P}_{5,5}^{j} - \mathbf{P}_{4,5}^{j}, \tag{19}$$

$$\mathbf{P}_{2,5}^{j+1} = 4\mathbf{P}_{5,5}^{j} - 4\mathbf{P}_{4,5}^{j} + \mathbf{P}_{3,5}^{j}. \tag{20}$$

where $P_{i,n}^{j}$ is the i-th control point of j-th curve $\mathbf{r}_5^j$. The last control point $\mathbf{P}_{5,5}^{j+1}$ is defined by the final position:

$$\mathbf{P}_{5,5}^{j+1} = \begin{bmatrix} x_5^{j+1} \\ y_5^{j+1} \end{bmatrix}. \tag{21}$$

The expressions for the control points $\mathbf{P}_{4,5}^{j+1}$ and $\mathbf{P}_{3,5}^{j+1}$ follow from evaluating the first and the second derivative of $\mathbf{r}_5^{j+1}$ in $\lambda = 1$ and setting its values to v_{spline} and a_{spline}, respectively:

$$\left.\frac{\mathrm{d}\mathbf{r}_5^{j+1}}{\mathrm{d}\lambda}\right|_{\lambda=1} = v_{\text{spline}}, \tag{22}$$

$$\left.\frac{\mathrm{d}^2\mathbf{r}_5^{j+1}}{\mathrm{d}\lambda^2}\right|_{\lambda=1} = a_{\text{spline}}. \tag{23}$$

$\mathbf{P}_{4,5}^{j+1}$ and $\mathbf{P}_{3,5}^{j+1}$ are defined by both the final position and the final orientation φ_5^{j+1}:

$$\mathbf{P}_{4,5}^{j+1} = \mathbf{P}_{5,5}^{j} - \tfrac{1}{5} v_{\text{spline}} \cdot \begin{bmatrix} \cos(\varphi_5^{j+1}) \\ \sin(\varphi_5^{j+1}) \end{bmatrix}, \tag{24}$$

$$\mathbf{P}_{3,5}^{j+1} = -\mathbf{P}_{5,5}^{j} + 2\mathbf{P}_{4,5}^{j} + \\ + \tfrac{1}{20} a_{\text{spline}} \cdot \begin{bmatrix} \cos(\varphi_5^{j+1}) \\ \sin(\varphi_5^{j+1}) \end{bmatrix}, \tag{25}$$

It should be pointed out that the values of the first and the second curve derivative in the $(j+1)$-th Bernstein-Bézier curve end point $\mathbf{P}_{5,5}^{j+1}$, v_{spline} and a_{spline}, are not true speed and tangential acceleration. The notation that we used simply reflects the analogy.

The flowchart in Fig.(2) describes the general principle of operation of the proposed trajectory planning algorithm. Corridor consists of m segments and j is the number of the current segment on which the calculation is performed: $j \in \{1, 2, \ldots m\}$. The given coordinates and values of velocity, angle and tangential acceleration in $\mathbf{P}_{0,5}^{j}$ according to Eqs.(19-20) also determine the second and the third control point of the Bézier curve, $\mathbf{P}_{1,5}^{j}$ and $\mathbf{P}_{2,5}^{j}$. The problem is to calculate the three remaining control points of the Bézier curve along which the travel time is the shortest. To find the solution, MATLAB's integrated function `fmincon` was used. This is a solver-based nonlinear optimization approach that finds a minimum of a constrained nonlinear multivariable function. The starting point of the optimization is travel time along a Bézier curve, given by setting the coordinates, velocity, angle and tangential acceleration in the final control point $\mathbf{P}_{5,5}^{j}$. The control points $\mathbf{P}_{3,5}^{j}$ and $\mathbf{P}_{4,5}^{j}$ are determined according to the Eqs.(24-25). The travel time t along this Bézier curve is calculated by applying the algorithm that generates the optimal velocity profile. The function `fmincon` then iteratively modifies the free parameters ($(x, y), v, \varphi$ and a_T) in $\mathbf{P}_{5,5}^{j}$ in a way that the travel time along the curve decreases until the local minimum is found (t_{MIN} in Fig.(2)). This is a basic concept of the gradient method: from the starting point the optimization function at each iteration takes either a direct (Newton) step or a conjugate gradient step to reach the local minimum of the cost function. The calculation is repeated for each segment to finally determine the complete trajectory $\mathbf{G}_1$ along the corridor.

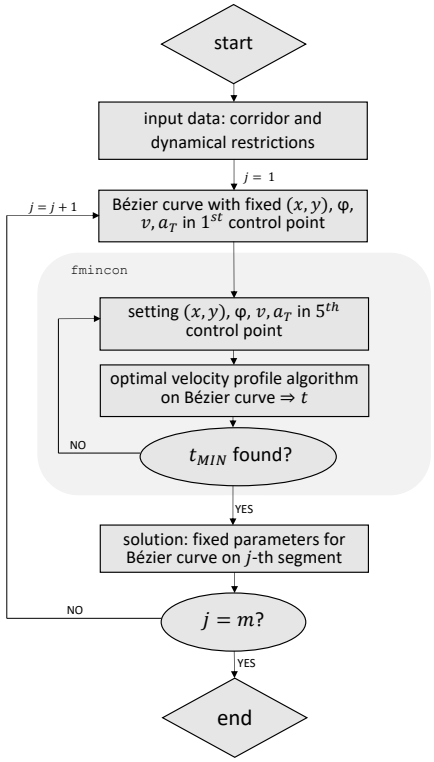


Figure 2: The basic principle of operation of the proposed trajectory planning algorithm.

In order to study the effectiveness of the proposed method, the time optimal trajectory planning algorithm is executed in two other versions. In the second version, minimal travel time $t_{\text{MIN},2seg}$ is calculated along Bézier curves for two successive corridor segments (for j and $j+1$) and in the third version $t_{\text{MIN},3seg}$ it is calculated for three successive corridor segments (for j, $j+1$ and $j+2$). Then, only the control points $\mathbf{P}_{i,5}^{j}$ ($i \in \{0, \ldots, n\}$) of the j-th segment are retained. This procedure is performed on the next segments until the the complete trajectory along the corridor is found: $\mathbf{G}_2$ in the second and $\mathbf{G}_3$ in the third version of the algorithm.

4 Simulation results

To demonstrate our trajectory planning method, we designed the problem as follows: compute the time optimal trajectory through the given 6-segment corridor with the following dynamical restrictions: $v_{\text{MAX}} = 1.5\,\text{m/s}$, $a_{\text{T}_{\text{MAX}}} = 2\,\text{m/s}^2$, $a_{\text{R}_{\text{MAX}}} = 4\,\text{m/s}^2$, $j_{\text{T}_{\text{MAX}}} = 6\,\text{m/s}^3$

and $j_{\mathrm{R_{MAX}}} = 8\,\mathrm{m/s^3}$. On the starting line, the values of velocity and tangential acceleration are $v = 1\,\mathrm{m/s}$ and $a_{\mathrm{T}} = 0.5\,\mathrm{m/s^2}$. The initial position is [0, 0] and the initial orientation is set parallel to the corridor center line.

Figure (3) shows the given six segment corridor and the optimal trajectories $\mathbf{G}_i$, depicted with a thick black line. The sets of discarded Bernstein-Bézier curves are depicted with thin grey lines. $\mathbf{G}_1$,$\mathbf{G}_2$, $\mathbf{G}_3$ are calculated based on knowledge of the shape of the current segment, two succesive segments and three successive segments, respectively.

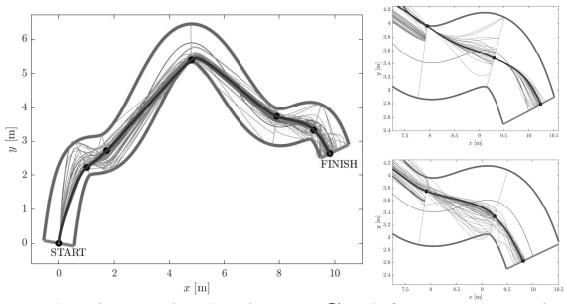


Figure 3: Time optimal trajectory $\mathbf{G}_3$ (left). For comparison: iteration steps in the last three segments for $\mathbf{G}_1$ (top right) and $\mathbf{G}_2$ (bottom right).

Table (1) lists the travel times t_i at the end of the separate corridor segments. The evidence from this study proves that the time optimal trajectory can be determined by our trajectory planning method. Furthermore, the more successive segments are taken into account when performing the calculation, the lower is the overall minimal travel time. Table (2) presents some computational properties of the proposed algorithm: number of iterations per segment $N_{\mathrm{it,seg}}$, number of points (per segment) `F-count` where the function evaluations took place, and the total computation time t_{COMP}. It should be pointed out that the problem has many constraints and the `F-count` can be significantly less than the total number of function evaluations.

$[s]$	t_1	t_2	t_3	t_4	t_5	t_6	$\sum_{i=1}^{6} t_i$
$\mathbf{G}_1$	1.79	0.80	2.93	2.48	0.96	0.85	9,81
$\mathbf{G}_2$	1,75	0,62	2,74	2,41	1,03	0,65	9,20
$\mathbf{G}_3$	1,73	0,62	2,75	2,43	1,01	0,64	9,18

Table 1: t_i and overall travel time $\sum_{i=1}^{6} t_i$ for $\mathbf{G}_i$.

	$N_{\mathrm{it,seg}}$	`F-count`	$t_{\mathrm{COMP}}[s]$
$\mathbf{G}_1$	25	250	246
$\mathbf{G}_2$	13	190	306
$\mathbf{G}_3$	19	305	738

Table 2: Computational properties of the proposed algorithm.

The current solutions to the time optimal trajectory generation problem under kinematic constraints (that include jerk limitations) exist mainly for robot manipulators, whereas the literature on the same topic in the case of autonomous vehicles is still scarce. A key problem with much of the literature is also that the approaches often do not support Cartesian constraints.

5 Conclusion

This paper presents a trajectory planning algorithm in a free space that is based on Bernstein-Bézier curves and considers corridor and dynamical (velocity, acceleration and jerk) constraints. Bézier curves provide an efficient way to generate the optimized trajectory and satisfy the constraints at the same time. The simulation results show that the minimal travel time through the corridor is the lowest in the case when three successive corridor segments are considered in the proposed method.

Despite some limitations, we believe that the results of this study could be the basis for our future research on similar receding horizon control methods to generate real-time optimal trajectories with dynamical restrictions. In the future, we will also consider a development of a more sophisticated way to determine whether the calculated path is within the corridor boundaries. Using higher order Bézier curves with more free parameters might also prove to be beneficial.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0219).

References

[1] A. Gasparetta, P. Boscario, A. Lanzutti, and R. Vidoni, *Motion and Operation Planning of Robot Systems, Chapter 1 - Path Planning and Trajectory Planning Algorithms: A General Overview*, vol. 29. 2015.

[2] G. Klančar, M. Seder, S. Blažič, I. Škrjanc, and I. Petrović, "Drivable Path Planning Using Hybrid Search Algorithm Based on E* and Bernstein-Bézier Motion Primitives," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, no. 1957, pp. 1–15, 2019.

[3] L. E. Dubins, "On Curves of Minimal Length with a Constraint on Average Curvature, and with Prescribed Initial and Terminal Positions and Tangents," *American Journal of Mathematics*, vol. 79, no. 3, p. 497, 1957.

[4] M. Shanmugavel, A. Tsourdos, B. White, and R. Zbikowski, "Co-operative path planning of multiple UAVs using Dubins paths with clothoid arcs," *Control Engineering Practice*, vol. 18, no. 9, pp. 1084–1092, 2010.

[5] F. Ghilardelli, G. Lini, and A. Piazzi, "Path Generation Using η^4-Splines for a Truck and Trailer Vehicle," *IEEE Transactions on Automation Science and Engineering*, vol. 11, pp. 187–203, jan 2014.

[6] T. Berglund, U. Erikson, H. Jonsson, K. Mrozek, and I. Söderkvist, "Automatic generation of smooth paths bounded by polygonal chains," no. August 2014, 2001.

[7] M. Lepetič, G. Klančar, I. Škrjanc, D. Matko, and B. Potočnik, "Time optimal path planning considering acceleration limits," *Robotics and Autonomous Systems*, vol. 45, no. 3-4, pp. 199–210, 2003.

[8] C. Kielas-Jensen and V. Cichella, "BeBOT: Bernstein Polynomial Toolkit for Trajectory Generation," *IEEE International Conference on Intelligent Robots and Systems*, no. 3, pp. 3288–3293, 2019.

Vpliv kinematičnih indeksov manipulabilnosti na sposobnost izvedbe trajektorije orodja z robotom UR5

Saša Stradovnik, Aleš Hace

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
E-pošta: sasa.stradovnik@um.si, ales.hace@um.si

The infuence of kinematic manipulability indices on the execution of the tool trajectory with the robot UR5

Abstract. *A detailed understanding of the movement capability of robots in their workspace is crucial for the successful planning and execution of complex robot tasks. The importance of the feasibility of robot tasks, which is determined by singularities, kinematic and dynamic limits or collisions, receives even more attention with the development of intelligent collaborative robotic systems.*

In the past, different robot performance indices have been developed to measure and qualify the different characteristic of the robot workspace. However, when planning a specific robot task, its actual influence on the kinematic ability to perform a robot task is questioned. In this paper an approach to compare different manipulability indices based on measurable kinematic characteristic during task execution is presented in order to optimally position the workpiece relative to the robot base frame. In this way, we can increase the probability of a successfully executed robot task, reduce the time required for its planning and achieve a more accurate execution.

1 Uvod

Kadar opisujemo delovni prostor robota, ga po navadi v grobem opišemo s pomočjo dosegljivega in priročnega delovnega prostora [1]. Dosegljiv delovni prostor robota obsega volumen okrog robota, ki ga lahko robot doseže z vsaj eno od vseh možnih orientacij. Čeprav lahko robot znotraj priročnega delovnega prostora, ki je podmnožica dosegljivega delovnega prostora, doseže posamezno točko s poljubno orientacijo, je pri načrtovanju robotske naloge ključnega pomena tudi poznavanje drugih parametrov, ki pomembno vplivajo na njeno izvedljivost. Pogosto se lahko zgodi, da robot prekorači svoje kinematične in dinamične limite, kot so položajne sklepne limite, limite sklepnih hitrosti, pospeškov ali navorov. Lahko se tudi zgodi, da je gibljivost robotskega mehanizma močno zmanjšana v primeru, da se konfiguracija robota nahaja v bližini singularne lege ali pa neizvedljivost robotske trajektorije povzročijo kolizije robota samega s sabo ali z njegovo okolico.

Da bi lahko natančneje opisali delovni prostor robota so bili razviti t.i. robotski performančni indeksi s katerimi lahko merimo in kvalificiramo različne karakteristike delovnega prostora robota [2]. Pogosto se nanašajo na splošno analizo delovnega prostora, v tem članku pa se osredotočimo na obravnavo indeksov, ki se nanašajo na konkretno pot gibanja orodja na robotu UR5. Sposobnost premikanja in gibljivosti v njegovem delovnem prostoru analiziramo za potrebe aplikacij natančnega sledenja obliki površine. Le-te predstavljajo eno izmed kompleksnejših nalog za robota, saj jih večina zahteva generiranje konstantne hitrosti gibanja orodja, ohranjanje pravokotne orientacije orodja na površino tekom izvajanja robotske trajektorije ter zagotavljanje zadostne togosti pravokotno na smer gibanja. Naš namen v tem članku je zasnovati kvalitativno vrednotenje poti pri obravnavi znane robotske naloge na podlagi različnih indeksov manipulabilnosti poti, s katerimi lahko potem načrtujemo optimalno postavitev obdelovanca. Pri tem vpeljemo nov utežen indeks smerne manipulabilnosti poti, primerjamo njegov vpliv na kinematične sposobnosti robota pri izvajanju podane robotske trajektorije z drugimi opisanimi indeksi manipulabilnosti poti ter ovrednotimo kateri izmed njih nam daje najbolj optimalne rezultate s stališča kinematičnih sposobnosti robota.

V nadaljevanju predstavimo različne izpeljanke kinematičnih indeksov manipulabilnosti na podlagi katerih lahko izrišemo barvne mape optimalne postavitve obdelovanca v območju obravnavanega delovnega prostora robota. Vpliv definiranih indeksov verificiramo z zajemom sklepnih hitrosti in na ta način ocenimo, kateri izmed njih daje najboljše rezultate pri obravnavi znane robotske naloge.

2 Indeksi kinematične manipulabilnosti

Kot je že znano, prenos gibanja od posameznega aktuatorja ne povzroči enakih hitrosti gibanja vrha robota v vseh točkah delovnega prostora. Prenos gibanja je odvisen predvsem od strukture in geometrije mehanizma, zato lahko robotski mehanizem zelo poenostavljeno obravnavamo kot prenosnik moči s spreminjajočim se prenosnim razmerjem. Skalarno predstavitev prenosa sklepnih hitrosti na hitrosti vrha robota opisuje indeks kinematične manipulabilnosti *w*, ki ga je prvi vpeljal Yoshikawa [3]:

$$w = \sqrt{\det(J(q)J(q)^T)} \tag{1}$$

Njegova vrednost podaja sposobnost spreminjanja pozicije vrha robota glede na spreminjanje pozicije sklepov pri premiku v poljubni smeri in na ta način nekako podaja bližino do singularne konfiguracije robota. Konfiguracija robota z boljšo manipulabilnostjo zahteva nižje sklepne hitrosti pri premiku z enako

hitrostjo vrha robota. Opisuje gibljivost točno določene konfiguracije robota v njegovem delovnem prostoru in je neodvisna od obravnavane poti. Z uporabo singularnega razcepa ga lahko opišemo tudi kot produkt singularnih vrednosti Jacobijeve matrike.

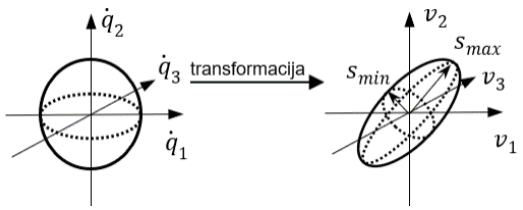
$$J = U\Sigma V^T \tag{2}$$

$$\Sigma = \mathrm{diag}(s_1, s_2, ..., s_m) \tag{3}$$

$$s_1 \geq s_2 \geq ... \geq s_m \geq 0 \tag{4}$$

$$w = \prod_{i=1}^{m} s_i \tag{5}$$

pri čemer U in V predstavljata ortogonalni matriki ter Σ diagonalno matriko singularnih vrednosti s_i. Singularne vrednosti Jacobijeve matrike določajo tudi glavne osi hitrostnega elipsoida v posamezni točki, ki nastane kot rezultat transformacije hitrostne hipersfere iz prostora notranjih koordinat v elipsoid v prostoru zunanjih koordinat, kot je prikazano na sliki 1. Glavne osi elipsoida predstavljajo smeri v katerih ima robotski mehanizem najboljše oz. najslabše sposobnosti generiranja hitrosti in sil.



Slika 1. Hitrostni elipsoid manipulabilnosti

V nadaljevanju predstavimo indekse manipulabilnosti za obravnavo celotne robotske poti.

Indeks kinematične manipulabilnosti poti

Za razliko od splošnega indeksa *w*, indeks kinematične manipulabilnosti poti w^{path} definiramo kot aritmetično sredino manipulabilnosti točk po celotni poti.

$$w^{path} = \frac{1}{N}\sum_{k=1}^{N} w(k) \tag{6}$$

N predstavlja število vseh kontrolnih točk na poti.

Smerni translacijski indeks kinematične manipulabilnosti poti

Sposobnost generiranja translacijskih hitrosti vrha robota ni odvisna le od položaja vrha robota v njegovem delovnem prostoru, temveč tudi od smeri v kateri se robot premika. To karakteristiko lahko najbolj interpretabilno predstavimo s pomočjo prej omenjenega hitrostnega elipsoida manipulabilnosti. Sposobnost generiranja hitrosti v smeri podani z vektorjem *u* določimo z indeksom smerne manipulabilnosti w_{dir}, ki je proporcionalen dolžini daljice od centra elipsoida do točke na površini elipsoida v tej smeri [4].

$$w_{dir} = \left[u^T (J(q)J(q)^T)^{-1} u\right]^{-1/2} \tag{7}$$

Pri obravnavi točno določene poti orodja računamo aritmetično sredino smernega translacijskega indeksa manipulabilnosti w_{dir}^{path} po celotni poti.

$$w_{dir}^{path} = \frac{1}{N}\sum_{k=1}^{N} w_{dir}(k) \tag{8}$$

pri čemer *N* predstavlja število vseh kontrolnih točk na robotski poti.

Indeks minimalne singularne vrednosti poti

Rezultat singularnega razcepa Jacobijeve matrike predstavlja produkt treh matrik, pri čemer diagonala matrike singularnih vrednosti Σ vključuje singularne vrednosti obravnavane matrike. Skalarna vrednost minimalne singularne vrednosti je proporcionalna najkrajši osi hitrostnega elipsoida manipulabilnosti. Indeks minimalne singularne vrednosti poti orodja $s_{\min}^k$ predstavlja aritmetično sredino minimalne singularne vrednosti točk, pri čemer *N* predstavlja število vseh kontrolnih točk na robotski poti.

$$w_s^{path} = \frac{1}{N}\sum_{k=1}^{N} s_{\min}(k) \tag{9}$$

Utežen smerni translacijskih indeks kinematične manipulabilnosti poti

Kljub temu, da je aritmetična sredina zelo občutljiva na ekstremne vrednosti, se lahko zgodi, da robotska pot z visoko vrednostjo povprečne smerne manipulabilnosti, vključuje tudi kritične točke v katerih bi lahko prišlo do prekoračitve kinematičnih ali dinamičnih limit robota. Le-te bi lahko povzročile neizvedljivost robotske naloge. To se lahko zgodi predvsem kadar je pot orodja sestavljena iz premikov v različnih smereh, saj ima robot v različnih smereh različne zmožnosti. V ta namen želimo iz obravnave izločiti točke v katerih je vrednost minimalne smerne manipulabilnosti poti nižja od tiste, ki bi lahko povzročila neizvedljivost robotske naloge. Indeks iz enačbe (8) zato modificiramo kot:

$$w_{dir}^{path} = \begin{cases} \frac{1}{N}\sum_{k=1}^{N} w_{dir}(k), & w_{dir}(k) > w_{dir}^{\min} \\ 0, & sicer \end{cases} \tag{10}$$

V vseh ostalih točkah želimo, da bi bil padec smernega translacijskega indeksa manipulabilnosti čim manjši. S tem namenom uvedemo nov kriterij ocenjevanja kvalitete poti, ki se nanaša na relativni enostranski odklon smerne manipulabilnosti $\sigma_{L,rel}$, kjer v izračun vključimo samo vrednosti pod aritmetično sredino.

$$\sigma_{L,rel} = \frac{1}{w_{dir}^{path}} \sqrt{\frac{\sum_{k=1}^{N} (w_{dir}(k) - w_{dir}^{path})^2}{N}} \tag{11}$$

Vrednost relativnega standardnega odklona smerne manipulabilnosti normiramo glede na celoten delovni prostor in tako oblikujemo utež $\sigma_{L,rel}^{norm}$,

$$\sigma_{L,rel}^{norm} = (\sigma_{L,rel} - \sigma_{L,rel}^{\max}) / (\sigma_{L,rel}^{\min} - \sigma_{L,rel}^{\max}) \quad (12)$$

kjer maksimalna (minimalna) vrednost $\sigma_{L,rel}^{\max}$ ($\sigma_{L,rel}^{\min}$) predstavlja največji (najmanjši) odklon od povprečnega indeksa v obravnavanem prostoru. Utežen smerni indeks manipulabilnosti poti definiramo v obliki (13).

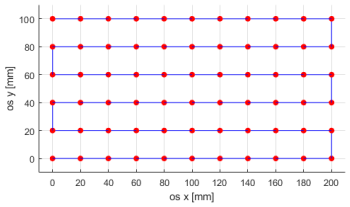
$$w_{dir}^{path\,'} = \sigma_{L,rel}^{norm} \cdot w_{dir}^{path} \quad (13)$$

3 Rezultati

Glede na to, da je pot gibanja orodja definirana z geometrijo obdelovanca in je le-ta v naprej poznana, je postavitev obdelovanca relativno glede na pozicijo robota lahko poljubna, vendar še kako pomembna s stališča izvedljivosti robotske trajektorije, saj mora robot, za dosego vseh točk trajektorije, ves čas delovati znotraj območja svojih kinematičnih in dinamičnih zmožnosti. Z namenom vrednotenja kvalitete poti se želimo prepričati kako posamezen opisan indeks manipulabilnosti vpliva na kinematične sposobnosti pri izvajanju znane robotske trajektorije ter kateri izmed opisanih indeksov manipulabilnosti nam daje najbolj optimalne rezultate.

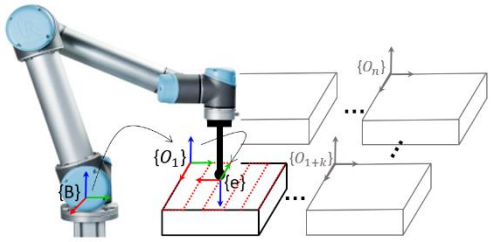
3.1 Robotska naloga

Vpliv obravnavanih indeksov manipulabilnosti na kinematične sposobnosti robota UR5 smo izvedli na primeru znane robotske poti orodja, sestavljene iz enakomerno razporejenih točk v x in y smeri, kot prikazuje slika 2.



Slika 2. Definirana pot orodja po površini obdelovanca

Ilustracijo obravnavanih postavitev surovcev prikazuje slika 3. Na prvem obdelovancu je ilustrirana pot obdelave, ki določa pot orodja. Pozicija in orientacija poti orodja sta definirani glede na lokalni koordinatni sistem obdelovanca {O}, pri čemer le-ta sovpada z začetno točko poti orodja. Koordinatni sistem obdelovanca {O}, ki ga premikamo po obravnavanem delovnem prostoru robota, je postavljen relativno glede na bazni koordinatni sistem robota {B}.

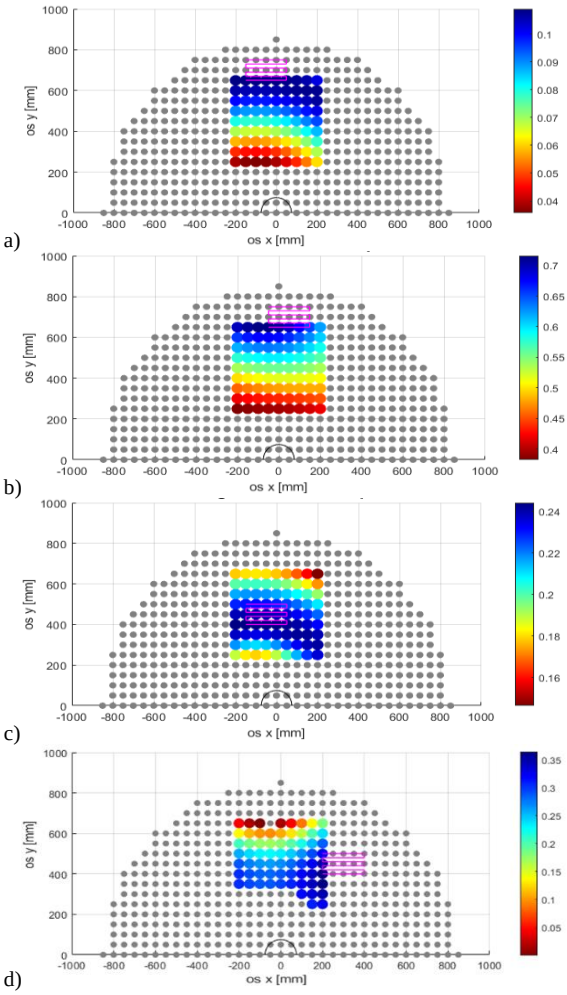


Slika 3. Robot UR5 in iskanje optimalne postavitve surovca za površinsko obdelavo

Tekom analize delovnega prostora robota, pot orodja premikamo s korakom 50 mm v x-y ravnini obravnavanega delovnega prostora robota UR5, ki zajema prostor v območju x = -200 mm…200 mm, y = 250 mm…650 mm in z = 100 mm relativno glede na bazni koordinatni sistem robota {B}. Za vsako posamezno pozicijo obdelovanca znotraj omenjenega območja izračunamo obravnavani indeks manipulabilnosti z uporabo Matlaba in knjižnice Robotic Toolbox [5], na podlagi katerega obravnavani točki delovnega prostora določimo pripadajočo barvo in na ta način grafično prikažemo karakteristiko obravnavanega indeksa manipulabilnosti v definiranem delovnem prostoru.

3.2 Barvna mapa manipulabilnosti

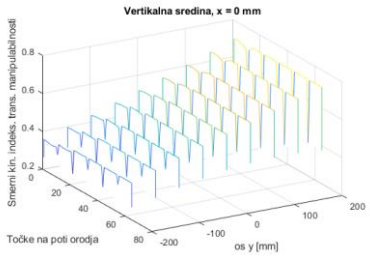
Barvna mapa manipulabilnosti predstavlja skupek točk z barvno upodobitvijo vrednosti posameznega indeksa manipulabilnosti poti orodja v obravnavanem delovnem prostoru robota v primeru, da začetno točko poti postavimo v opazovano točko. Primerjavo barvnih map manipulabilnosti smo naredili na podlagi štirih različnih indeksov, ki smo jih opisali v poglavju 2 (slika 4).



Slika 4. Barvna mapa: a) indeksa kinematične manipulabilnosti poti, b) smernega indeksa kinematične manipulabilnosti poti, c) minimalne singularne vrednost poti, d) uteženega smernega indeksa kinematične manipulabilnosti poti

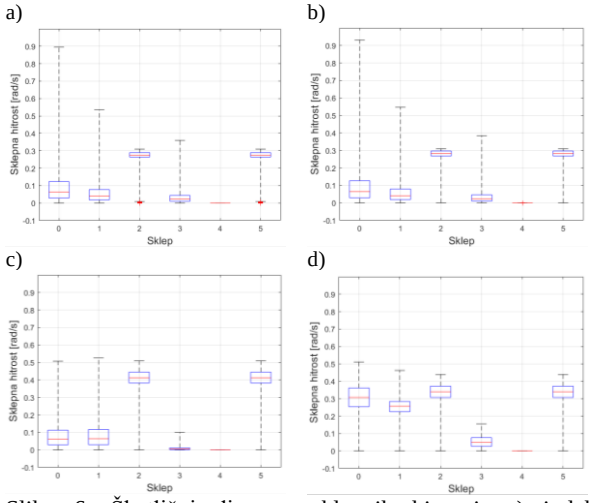
Pri tem območja z modro barvo predstavljajo območja z dobro manipulabilnostjo poti orodja, medtem ko rdeče

obarvana območja predstavljajo območja s slabo manipulabilnostjo poti orodja. Optimalna postavitev obdelovanca glede na najvišjo vrednost posameznega obravnavanega indeksa je prikazana z magenta barvo. Opazimo lahko, da se optimalne postavitve obdelovanca glede na posamezen indeks med sabo precej razlikujejo. Postavlja se torej vprašanje, katero optimalno rešitev je smiselno izbrati. V primeru smernega translacijskega indeksa kinematične manipulabilnosti opazimo, da se njegova povprečna vrednost v smeri pozitivne osi y povečuje, kar je razvidno tudi s slike 5. Poleg tega lahko na tej sliki opazimo, da na mestih, kjer pride do spremembe gibanja, smerna manipulabilnost močno pade. Z namenom, da bi poleg povprečja smerne manipulabilnosti upoštevali še njen padec, za katerega želimo, da bi bil čim manjši, smo vpeljali uteženi smerni translacijski indeks manipulabilnosti.



Slika 5. Diagram spreminjanja smerne manipulabilnosti v smeri y-smeri obravnavanega območja

3.3 Verifikacija



Slika 6. Škatlični diagram sklepnih hitrosti: a) indeks kinematične manipulabilnosti poti b) smerni indeks kinematične manipulabilnosti poti, c) minimalna singularna vrednost poti, d) utežen smerni indeks kinematične manipulabilnosti poti

Primerjavo učinkovitosti definiranih indeksov manipulabilnosti smo naredili na podlagi zajetih sklepnih hitrosti za pozicije poti orodja, kjer je vrednost posameznega indeksa najvišja. Simulacijo gibanja robota po definirani poti, katere sklepne hitrosti smo zajeli, smo generirali s pomočjo simulacijskega orodja URSim. Pri tem smo v vseh štirih primerih zagotavljali enake pogoje izvajanja robotske trajektorije. Robotsko trajektorijo smo generirali s hitrostjo 200 mm/s in s pospeškom 1200 mm/s^2. Absolutne sklepne hitrosti vsakega posameznega sklepa robota smo prikazali s pomočjo škatličnega diagrama, ki omogoča nazorno grafično ponazoritev in statistično analizo številčnih podatkov, kot prikazuje slika 6. V primeru a) in b) lahko opazimo veliko razpršenost vrednosti sklepnih hitrosti okrog mediane. Najmanjšo razpršenost sklepnih hitrosti lahko dosežemo s kvalitativnim vrednotenjem poti na podlagi indeksa c) in d). V primeru c) je sicer večina sklepnih hitrosti nižja kot v primeru d), vendar je maksimalna dosežena sklepna hitrost v primeru c) višja kot v primeru d).

4 Zaključek

Na podlagi prikazanih rezultatov je razvidno, da lahko s takšno analizo delovnega prostora predvidimo kinematično obnašanje robota za potrebe izvajanja želene robotske trajektorije, vendar opazimo tudi, da se območja z najvišjo vrednostjo različnih indeksov manipulabilnosti med sabo razlikujejo. Z namenom iskanja odgovora na to, kateri indeks ima največji vpliv na optimalno izvajanje robotske naloge s stališča njegovih kinematičnih sposobnosti, smo naredili njihovo primerjavo na podlagi zajema sklepnih hitrosti. Najbolj optimalne rezultate s stališča generiranih sklepnih hitrosti smo dosegli z upoštevanjem uteženega smernega translacijskega indeksa manipulabilnosti poti, ki smo ga predstavili v članku. Na podlagi analize smo ugotovili, da je povprečje smerne translacijske manipulabilnosti dobra osnova za iskanje optimalne poti, vendar je pri tem potrebno vključiti dodatne kriterije na podlagi katerih izločimo pozicije obdelovanca, ki vključujejo kritične točke v katerih pade smerna manipulabilnost pod dopustno mejo. Za nadaljnje kvalitativno razvrščanje poti je to povprečje smiselno utežiti še z enostransko standardno deviacijo, ki nam v povprečju podaja merilo za prisotnost točk z nižjo vrednostjo smerne manipulabilnosti.

Zahvala

To delo je bilo delno sofinancirano z raziskovalnim programom št. P2-0028 Javne agencije za raziskovalno dejavnost Republike Slovenije, in v okviru projekta ROBOTOOL-1, OP20.03540, kjer naložbo sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj.

Literatura

[1] O. Porges, R. Lampariello, J. Artigas, A. Wedler, C.Borst, and M. A. Roa, "Reachability and Dexterity: Analysis and Applications for Space Robotics,"(ASTRA), 2015.

[2] S. Patel and T. Sobh, "Manipulator Performance Measures - A Comprehensive Literature Survey," 2014.

[3] T. Yoshikawa, "Manipulability of Robotic Mechanisms," *The International Journal of Robotics Research*, 1985.

[4] S. Chiu, "Task Compatibility of Manipulator Postures," *The International Journal of Robotics Research*, vol. 7, no. 5, pp. 13-21, 1988.

[5] P. Corke, *Robotic Toolbox for Matlab*. 2018.

Optimizing robot motion synthesis from digital images using neural networks

Matija Mavsar[1]

[1]*Humanoid and Cognitive Robotics Lab, Dept. of Automatics, Biocybernetics, and Robotics, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana*
E-mail: matija.mavsar@ijs.si

Abstract

Teaching robots to navigate in an unfamiliar environment is a demanding but important step in enhancing robot autonomy. To achieve this, robots must be able to respond to visual representations of their surroundings with desired motions. Recent work has shown success in using deep encoder-decoder networks to map raw digit images to dynamic movement primitives, allowing the robot to reproduce the digits by writing. In our work, we modify the latent space size of the proposed network architecture in order to improve the network's performance on new digital image data. Additionally, we examine the latent space to extract average image representations of individual digits and include a simple classification network.

1 Introduction

To enable humanoid robots to effectively carry out neccessary actions in a dynamic environment, they must be able to translate visual representations into movement responses. An appropriate method for learning perception-to-action skills is therefore required.

One possible problem in this context is performing handwriting trajectories based on presented raw images of characters. Due to their ability to learn highly nonlinear transformations, neural networks are a good candidate for this type of translation. Ridge *et al*. [1] proposed an encoder-decoder network architecture, capable of converting raw images of digits into dynamic movement primitives (DMPs) [2], which define a robot trajectory, to facilitate direct reconstruction of handwriting motions from given number images. Additionally, in [3], a spatial transformer network and a motion transformer network were included to achieve an improved performance on various datasets and enable a humanoid robot to read and write digits, presented in arbitrary poses.

However, the generalization capability of the architecture in [1] can additionally be improved by modifications of the latent space. Furthermore, the current architecture is not capable of digit classification; it can only translate images into trajectories. Another possible requirement is for the robot to write a specific number, without an image present. Therefore, an average representation of individual digits must be extracted, which can be achieved by analyzing the clustering inside the latent space.

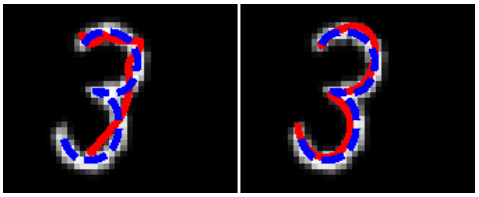


Figure 1: The output trajectory (solid line) of the neural network architecture designed by Ridge *et al*. [1], compared to the ground truth trajectory (dashed line), where left and right images show the results after 10 and 3610 epochs, respectively.

In this paper we focus on the network architecture from [1], where we mainly investigate the optimal latent space or bottleneck size that can be introduced into other architectures and additionally trained. Similarly, the classification feature can easily be implemented in a different network and retrained. The contributions presented in this paper are as follows: 1) We analyze and adjust the latent space of the network architecture introduced in [1] to achieve optimal performance; 2) We examine the latent space and achieve classification capability; and 3) We extract average representations of individual digits.

2 Related work

In the context of translating perception into motion several methods based on extracting task parameters (object pose) and using them to generate trajectories were implemented. Stulp *et al*. [4] used a Kinect camera to obtain the parameters for the object pose and passed them directly to DMP's function approximator to perform transport and grasping tasks. Pervez *et al*. [5] proposed a Deep-DMP formulation to model forcing terms of a DMP using a CNN, which allows for direct processing of camera images and removes the need for extraction of task parameters during motion reproduction. For accomplishing folding tasks, Yang *et al*. [6] utilized a deep convolutional autoencoder to extract image features and combined them with robot motor angles in a fully connected time delay network, in order to obtain online predictions of next motion steps. Finn and Levine [7] presented a method for object manipulation by using a deep predictive neural network to plan pushing motions. The model predicts future images, resulting from executing proposed actions, and chooses the most beneficial movement. The first method

requires separate parameter extraction and the others produce the next step from the current image, while in [1] complete trajectories are generated based on one image.

Several attemps to implement calligraphy with robots were made; Zhang *et al*. [8] developed a system that uses a CNN to identify parts of sketched characters, retrieves corresponding standard strokes from a pre-built library and utilizes global optimization approach to achieve a visually pleasing result; Chao *et al*. [9] employed a generative adversarial network to generate robotic writing trajectories, which are then compared to textbook characters by the discriminative network and the results are used to further train the generative network; and Gao *et al*. [10] used a cyclical framework, where a convolutional auto-encoder network evaluates an image of robot-written strokes and a learning subsystem attempts to improve the writing trajectory model.

3 Methods

3.1 Data structure and motion representation

The network, presented in [1], takes images of digits as inputs and outputs corresponding movement parameters. The data pairs thus have the following form:

$$\mathbf{D} = \{\mathbf{C}_j,\ \mathbf{k}_j\}_{j=1}^{M}, \tag{1}$$

where M is the number of training pairs, $\mathbf{C}_j$ input images and $\mathbf{k}_j$ are DMP parameters, calculated from trajectories that are used to replicate a specific input image.

A robot trajectory can be described with a DMP using a system of differential equations [2] with several parameters, including initial position $\boldsymbol{y}_0 \in \mathbb{R}^d$, goal position $\boldsymbol{g} \in \mathbb{R}^d$ and a nonlinear forcing term $\mathbf{F}(x) \in \mathbb{R}^d$ defined as a linear combination of radial basis functions (RBFs):

$$\mathbf{F}(x) = \frac{\sum_{k-1}^{N} \boldsymbol{w}_k \boldsymbol{\Psi}_k(x)}{\sum_{k=1}^{N} \boldsymbol{\Psi}_k(x)} x, \tag{2}$$

where $x \in \mathbb{R}$ is the phase used to attain time independency, $\boldsymbol{\Psi}_k(x)$ are RBFs (exponential functions in our case), $\boldsymbol{w}_k$ their weights, and d the number of degrees of freedom. $\mathbf{F}(x)$ can thus be used to reproduce any smooth motion between the initial and final robot position.

The parameters $\mathbf{k}_j$ that the neural network must learn are therefore characterized as

$$\mathbf{k}_j = \{\{\boldsymbol{w}_k\}_{k=1}^{N},\ \tau,\ \boldsymbol{g},\ \boldsymbol{y}_0\}, \tag{3}$$

while the rest of the parameters required to fully specify a DMP are set to fixed values to ensure critical damping and convergence.

3.2 Encoder-decoder network architecture

The structure of the image-to-motion encoder-decoder network (CIMEDNet) that was used in our experiments is the same as presented in [1] and shown in Figure 2.

The encoder part of the architecture consists of two convolutional layers and two fully-connected layers. The convolutional layers take a $40 \times 40 \times 1$ pixel image as input, where the first layer has a 5×5 kernel size and 10 feature maps, and the second has the same kernel size but 20 feature maps. The two fully-connected layers have 600 and 200 neurons, respectively, and the bottleneck of the network, which represents the latent space, has 20 neurons. The following part is the decoder with a layer of size 35 and finally the output layer with 54 output values that represent the DMP parameters of a trajectory.

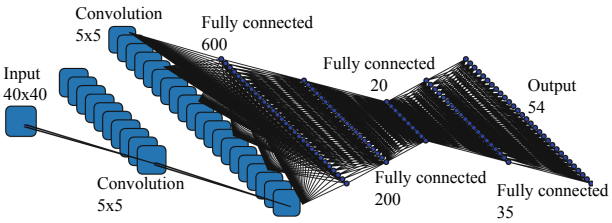


Figure 2: The CIMEDNet neural network architecture proposed by Ridge *et al*. [1], used in the experiments.

To evaluate the network output, we implemented a mean squared error loss function that measures the difference between the training DMP parameters and the network output parameters and thus minimizes trajectory error through the error of the encoding DMP parameters.

3.3 Classification network

To implement classification capability, a simple neural network was designed (Figure 3, right). This way, we also tested whether the latent space representations of individual digits can be distinguished from one another. The network consists of three fully-connected layers; the first one with the same number of neurons as the latent space layer of the CIMEDNet, the second with 80 neurons and the output layer with 10 neurons, representing detection probabilities for each digit. The network takes the output of the CIMEDNet latent space layer as input and outputs the corresponding probability vector. The cost function used during training was a mean squared error, comparing network digit prediction vector to the actual label vector of an image.

4 Experiments

4.1 Latent space size variation

We trained the CIMEDNet architecture with latent space sizes varying from 2 to 35, where we observed the convergence rate and final test error for each size. The training data were 40×40 images of digits and corresponding trajectory parameters from the s-MNIST dataset, described later, while the models were trained end-to-end.

We used a learning rate of 0.0005 and the batch size was set to 140. The training was automatically halted if the best validation loss remained unchanged for 60 consecutive epochs and the data was split in $70\%/15\%/15\%$ ratios for training, validation and test data, respectively.

The datasets used in the experiments with latent space variation were the same as described in [1], where synthetic trajectories and grayscale image data were generated by combining geometric elements with varying parameters. Additionally, salt-and-pepper noise was added

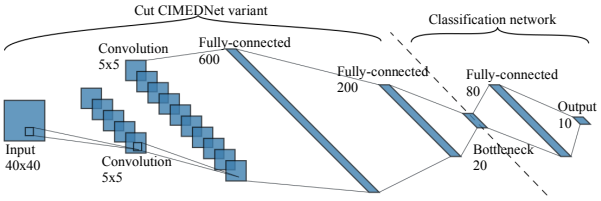


Figure 3: The architecture used to train the classification network. The CIMEDNet structure was frozen and cut at the latent space layer and the outputs were fed into additional layers of the classification network.

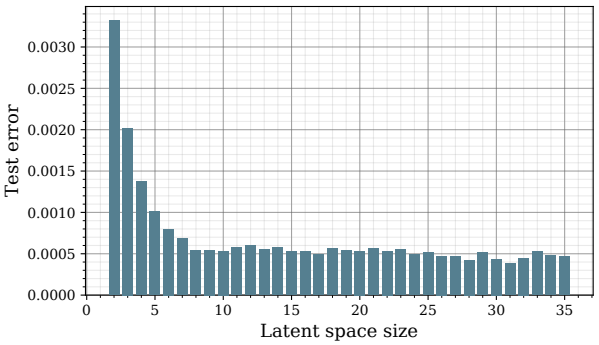


Figure 4: Test errors for different neural network architectures with latent space layer sizes varying from 2 to 35.

and the images were transformed using various affine transformations. The DMP trajectories are represented using 25 RBFs for each of the dimensions (50 for 2D images) along with 4 parameters for start and goal positions. The CIMEDNet was trained on noiseless s-MNIST dataset and the optimal network was tested on additional datasets with added Gaussian noise and reduced contrast.

4.2 Classification network

In order to train the neural network for classification of digits, we first selected the CIMEDNet model with the optimal latent space size. The model was then frozen and cut after the latent space layer and the s-MNIST dataset was forward passed through network. The classification network was included as shown in Figure 3. This way latent space representations of digit images were obtained that could then be used along with digit labels for the training of the classification network.

The training parameters were the same as described in Section 4.1, where images along with the corresponding digit values were used in the training process. The classification capability was tested on the optimal architecture by using the same test data as for testing the CIMEDNet, where we calculated the number of wrong predictions.

4.3 Average digit representation

In order for robot to be able to write a specific digit, we passed the dataset through the optimal trained CIMEDNet architecture and calculated the average latent space values for each digit. The values were fed through the decoder part of the network structure and the resulting DMP parameters integrated into trajectories. This approach stems from valuable arithmetic characteristics of latent space [11].

5 Results

5.1 Optimization of latent space layer size

The test error for architectures with variable latent space size is shown in Figure 4 and represents the mean squared difference between the actual DMP parameters of the trajectory and the predicted DMP parameters. The average error on the test dataset generally decreases with larger sizes and converges to approximately 5×10^{-4}, where a minimum is obtained with 31 neurons in the latent space layer (3.78×10^{-4}). To achieve reduced architecture complexity, 8 neurons may be enough for a sufficiently low test error compared to the optimal one.

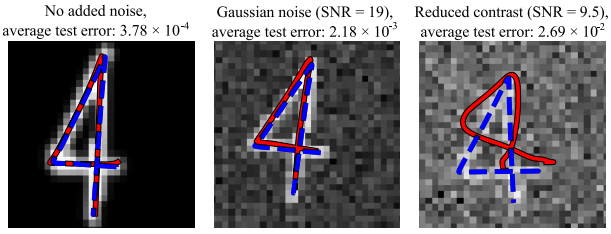


Figure 5: The average test errors for different datasets along with example digit images and the corresponding ground truth trajectory (dashed line) and predicted trajectory (solid bordered line).

In addition to test errors, the validation errors of each architecture were also observed. The rate of validation error convergence noticeably increased with the latent space size, and the error reached lower values at larger sizes. However, for layers of size 8 and more, the final validation error varied considerably less and no significant learning rate improvement was achieved.

To evaluate the ability of the architecture to generalize from noiseless data to more realistic images, more likely to be encountered in the real world, the architecture with the lowest test error (latent space size 31) was additionally tested on several datasets. Figure 5 shows average test error and an example digit image with the ground truth trajectory (dashed line) and the predicted trajectory (solid line) for each dataset.

The results indicate that the network is most successful at replicating the digits with no added noise and digits with gaussian noise where a SNR of 19 was used. On the other hand, the architecture has noticeable difficulties generating trajectories for images with SNR of 9.5 and reduced contrast, where the average test errors are higher by an order of magnitude. This indicates a satisfactory ability of the network to generalize to low noise data, but is much less successful at transforming heavily distorted images.

5.2 Implementation of classification capability

The addition of the classification network was assessed by passing the s-MNIST test data through the optimal CIMEDNet architecture, which was cut after the latent space layer of 31 neurons, as well as through some of the other networks with different latent space sizes. Ta-

	Latent space size					
	2	8	15	20	31	35
N_{WR}	8	2	2	2	1	2
$\%_{OK}$	99.7	99.93	99.93	99.93	99.97	99.93

Table 1: Prediction results of the classification network, where N_{WR} represents the number of incorrect digit predictions and $\%_{OK}$ represents the percentage of correct predictions among a test dataset of 3000 samples.

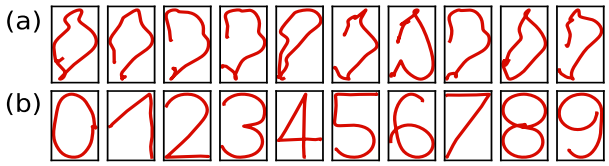


Figure 6: The average digit trajectories, obtained by averaging the latent space representations of test dataset: (a) Latent space size 31. (b) Latent space size 9.

ble 1 shows the results of the classification process, with the numbers of incorrect predictions and percentages of correct predictions shown for each of the structures.

The classification network demonstrates fairly accurate behavior, where the percentage of correct predictions is always above 99.7 % and the accurracy achieved with the optimal size of 31 is 99.97 %. High classification accuracy suggests that the CIMEDNet creates distinct latent space representations of digit images during training, despite the fact that it is only concerned with correct replication of the curves on presented images.

5.3 Average digit trajectory extraction

The classification process showed good clustering of the latent space, therefore we expected to obtain useful average digit trajectories by averaging the latent space representations of samples for each digit.

We calculated the average value of the latent space layer output for layer size of 31, where the input was test dataset. The result was passed through the decoder part of the network and the output DMP parameters were integrated into a trajectory for each of the ten digits (Figure 6a). However, the resulting curves were not recognizable as individual digits.

Additional structures with smaller latent space sizes were tested and drastically better results were obtained, where the most distinguishable digit trajectories were achieved by sizes of 5 up to 13, with size 9 being the most legible (Figure 6b), while results of larger sizes were considerably worse. This indicates that small latent spaces are more suitable for extraction of average trajectories, which may be attributed to more effective clustering of the abstract image representations. A strategy to simply take the network architecture with the lowest test error may therefore not always be the most rewarding.

6 Conclusion

In this work, we modified a neural network architecture for generation of robot trajectories based on images of digits in order to achieve the optimal latent space size of the encoder-decoder structure. Lowest error was achieved with size 31, however all sizes above 8 showed similar performance. The optimal architecture was further tested on distorted datasets, where significant accuracy reduction was observed for highly noisy images.

The network's functionality was enhanced by adding classification capability, where high accuracy was achieved, presumably due to efficient clustering of digit latent space representations. Latent space was further utilized to extract average digit trajectories by calculating the average latent space values for all samples of each digit. Smaller latent space sizes were drastically more successful, generating clear digit shapes, while larger sizes produced unintelligible curves, suggesting that a complex architecture is not neccessarily the most advantageous.

References

[1] B. Ridge, R. Pahič, A. Ude, and J. Morimoto, "Convolutional encoder-decoder networks for robust image-to-motion prediction", in *International Conference on Robotics in Alpe-Adria Danube Region*, 2019, pp. 514–523.

[2] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: Learning attractor models for motor behaviors", *Neural computation*, vol. 25, no. 2, pp. 328–373, 2013.

[3] B. Ridge, R. Pahič, A. Ude, and J. Morimoto, "Learning to write anywhere with spatial transformer image-to-motion encoder-decoder networks", in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2111–2117.

[4] F. Stulp, G. Raiola, A. Hoarau, S. Ivaldi, and O. Sigaud, "Learning compact parameterized skills with a single regression", in *2013 13th IEEE-RAS International Conference on Humanoid robots (Humanoids)*, 2013, pp. 417–422.

[5] A. Pervez, Y. Mao, and D. Lee, "Learning deep movement primitives using convolutional neural networks", in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 2017, pp. 191–197.

[6] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable folding task by humanoid robot worker using deep learning", *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 397–403, 2016.

[7] C. Finn and S. Levine, "Deep visual foresight for planning robot motion", in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2786–2793.

[8] X. Zhang, Y. Li, Z. Zhang, K. Konno, and S. Hu, "Intelligent chinese calligraphy beautification from handwritten characters for robotic writing", *The Visual Computer*, vol. 35, no. 6–8, pp. 1193–1205, 2019.

[9] F. Chao, J. Lv, D. Zhou, L. Yang, C.-M. Lin, C. Shang, and C. Zhou, "Generative adversarial nets in robotic chinese calligraphy", in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1104–1110.

[10] X. Gao, C. Zhou, F. Chao, L. Yang, C.-M. Lin, TaoXu, C. Shang, and Q. Shen, "A data-driven robotic chinese calligraphy system using convolutional auto-encoder and differential evolution", *Knowledge-Based Systems*, 2019.

[11] P. Bojanowski, A. Joulin, D. Lopez-Pas, and A. Szlam, "Optimizing the latent space of generative networks", in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, Stockholm, Sweden, 2018, pp. 600–609.

Learning of the Velocity Profile for Quality Inspection Motion Using PoWER

Zvezdan Lončarević, Rok Pahič, Andrej Gams

All authors are with the Jožef Stefan Institute,
and with the Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: zvezdan.loncarevic@ijs.si

Abstract

Modern factories and factories of the future are increasing their demands in terms of productivity and time of production line adaptation for the each new product. Products from automated production lines are often checked with machine vision, but the process of inspection is set-up manually. Manually set-up processes are not optimal so they represent the bottleneck in the production capacities. That is our motivation for automatizing the set-up of the visual inspection process. Neural network based-algorithms are already capable of reliably detecting defects in products, but they should be provided with sharp images. Manually determining the inspecting motion that would provide us with sharp images is a long-lasting and tedious task for a programmer. In this paper, we test the possibilities of using RL algorithm named PoWER for the purpose of automatizing this process.

1 Introduction

For industrial tasks such as manipulation and visual quality control, hardware and the environment need to be properly tuned. Usually this process is done only once, and after that, the production line setup stays the same during the whole life cycle [1]. Hardware is often designed in a way that some adjustments can only be carried out manually. Although hardware is set up and optimized for production only once, this is a tedious and demanding task and it heavily depends on the programmer's experience. The required cycle times usually set high demands on how fast the quality control has to be, and how it can be performed [2]. As a consequence of the fact that optimality of the task depends on the programmer's skills and intuition, it is obvious that learning algorithms have a great potential for reducing the cycle times of the production thus saving a lot of money in both set-up and increased productivity.

Learning the required movement of the robot for the quality inspection task can be done by off-line or on-line programming. However, none of the methods provides us with the optimal velocity profile of the camera so it needs to be learned or manually calculated. Calculating this velocity for the objects where the robot has to change its orientation is far from trivial. That is why in this paper we propose an algorithm that uses reinforcement learning (RL) to optimize the velocity of robot motion for visual quality inspection in order to achieve fastest robot motion that does not introduce blur to the image.

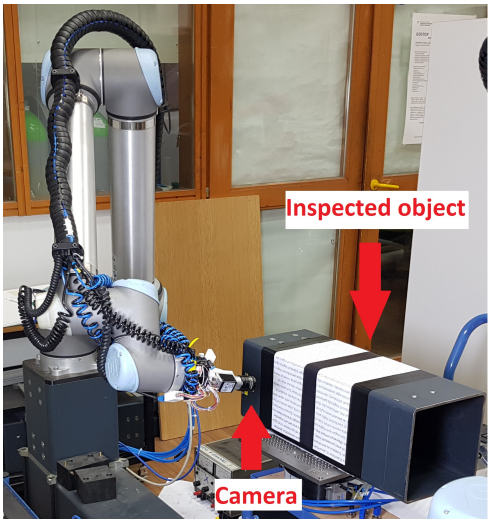


Figure 1: Experimental setup with the UR10 robot and object with the sharp edge

Reinforcement learning is a branch of machine learning that is capable of solving the optimization problems without having explicitly defined all the parameters or the model of the environment. In this paper we explore possibilities of using the PoWER algorithm [3].

In order to deal with high dimensionality of the search space for RL, we encode our trajectory obtained from the CAD model into Cartesian Dynamical Movement Primitives (CDMPs) with explicitly defined velocity profile [4].

We present our approach on the example where an in-hand camera moves around the object of inspection and provides a continuous visual stream. After that, visual feedback can be processed with different methods, including deep learning methods [5], or with comparing the obtained images with the predefined pattern [6]. This is, however, not the focus of this paper.

The paper is organized as follows: In the next section, short recapitulation of CDMPs is given. In section 3 we briefly present PoWER algorithm. Section 4 shows experimental setup and the parameters of the system and Section 5 presents obtained results. The paper concludes with a short outlook on the obtained results and suggestion for the future work.

2 Cartesian Dynamical Movement Primitives (CDMPs)

CDMPs are composed of the position and the orientation part. The position part of the trajectory is the same as in standard DMPs [7]. Since the orientation part is given in unit quaternions, it requires special treatment for the nonlinear dynamic equations and for their integration.

CDMPs encode the trajectory into the following parameters: weights $\boldsymbol{w}_k^p,\ \boldsymbol{w}_k^o \in \mathbb{R}^3,\ k = 1, \ldots, N$, which correspond to the position and orientation parts of the trajectory, respectively; trajectory duration τ and the final, goal position $\boldsymbol{g}^p$ and the final orientation $\boldsymbol{g}^o$. Variable N stands for the number of radial basis functions used for encoding the trajectory. The orientation is represented by a unit quaternion $\boldsymbol{q} = v + \boldsymbol{u} \in \mathrm{S}^3$, where S^3 is a unit sphere in $\mathbb{R}^4$. $v \in \mathbb{R}$ is its scalar and $\boldsymbol{u} \in \mathbb{R}^3$ its vector part. As in standard DMPs, we use transformation and phase systems for encoding the trajectory. Position ($\boldsymbol{p}$) and orientation ($\boldsymbol{q}$) of the trajectories are encoded using a transformation system that consists of the following differential equations:

$$\tau \dot{\boldsymbol{z}} = \alpha_z(\beta_z(\boldsymbol{g}^p - \boldsymbol{p}) - \boldsymbol{z}) + \boldsymbol{f}_p(x), \quad (1)$$

$$\tau \dot{\boldsymbol{p}} = \boldsymbol{z}, \quad (2)$$

$$\tau \dot{\boldsymbol{\eta}} = \alpha_z\left(\beta_z 2 \log\left(\boldsymbol{g}^o * \bar{\boldsymbol{q}}\right) - \boldsymbol{\eta}\right) + \boldsymbol{f}_o(x), \quad (3)$$

$$\tau \dot{\boldsymbol{q}} = \frac{1}{2}\boldsymbol{\eta} * \boldsymbol{q}, \quad (4)$$

In these equations, $\boldsymbol{z}$ and $\boldsymbol{\eta}$ denote the scaled linear and angular velocity ($\boldsymbol{z} = \tau\dot{\boldsymbol{p}}$, $\boldsymbol{\eta} = \tau\boldsymbol{\omega}$). For details on quaternion product $*$, conjugation $\bar{\boldsymbol{q}}$, and the quaternion logarithm $\log(\boldsymbol{q})$, see [4]. The nonlinear parts, termed also forcing terms, $\boldsymbol{f}_p$ and $\boldsymbol{f}_o$ are defined as

$$\boldsymbol{f}_p(x) = \boldsymbol{D}_p \frac{\sum_{k=1}^N \boldsymbol{w}_k^p \Psi_k(x)}{\sum_{k=1}^N \Psi_k(x)} x, \quad (5)$$

$$\boldsymbol{f}_o(x) = \boldsymbol{D}_o \frac{\sum_{k=1}^N \boldsymbol{w}_k^o \Psi_k(x)}{\sum_{k=1}^N \Psi_k(x)}, \quad (6)$$

where the nonlinear forcing terms are defined as a linear combination of radial basis functions Ψ_k:

$$\Psi_k(x) = \exp\left(-h_k\left(x - c_k\right)^2\right). \quad (7)$$

Forcing terms contain parameters $\boldsymbol{w}_k^p,\ \boldsymbol{w}_k^o \in \mathbb{R}^3$. They have to be learned, for example directly from an input Cartesian trajectory $\{\boldsymbol{p}_j, \boldsymbol{q}_j, \dot{\boldsymbol{p}}_j, \boldsymbol{\omega}_j, \ddot{\boldsymbol{p}}_j, \dot{\boldsymbol{\omega}}_j, t_j\}_{j=1}^T$. The scaling matrices $\boldsymbol{D}_p,\ \boldsymbol{D}_o \in \mathbb{R}^{3\times3}$ can be set to $\boldsymbol{D}_p = \boldsymbol{D}_o = \boldsymbol{I}$. Other possibilities are described in [4]. Here c_k are the centers and h_k the widths of the radial basis functions. The distribution of weights can be, as in [8], $c_k = \exp\left(-\alpha_x \frac{k-1}{N-1}\right)$, $h_k = \frac{1}{(c_{k+1} - c_k)^2}$, $h_N = h_{N-1}$, $k = 1, \ldots, N$. The time constant τ is set to the desired duration of the trajectory, i. e. $\tau = t_T - t_1$. The goal position and orientation are usually set to the final position and orientation on the desired trajectory, i. e. $\boldsymbol{g}^p = \boldsymbol{p}_{t_T}$ and $\boldsymbol{g}^o = \boldsymbol{q}_{t_T}$. In order to avoid explicit dependency of the trajectory from the time and to synchronize position and orientation of the trajectory, phase system is used:

$$\nu(x)\tau\dot{x} = -\alpha x. \quad (8)$$

In the above equation, α is positive constant, x is the phase that starts at $x_1 = 1$ and converges to $x_J = e^{-\alpha}$ as the trajectory goal is reached and $\nu(x)$ is the velocity profile of the trajectory. It is used for temporal scaling because it allows us to separate the position and velocity and consequently allow optimization of one without changing the other. We encoded the velocity profile as a weighted combination of kernel functions:

$$\nu(x) = \frac{\sum_{k=1}^R w_k^\nu \Psi_k(x)}{\sum_{k=1}^R \Psi_k(x)}. \quad (9)$$

Here, R defines the number of radial basis functions given in Equation 7 used to encode the velocity profile. We used $R = 25$ to define the profile of the velocity. CDMP derivation and auxiliary math are explained in [4].

3 PoWER

Policy Learning by Weighting Exploration with the Returns (PoWER) [3] is an Expectation-Maximization based RL algorithm that tries to maximize the expected return of the trials by updating its policy over the iterations. If PoWER is combined with importance sampling [9], weights defining the velocity profile are updated using:

$$\boldsymbol{w}_{n+1}^\nu = \boldsymbol{w}_n^\nu + \frac{\sum_{i=1}^m \left\{ (\boldsymbol{w}_i^\nu - \boldsymbol{w}_n^\nu) \sum_{j=1}^J \boldsymbol{M}_n(x_j) Q_i^\pi(x_j) \right\}}{\sum_{i=1}^m \left\{ \sum_{j=1}^J \boldsymbol{M}_n(x_j) Q_i^\pi(x_j) \right\}}. \quad (10)$$

In order to explore the response of the system to the new parameters, exploration noise ($\boldsymbol{\epsilon} = \mathcal{N}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}})$) is added $\hat{\boldsymbol{w}}_{n+1}^\nu = \boldsymbol{w}_{n+1}^\nu + \boldsymbol{\epsilon}$ The first sum in the above equation is importance sampler that chooses and sums m examples with the highest terminal reward from the previous rollouts. The corresponding weights from those rollouts are $\boldsymbol{w}_i^\nu$. Second summation sums over the phase steps through the intermediate rewards (Q_i^π) weighted by weighting vector defined as:

$$\boldsymbol{M}_n(x) = \frac{\boldsymbol{\phi}(x)\boldsymbol{\phi}^T(x)}{\boldsymbol{\phi}^T(x)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}\boldsymbol{\phi}(x)}, \quad (11)$$

where value of ϕ for k-th dimension is computed as:

$$\phi_k(x) = \frac{\Psi_k(x)}{\sum_{k=1}^R \Psi_k(x)}. \quad (12)$$

The maximization of the expectation can also be achieved with varying the covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_{n+1} = \frac{\sum_{i=1}^m \left\{ (\boldsymbol{w}_i^\nu - \boldsymbol{w}_n^\nu)(\boldsymbol{w}_i^\nu - \boldsymbol{w}_n^\nu)^T \sum_{j=1}^J Q_i^\pi(x_j) \right\}}{\sum_{i=1}^m \left\{ \sum_{j=1}^J Q_i^\pi(x_j) \right\}}. \quad (13)$$

Equation (10) can be simplified by choosing $\hat{\boldsymbol{\Sigma}}$ to be diagonal matrix with all the same values on its diagonal. Note that this makes the search noise for the parameters same and pairwise independent. This simplifies the weighting vector part from Equation (11) to:

$$\boldsymbol{M}_n(x) = \frac{\boldsymbol{\phi}(x)\boldsymbol{\phi}^T(x)}{\boldsymbol{\phi}^T(x)\boldsymbol{\phi}(x)}. \quad (14)$$

If we use only terminal reward $R = \sum_{j=1}^{J} Q^{\pi}(x_j)$, then Equation (10) simplifies to reward weighted policy learning:

$$\boldsymbol{w}_{n+1}^{\nu} = \boldsymbol{w}_{n}^{\nu} + \frac{\sum_{i=1}^{m} (\boldsymbol{w}_{i}^{\nu} - \boldsymbol{w}_{n}^{\nu}) R_{i}^{\pi}}{\sum_{i=1}^{m} R_{i}^{\pi}}. \quad (15)$$

4 Experimental evaluation

The goal of our experiment was to optimize the velocity of the motion that would provide us clear images for the quality inspection. The sharper images we get, the more reward was given to the motion. All our experiments were performed on a UR10 robot and an industrial grade GigE camera - Basler acA1300-60gm (Figure 1). Camera resolution was set to 1282x1226 and frame-rate to 20 fps. In order to have stable and repeatable measurements, we need constant lighting conditions. That is the reason why we used additional dedicated LED lighting. Black line on its surface of the object (see Fig. 1 represents the object that needs to be inspected and the surrounding text represents a random background or noise. To make the problem for the optimization algorithm more difficult, we used an object with a sharp edge.

The path of the motion was obtained from the CAD model of the object so that at each point, the camera is perpendicular to the object being inspected. We decided to assign time dependency of the path points obtained from the CAD model so that the formed trajectory is minimal-jerk. This way, performed movements cause the least wear of the equipment and make the task of controllers following the trajectory easier. The biggest degradation of the focus measure appears at the points where the robot begins to change its orientation, so we decided that on that points the end-effector of the robot's velocity should be zero. Although this way we add a small deviation from the extracted path in the time domain, the relation between coordinates of the trajectory stays intact. Since the velocity profile of the minimal jerk trajectory is smooth, the focus measure does not have dramatically big changes or jumps. As the final step of the trajectory preparation, the calculated minimal-jerk trajectory was encoded in CDMPs. This initial minimal-jerk trajectory was set to last 10 s.

As the measure of image quality, we used squared horizontal gradient focus measure because it is one of the most reactive to the change in velocity. Reference focus measure was recorded by executing the initial minimal-jerk trajectory in 70 s. To do this we set all the velocity weights to $\boldsymbol{w}_k = 7$ and consequently the trajectory was linearly slowed down (from initial 10 s to 70 s). This represents our quasi-static focus measure, and the goal of the algorithm was to try to adapt the velocity profile weights $\boldsymbol{w}_k$ in order to reach as close as possible to this focus-measure by performing the trajectory in the shortest possible time. The initial condition for our learning process was minimal-jerk trajectory lasting 25 s (all $\boldsymbol{w}_k = 2.5$).

We tested three different versions of PoWER and compared their results. First, we tested PoWER without any of simplifications. In this case, the overlapping of CDMP weights is taken into account when using the update rule, given by Equations (10), (11), and (12). Besides this, we also varied covariance matrix as shown in Equation (13). In the second set of experiments, the covariance matrix is chosen to be diagonal and constant so that the weighting vector in Equation (10) was set as in Equation (14). In this experiments, the overlapping of weights is still taken into the account. In the third set of experiments, we used only the terminal reward and did not take into the account overlapping of weights. This way, PoWER algorithm was simplified to reward weighted policy learning as in Equation (15). All the aforementioned experiments were performed under the same conditions and with the same initial search noise set to $\boldsymbol{\varepsilon} = diag\{1.5\}$. The length of the importance sampler was chosen to be $m = 3$. All three sets of experiments were performed 5 times and lasted 50 iterations (updates of the weights).

In order to test the influence of the reward, we repeated the experiments with full implementation of PoWER and the implementation that takes into account overlapping of the weights with using the accumulated instead of the intermediate reward. As the initial search noise is crucial for the algorithm not to get stuck in local minimum, we also repeated the experiment with the full implementation of PoWER but using twice bigger initial search noise, i.e., $\boldsymbol{\varepsilon} = diag\{3\}$.

5 Results

Figure 2 shows the camera focus measure obtained when inspecting the object. Execution times through the experiments and their average as well as an average error compared to the reference focus measure are given in Table 1.

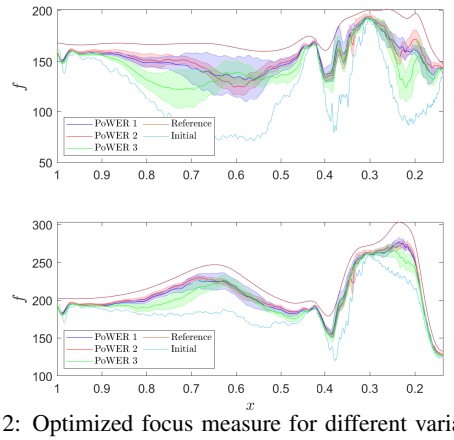


Figure 2: Optimized focus measure for different variations of PoWER: Top line shows our reference focus measure obtained by slowly moving the robot (quasi-static trajectory lasting 70 s). The bottom line is starting focus measure for all our learning algorithms (recorded with the minimal jerk trajectory lasting 25 s). In between are the final results of all variations of the algorithm. As all the experiments were repeated 5 times, shaded area on the graphs shows the spreading of the results.

The simplest version of PoWER that takes only the terminal reward into account and does not take into account overlapping of weights is labeled as PoWER 1. The one that takes into the account overlapping of weights and has exponentially fading exploration noise is labeled

as PoWER 2. The full implementation that additionally calculates the full covariance matrix to obtain the exploration noise is labeled as PoWER 3.

Test iteration	PoWER 1	PoWER 2	PoWER 3
1	46.23 s	50.94 s	46.97 s
2	47.97 s	48.82 s	46.28 s
3	49.17 s	51.34 s	41.89 s
4	45.49 s	54.19 s	41.58 s
5	52.26 s	46.34 s	41.54 s
Average	48.22 s	50.32 s	43.65 s
Error	8.96 %	8.27 %	13.3 %

Table 1: Time of trajectory execution over learning iterations

Figure 3 and Table 2 show the same data for the situation when accumulated reward is used instead of the intermediate reward.

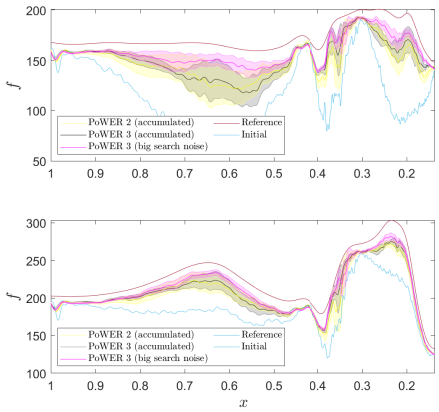


Figure 3: Optimized focus measure using accumulated reward and big search noise (check the description of Figure 2)

Test iteration	PoWER 2 acc. reward	PoWER 3 acc. reward	PoWER 3 big ϵ
1	49.37 s	46.12 s	57.37 s
2	49.49 s	43.21 s	58.59 s
3	51.09 s	49.04 s	55.90 s
4	48.77 s	42.95 s	58.70 s
5	44.13 s	44.09 s	58.84 s
Average	48.57s	45.08 s	57.88 s
Error	9.51 %	10.99 %	5.26 %

Table 2: Time of trajectory execution over learning iterations for accumulated reward and big search noise

Results show that the algorithm that used intermediate rewards and took into account the overlapping of the weights achieved only slightly better results in the terms of focus than the simplest version of the algorithm. However, the execution time of the trajectories was slower with terminal reward only because convergence was faster. The full version of PoWER achieved the worst results in the terms of focus measure because its adaptation of exploration noise makes the update steps smaller and consequently the time of execution stays faster. However, when we tested the same algorithm with double initial search noise, it outperformed all other simplified algorithms in the terms of focus measure. For this use-case, the difference between using intermediate or accumulated reward was small. This was expected because in this case, the rewards obtained in the beginning are equally important as the ones at the end.

6 Conclusion

Our result show that for the purpose of velocity optimization, the most crucial parameter that determines the trade-off between accuracy and the convergence speed is the search noise. We have shown that for some use-cases, the simplified version of the algorithm is sufficient for obtaining a good results. We have also showed that each simplification of the algorithm requires different initial settings.

PoWER algorithm has a great potential in optimization tasks. However, it would require a lot of iterations until it finds the optimal solution and it would extend the time needed to set up new production line. The main reason for that is a too big search space. That is why in the future we plan to use algorithms such as Iterative Learning Control to reduce the search space and only afterwards to use RL to fine tune the performance.

Acknowledgement: The research leading to these results has received partial funding from the Horizon 2020 RIA Programme grant 820767 CoLLaboratE.

References

[1] A. Gams, S. Reberšek, B. Nemec, J. Škrabar, R. Krhlikar, J. Skvarč, and A. Ude, "Robotic learning for increased productivity: Autonomously improving speed of robotic visual quality inspection," in *2019 IEEE 15th International Conference on Automation Science and Engineering*, pp. 1275–1281, 2019.

[2] O. Semeniuta, S. Dransfeld, and P. Falkman, "Vision-based robotic system for picking and inspection of small automotive components," in *2016 IEEE International Conference on Automation Science and Engineering*, pp. 549–554, 08 2016.

[3] J. Kober and J. Peters, "Policy search for motor primitives in robotics," *Mach. Learn.*, vol. 84, p. 171–203, July 2011.

[4] A. Ude, B. Nemec, T. Petrič, and J. Morimoto, "Orientation in cartesian space dynamic movement primitives," in *IEEE Int. Conference on Robotics and Automation (ICRA)*, pp. 2997–3004, 2014.

[5] D. Racki, D. Tomazevic, and D. Skocaj, "A compact convolutional neural network for textured surface anomaly detection," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1331–1339, March 2018.

[6] T. Ivanovska, S. Reich, R. Bevec, Z. Gosar, M. Tamosiunaite, A. Ude, and F. Wörgötter, "Visual inspection and error detection in a reconfigurable robot workcell: An automotive light assembly example," in *VISIGRAPP*, 2018.

[7] A. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: Learning attractor models for motor behaviors," *Neural computation*, vol. 25, 11 2012.

[8] A. Ude, A. Gams, T. Asfour, and J. Morimoto, "Task-specific generalization of discrete and periodic dynamic movement primitives," *IEEE Transactions on Robotics*, vol. 26, pp. 800–815, Oct 2010.

[9] S. Calinon, P. Kormushev, and D. Caldwell, "Compliant skills acquisition and multi-optima policy search with em-based reinforcement learning," *Robotics and Autonomous Systems*, vol. 61, p. 369–379, 04 2013.

Turingov test zaznave gibalne sposobnosti pri fizičnem sodelovanju med človekom in inteligentnim robotskim agentom

Rebeka Kropivšek Leskovar[1,2], **Tadej Petrič**[1,2]

[1]*Odsek za avtomatiko, biokibernetiko in robotiko, Inštitut "Jožef Stefan", Jamova cesta 39, 1000 Ljubljana*
[2]*Mednarodna podiplomska šola Jožefa Stefana, Jamova cesta 39, 1000 Ljubljana*
E-pošta: {rebeka.leskovar, tadej.petric}@ijs.si

Turing test of motor ability perception in physical collaboration between a human and an intelligent robot agent

In this paper we propose a novel robot control method for human-robot collaboration tasks that takes into account the leader-follower relationship found in human interactions. Taking into account the leader-follower dynamics, learnt during a study on human-human collaboration, the control method replicates human behaviour when performing collaborative tasks. The performance of the proposed control method was evaluated using a 2D reaching task where we compared task performance between individual tasks, tasks in collaboration with a human and tasks in collaboration with a robot. The subjects in the evaluation were asked to grade their perceived task load for each experiment as well as specify if they thought they performed the task alone, with a robot or with a human partner as a Turing test to determine whether the subjects were able to distinct between a robot and a human partner. The results of the evaluation showed, that the robot control method is capable of replicating human behavior to benefit overall task performance of the subject in collaboration, however it is not capable of replicating this behaviour to the degree that the subject in collaboration would not be able distinct whether they were collaborating with a robot or a human partner.

1 Uvod

V zadnjem desetletju pridobiva raziskovalno področje interakcij med človekom in robotom veliko zanimanja z raziskavami, kot so [1][2][3]. Da bi bolje razumeli interakcije med človekom in robotom, je potrebno prvo razumeti interakcije med ljudmi. Študija Ganesha in drugih [4] je pokazala, da se ljudje v izvajanju naloge izboljšajo, če sodelujejo s partnerjem. Poleg tega so ugotovili, da se človek izboljša bolj, če sodeluje s človekom kot, če sodeluje z umetnim agentom (npr. robotom). Iz tega lahko izhajamo, da je za boljše sodelovanje med človekom in robotom potrebno razviti sistem vodenja robota, ki implementira človeku podobna obnašanja.

V dosedanjih študijah [5][6] smo se človeškemu obnašanju poskušali približati z implemetacijo že obstoječih nevromehanskih modelov človeka v sisteme vodenja. V omenjenih študijah je bila naloga robota, da sledi gibanju človeka in prilagodi svoje obnašanje tako, da izboljša njegovo uspešnost. Tako je človek med izvajanjem naloge vedno prevzel vlogo vodje. V primerjavi s tem se v interakcijah med ljudmi vodjo določi med izvajanjem naloge. Da bi lahko posnemali razmerje vodja-spremljevalec, ki se pojavi v sodelovanju med ljudmi, potrebujemo v interakcijah med človekom in robotom sistem vodenja robota, ki poleg stereotipno človeškega obnašanja vzame v zakup tudi razmerje vodja-spremljevalec.

V sklopu tega prispevka bomo predstavili nov sistem vodenja za kolaborativne robote, ki poleg človeku podobnega gibanja med interakcijo s človekom upošteva tudi strategijo vodja-spremljevalec, ki se naravno pojavi v sodelovanju med ljudmi. Predlagan sistem bomo prav tako ovrednotili na podlagi uspešnosti človeka v sodelovanju s predlaganim sistemom vodenja robota v primerjavi z uspešnostjo človeka, ko nalogo opravlja sam ter z uspešnostjo človeka, ko sodeluje s človeškim partnerjem. Pri tem predpostavljamo, da bo predlagan sistem vodenja izboljšal uspešnost človeka s katerim sodeluje glede na njegovo individualno uspešnost. Prav tako predpostavljamo, da bo uspešnost kolaborativne naloge približno enaka v primeru sodelovanja z robotom in sodelovanja s človeškim partnerjem. Za konec bomo s pomočjo preprostega Turnigovega testa ovrednotili ali človek prepozna razliko med sodelovanjem s človeškim ali robotskim partnerjem.

2 Sistem vodenja robota

Predlagani sistem vodenja temelji na dinamičnem modelu, ki smo ga razvili v sklopu članka [7]. Le-ta je sestavljen iz dveh točk, ki predstavljata končna člena robotskih rok in sta med seboj povezani z virtualno vzmetjo. Pri tem je koeficient virtualne vzmeti nastavljen tako, da vzmet posnema palico, zato lahko silo vzmeti v sistemu zanemarimo. Dinamični model generira silo F v središču vzmeti glede na sili, ki sta aplicirani na končna člena robotskih rok. Tako dobimo enačbo:

$$\boldsymbol{F} = \boldsymbol{F_1} + \boldsymbol{F_2} \tag{1}$$

kjer je F_1 sila na prvem končnem členu in F_2 sila na drugem končnem členu.

Zgoraj opisani dinamični model omogoča povezavo dveh robotskih rok v primerih, ko med seboj sodelujeta dve osebi, uporablja pa se lahko tudi za individu-

alno izvajanje nalog, pri čemer se robotski roki med seboj razčlenita tako, da upoštevamo silo na drugem končnem členu kot 0.

Da lahko uporabimo zgoraj opisani dinamični model v primerih, ko človek sodeluje z robotskim agentom, je potrebno enačbo 1 zapisati kot:

$$\boldsymbol{F} = \boldsymbol{F_h} + \boldsymbol{F_r} \tag{2}$$

kjer je F_h sila, ki jo človek aplicira na končno točko robota, F_r pa sila, ki jo proizvede robotski agent.

Sila F_r je v predlaganem sistemu vodenja sestavljena iz dveh delov in sicer:

$$\boldsymbol{F_r} = K_l \cdot (\boldsymbol{F_{r,fb}} + \boldsymbol{F_{r,ff}}) \tag{3}$$

kjer je $\boldsymbol{F_{r,ff}}$ sila, naučena iz povprečne trajektorije sil posameznega subjekta s katerim sodeluje, $\boldsymbol{F_{r,fb}}$ pa je sila, ki temelji na povratni zanki in je odvisna od napake med željeno in dejansko pozicijo robota v danem času. Sledeča je definirana kot:

$$\boldsymbol{F_{r,fb}} = K_p \cdot (\boldsymbol{p_{dej}} - \boldsymbol{p_t}) + K_d \cdot (\boldsymbol{v_{dej}} - \boldsymbol{v_t}) \tag{4}$$

kjer sta K_p in K_d koeficienta ojačanja, $\boldsymbol{p_{dej}}$ dejanska pozicija in $\boldsymbol{p_t}$ željena končna pozicija ter $\boldsymbol{v_{dej}}$ dejanska hitrost in $\boldsymbol{v_t}$ željena končna hitrost, ki je enaka 0.

Silo $\boldsymbol{F_{r,ff}}$ se robot nauči tako, da se najprej nauči karakteristično trajektorijo pomika skozi fazo naloge $\boldsymbol{p}(\phi)$, ko subjekt izvaja nalogo individualno. Iz naučene trajektorije pomika, nato izračuna silo kot:

$$\boldsymbol{F_{r,ff}}(\phi) = m\ddot{\boldsymbol{p}}(\phi) + D\ddot{\boldsymbol{p}}(\phi) \tag{5}$$

pri čimer je m masa dinamičnega sistema in D faktor dušenja dinamičnega sistema.

K_l predstavlja vodilno konstanto, ki določi kolikšen vpliv ima virtualna sila robota na celotno gibanje dinamičnega modela. Vodilna konstanta je odvisna od zadane naloge in je definirana na podlagi rezultatov raziskave o sodelovanju med ljudmi, katere eksperiment je predstavljen v [7]. V tej raziskavi, je bila zadana naloga subjektov doseči devet različnih tarč, medtem ko so bili med seboj povezani v pare. V sklopu raziskave smo ugotovili, da subjekta v sodelovanju ne vložita enako sile v zadano nalogo, temveč eden izmed subjektov vedno vloži večjo silo in posledično vodi opravljanje naloge. Kakšna je proporcionalna razlika med vloženima silama posameznega subjekta smo definirali kot:

$$L = \int_0^T \boldsymbol{F_f}(t) - \boldsymbol{F_s}(t)dt \tag{6}$$

kjer je $\boldsymbol{F_f}$ sila hitrejšega subjekta, $\boldsymbol{F_s}$ pa sila počasnejšega subjekta, pri čimer sta bila hitrejši in počasnejši subjekt določena glede na njuno hitrost izvajanja naloge, ko sta naloge opravljala sama.

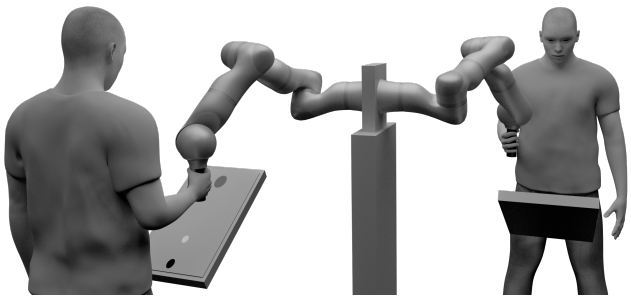
Da bi lahko robot opravljal vlogo tako partnerja, ki vodi kot partnerja, ki sledi, smo K_l glede na željeno vlogo definirali kot:

$$\Delta K_l = \begin{cases} 1 - \frac{L}{max(L)} & \text{; robot sledi človeku} \\ 1 & \text{; človek in robot sta ekvivalentna} \\ 1 + \frac{L}{max(L)} & \text{; robot vodi človeka} \end{cases} \tag{7}$$

Tako koeficient vodenja K_l kot sili $F_{r,ff}$ in $F_{r,fb}$ so odvisni od končnega cilja naloge. Zaradi tega predlagan sistem vodenja pred začetkom izvajanja naloge najprej oceni, kaj je cilj naloge, ter cilju primerno prilagodi svojo strategijo izvajanja naloge.

3 Vrednotenje sistema vodenja

Za vrednotenje sistema vodenja smo izvedli štiri vrste eksperimentov in sicer individualen eksperiment, kolaborativni eksperiment z robotom, kjer robot sledi subjektu, kolaborativni eksperiment z robotom, kjer robot vodi subjekta ter kolaborativni eksperiment med dvema subjektoma. V vrednotenju je sodelovalo 12 subjektov, oziroma 6 parov, pri čemer je bilo 8 moških in 4 ženski subjekti. Vsi subjekti so bili desničarji.



Slika 1: 3D prikaz eksperimentalne postavitve uporabljene za vrednotenje krmilnika. Na ekranu je prikazan grafični vmesnik, ki se je uporabljal v eksperimentu. Le-ta prikazuje začetno pozicijo (črna točka), tarčo (siva točka) in točko, ki jo subjekta premikata s pomočjo haptičnega vmesnika (bela točka).

3.1 Postavitev eksperimenta

Eksperiment se je izvajal na dveh robotskih rokah Kuka LWR, ki sta služili kot haptični vmesnik med človekom in virtualnim okoljem. Za prikaz virtualnega okolja sta bila uporabljena dva računalniška ekrana, na katerih je bil prikazan preprost grafični vmesnik, predstavljen na sliki 1.

3.2 Postopek eksperimenta

Subjekta sta se postavila pred računalniški zaslon in prijela ročaj haptičnega vmesnika, kot je prikazano na sliki 1. Eksperiment se je začel, ko sta se subjekta s haptičnim vmesnikom pomaknila v začetno točko. Ko sta bila oba subjekta v začetni točki, se je na grafičnem vmesniku prikazala tarča. Subjekta sta imela nalogo, da zadeneta tarčo v najkrajšem možnem času ter ostaneta v tarči dokler leta ne izgine iz zaslona. Ko je tarča izginila, sta se morala vrniti nazaj v začetno točko ter tam počakati, dokler se ni pojavila nova tarča.

Vsak eksperiment je trajal približno 10 minut in je bil sestavljen iz pet ciklov. V sklopu vsakega cikla so morali subjekti doseči 9 različnih tarč, ki so se na ekranu pojavljale v naključnem zaporedju. Vsak eksperiment je imel v celoti 45 ponovitev. Da bi se izognili vplivu učenja naloge na končni rezultat med posameznimi eksperimenti, smo za vsak par spremenili vrstni red izvajanja eksperimentov, pri čemer je bil prvi eksperiment vedno individualen zaradi učenja trajektorij robotskega agenta.

Med posameznim eksperimentom so imeli subjekti 5 min pavze, v katerih so za vsak eksperiment posebej izpolnili *Nasa Task Load index* oziroma Nasa-TLX vprašalnik, ki je namenjen ocenjevanju zaznane obremenitve pri izvajanju nalog [8]. Kot Turingov test so morali subjekti, po izvedbi vseh štirih eksperimentov, odgovoriti še na dva vprašanja in sicer:

1. *če bi izbrali enega izmed eksperimentov, kateri se vam je zdel najlažji za izvesti?*

2. *Za posamezni eksperiment izberite ali ste eksperiment opravljali sami, s človekom ali z robotom.*

3.3 Analiza podatkov

Poleg subjektivnega vrednotenja na podlagi Nasa-TLX, smo uspešnost izvajanja naloge v posameznih eksperimentih vrednotili še na podlagi Fittsovega zakona (ang.*Fitts' law*) [9]. Fittsov zakon opisuje razmerje med hitrostjo in natančnostjo človeka pri izvajanju nalog, kjer je potrebno doseči določeno tarčo. V sklopu tega zakona je uspešnost izvajanja naloge za različne tarče opisana z vrednostjo IP oziroma *index of performance*. Le-ta je definiran kot:

$$IP = \frac{ID}{T} \tag{8}$$

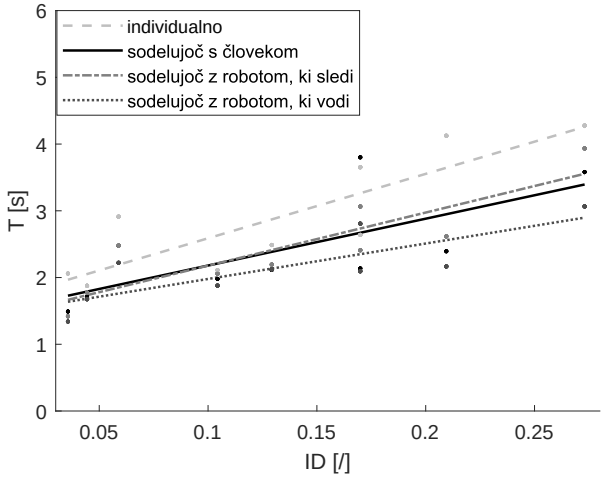
kjer je T merjen čas za doseg določene tarče in ID *index of difficulty*, ki definira zahtevnost tarče.

V literaturi je ID definiran na različne načine, v našem primeru pa smo uporabili Shannonovo formulacijo, ki je definirana kot:

$$ID = log_2(\frac{D}{W} + 1) \tag{9}$$

V tej enačbi predstavlja D razdaljo tarče, W pa velikost tarče.

4 Rezultati



Slika 2: Primerjava modelov Fittsovega zakona za posamezne eksperimente.

Na sliki 2 so prikazani modeli po Fittsovem zakonu za posamezne tipe eksperimentov. Nižji čas izvedene naloge predstavlja boljšo učinkovitost izvajanja naloge. Iz slike lahko vidimo, da se je uspešnost pri vseh eksperimentih, kjer je subjekt sodeloval z robotom ali človekom, izboljšala v primerjavi z individualnim eksperimentom.

Poleg tega lahko vidimo, da je učinkovitost izvajana naloge bila najboljša v eksperimentih, kjer je subjekt sodeloval z robotom, ki je vodil, medtem ko je bila učinkovitost v eksperimentih, kjer je subjekt sodeloval z drugim človekom in z robotom, ki je sledil subjektu, v povprečju približno enaka.

Če si ogledamo ocene pridobljene z Nasa-TLX vprašalnikom, ki so prikazane na sliki 3, lahko vidimo, da je skupno najboljšo oceno dobil eksperiment, v katerem je subjekt izvajal nalogo v sodelovanju z robotom, ki je vodil. To potrjujejo tudi odgovori subjektov na vprašanje 1, kjer je 7 od 12 subjektov odgovorilo, da jim je bil najlažji eksperiment, kjer so sodelovali z robotom, ki vodi. Pri tem sta 2 od 12 subjektov odgovorila z individualnim eksperimentom ter eksperimentom, kjer so sodelovali s človekom, 1 od 12 subjektov pa je odgovoril z eksperimentom, kjer so sodelovali z robotom, ki je sledil.

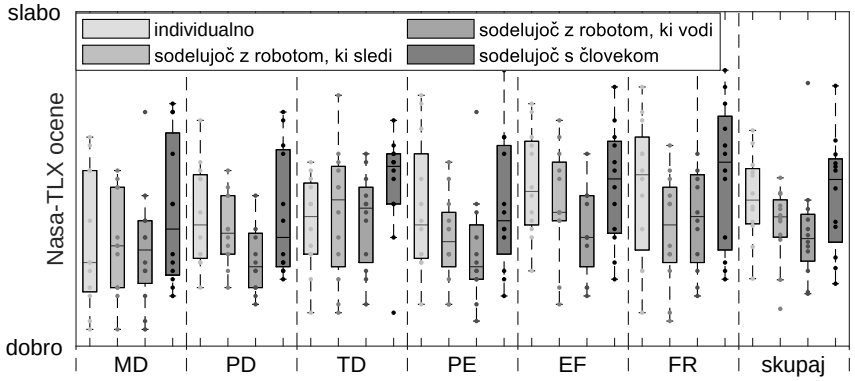
Poleg tega lahko opazimo, da ima eksperiment, kjer so subjekti sodelovali s človekom največjo razpršenost ocen skozi vse faktorje razen časovne zahtevnosti (ang. *temporal demand*), medtem ko imata eksperimenta, kjer je subjekt sodeloval z robotom najnižjo razpršenost skozi vse faktorje razen časovne zahtevnosti.

Pri vprašanju 2 je tipe eksperimentov pravilno določilo 9 od 10 subjektov za individualni eksperiment, 4 od 10 za eksperiment, kjer so sodelovali z robotom, ki jim je sledil, 7 od 10 za eksperiment, kjer so sodelovali z robotom, ki je vodil in 9 od 10 za eksperiment, kjer so sodelovali s človekom. Dva od desetih subjektov je eksperiment, kjer so sodelovali z robotom, ki vodi zamenjalo za eksperiment, kjer so sodelovali s človekom in en za individualen eksperiment. Eksperiment, kjer so sodelovali z robotom, ki je subjektom sledil, pa je 5 od 10 subjektov zamenjalo za individualen eksperiment in 1 od 10 za sodelovanje s človekom. Pri tem smo 2 od 12 subjektov zanemarili zaradi nepopolnega odgovora na vprašanje.

5 Diskusija

Namen raziskave je bilo razviti nov sistem vodenja kolaborativnega robota, ki vzame poleg človeku podobnega gibanja v zakup tudi strategijo vodja-spremljevalec in omogoča zamenjavo človeškega partnerja v kolaborativnih nalogah s človeku podobnim obnašanjem robota. Predlagan sistem je bil preizkušen na preprosti nalogi, kjer je moral subjekt doseči tarčo in sicer na štiri različne načine – sam, s pomočjo robota, ki je sledil subjektu, s pomočjo robota, ki je vodil subjekta in s pomočjo človeškega partnerja.

Analiza učinkovitosti izvajanja naloge, ki smo jo določili na podlagi Fittsovega zakona, je pokazala, da se je subjektom med izvajanjem naloge v sodelovanju s tako robotskim kot človeškim partnerjem učinkovitost izboljšala, kar se ujema z rezultati, prikazanimi v [4] in [7]. Iz tega lahko sklepamo, da predlagan sistem vodenja ne ovira subjekta pri izvajanju naloge, temveč celo dodatno



Slika 3: Nasa-TLX ocene vseh subjektov. Ocene so predstavljene za vse faktorje posebej (MD: mentalna zahtevnost (ang. *mental demand*), PD: fizična zahtevnost (ang. *physical demand*), TD: časovna zahtevnost (ang. *temporal demand*), PE: učinkovitost (ang. *performance*), EF: napor (ang. *effort*), FR: frustracija (ang. *frustration*)) ter kot skupna ocena (tj. povprečna ocena vseh faktorjev z obteženimi ocenami glede na vsakega posameznika).

pripomore k učinkovitosti.

Pri kolaborativnih nalogah pa je poleg učinkovitosti pomembna tudi obremenitev subjekta pri izvajanju naloge. Le-to smo vrednotili s pomočjo Nasa-TLX vprašalnika, v katerem je najboljšo oceno dobil eksperiment, kjer so subjekti sodelovali z robotom, ki jih je vodil, najslabšo oceno pa eksperiment, kjer so sodelovali s človeškim partnerjem. Razlog za to izhaja predvsem iz uteži posameznih faktorjev, ki se jih upošteva med računanjem skupne ocene TLX. Uteži posameznih faktorjev določi vsak subjekt posebej glede na to, kateri faktor se mu zdi najbolj pomemben pri izvajanju naloge [8]. Pri izbiri uteži so v povprečju subjekti za najbolj pomemben faktor izbrali učinkovitost (PE) in frustracijo (FR). Iz slike 3 lahko vidimo, da ima eksperiment, kjer so subjekti sodelovali s človeškim partnerjem, najslabšo oceno tako v PE kategoriji kot v FR kategoriji, medtem ko ima eksperiment z robotom, ki vodi, najboljšo oceno v PE kategoriji in drugo najboljšo oceno v FR kategoriji. Razlaga za prikazane ocene izhaja predvsem iz tega, da je človek manj predvidljiv, kot robot s predlaganim sistemom vodenja, kar povzroči večjo frustracijo ter mentalno breme subjekta in posledično zmanjša občutek učinkovitosti izvajanja naloge.

Iz odgovorov na vprašanja Turingovega testa lahko ugotovimo, da je večina subjektov pravilno ugotovila ali so nalogo izvajali sami, z robotom ali s človeškim partnerjem, pri čimer je bil izjema eksperiment, kjer je subjekt sodeloval z robotom, ki sledi. Iz tega lahko sklepamo, da sistem vodenja, predlagan v tem članku, ne generira človeku dovolj podobnega obnašanja, da bi prestal Turingov test. V primeru sistema vodenja, kjer robot vodi je namreč bilo obnašanje dovolj drugačno od človeškega partnerja, da sta ga za človeka zamenjala le 2 od 10 subjektov, v primeru sistema vodenja, kjer je robot sledil subjektu, pa ga je za človeka zamenjal 1 od 10 subjektov, pri čemer je 5 od 10 subjektov mislilo, da so nalogo izvajali sami.

Študija predstavljena v tem članku nakazuje, da je predlagan sistem vodenja robota sposoben nadomestiti človeškega partnerja, saj se učinkovitost izvajanja naloge v primeru sodelovanja z robotom, tako kot pri sodelovanju s človeškim partnerjem, izboljša. Kljub temu pa predlagan sistem vodenja robota ni sposoben oponašati človeškega obnašanja v kolaborativnih nalogah do te mere, da ne bi bil prepoznaven v primerjavi s človekom, saj je njegovo obnašanje tipično drugačno kot obnašanje človeškega partnerja, subjekt pa ga je zmožen razločiti.

Acknowledgement: This work was supported by Slovenian Research Agency grant N2-0130.

Literatura

[1] S. Ikemoto, H. B. Amor, T. Minato, B. Jung, and H. Ishiguro, "Physical human-robot interaction: Mutual learning and adaptation," *IEEE Robotics Automation Magazine*, vol. 19, no. 4, pp. 24–35, 2012.

[2] C. Thomas, L. Stankiewicz, A. Grötsch, S. Wischniewski, J. Deuse, and B. Kuhlenkötter, "Intuitive work assistance by reciprocal human-robot interaction in the subject area of direct human-robot collaboration," *Procedia CIRP*, vol. 44, pp. 275 – 280, 2016, 6th CIRP Conference on Assembly Technologies and Systems (CATS).

[3] X. Yu, Y. Li, S. Zhang, C. Xue, and Y. Wang, "Estimation of human impedance and motion intention for constrained human–robot interaction," *Neurocomputing*, 2019.

[4] G. Ganesh, A. Takagi, R. Osu, T. Yoshioka, M. Kawato, and E. Burdet, "Two is better than one: Physical interactions improve motor performance in humans," *Scientific Reports*, vol. 4, no. 1, p. 3824, 2014.

[5] T. Petrič, R. Goljat, and J. Babič, "Cooperative human-robot control based on fitts' law," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 345–350.

[6] T. Petrič, M. Cevzar, and J. Babič, "Utilizing speed-accuracy trade-off models for human-robot coadaptation during cooperative groove fitting task," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 2017, pp. 107–112.

[7] R. Kropivšek Leskovar, J. Čamernik, and T. Petrič, "Dyadic human-human interactions in reaching tasks: Fitts' law for two," in *Advances in Service and Industrial Robotics*. Cham: Springer International Publishing, 2020, pp. 199–207.

[8] S. G. HART, "Development of nasa-tlx (task load index) : Results of empirical and theoretical research," *Human Mental Workload*, 1988. [Online]. Available: https://ci.nii.ac.jp/naid/10015399884/en/

[9] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Experimental Psychology*, vol. 47, no. 6, pp. 381–391, 1954.

Ocenjevanje dvoročne vadbe z robotom

Ana Mandeljc[1], Eva Čebašek[1], Marko Munih[1], Janez Podobnik[1], Matjaž Mihelj[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: ana.mandeljc@fe.uni-lj.si

Evaluation of two-handed exercise with a robot

Abstract. *The aim of our study was to evaluate bimanual movements and to calculate the parameters of movement coordination using a two-handed gripper and pressure sensors on a robotic system. The research involved designing a measurement protocol on a HapticMaster robot, executing the measurements and analyzing the forces of interaction between the upper limbs. While designing the measurement protocol, we had to consider different values of damping. Lastly, we tested our measurement protocol on a sample of post-stroke patients, evaluated the interaction parameters and compared the results with a control group of healthy subjects.*

1 Uvod

Možganska kap letno prizadene vsaj 0,2% populacije ter več kot 1% ljudi, starejših od 65 let. Med preživelimi po kapi jih je vsaj polovica trajno onesposobljenih, zato je možganska kap vodilni vzrok za zmanjšanje motoričnih, kognitivnih, govornih ter socialnih zmožnosti − tako pri nas, kot po svetu [**1**].

V zgodnjem obdobju po možganski kapi je pomembna čimprejšnja rehabilitacija, s katero pacient nadaljuje tudi po odhodu iz bolnišnice. Zaradi različnih stopenj telesne prizadetosti, pacienti že takoj po nastopu akutne faze pričnejo z ustrezno rehabilitacijo [**1**]. Z napredkom v robotiki so terapevti v proces rehabilitacije začeli vključevati tudi robote, ki jih uporabljajo predvsem kot vadbene pripomočke, s katerimi lahko pacientom pomagajo pri hitrejšem doseganju ciljev fizikalne rehabilitacije.

Ker je velika večina vsakodnevnih dejavnosti dvoročnih, je za njihovo izvedbo potrebno koordinirano gibanje obeh zgornjih udov − načeloma gre za koordinirano potiskanje ali vlečenje obeh rok ali pa ena roka potiska in druga vleče. Med izvajanjem takšnih nalog ljudje težimo k simetriji gibanja zgornjih udov, ki jih med takšnim gibanjem nadzorujemo kot eno funkcijsko enoto. Pacienti med dvoročno vadbo z zdravim udom razgibavajo, oziroma vodijo okvarjen ud, ter kasneje začnejo z aktivnim gibanjem okvarjenega uda. Pri tem je pomembno, da v tem večji meri poskušajo nalogo opraviti z okvarjenim udom, z zdravim pa si pomagajo le toliko, da nalogo uspešno opravijo. Takšna vadba izboljša koordinacijo med udoma, moč prijema in ostale funkcijske sposobnosti okvarjenega uda. Pozitivni učinki dvoročne vadbe se prenesejo tudi na enoročno izvajanje nalog s prizadetim udom [**2**].

Poleg vadbe za izboljšanje motoričnih sposobnosti, povečanja delovnega prostora ter izboljšanja spretnosti okvarjenega uda, uporaba rehabilitacijskih robotov s senzorji pozicij in sil omogoča tudi sledenje različnim parametrom gibanja, kot so sile, navori, hitrosti, pospeški in podobno. To omogoča objektivno ocenjevanje pacientovih gibalnih sposobnosti ter napredka tekom okrevanja.

Cilj članka je bila izvedba enostavnih dvoročnih gibalnih nalog z robotskim sistemom ter obdelava in analiza surovih signalov. S pomočjo meritev smo želeli objektivno oceniti nekatere izmed parametrov ocenjevanja gibanja in ugotoviti, kako se njihove vrednosti razlikujejo med zdravimi osebami ter pacienti po možganski kapi.

2 Metodologija

2.1 Preiskovanci

Za potrebe meritev smo oblikovali skupino zdravih preiskovancev, ki je služila za referenco, ter skupino pacientov po možganski kapi. V skupino zdravih preiskovancev je bilo vključenih 26 oseb, od tega 7 žensk in 19 moških, povprečne starosti 37,6 $\pm$ 16,8 let. Skupino pacientov po kapi je na začetku sestavljalo 16 oseb, a smo v analizo meritev, za potrebe tega članka, vključili le osebe z višjo okvaro desnega uda. Analizirano skupino pacientov po kapi je tako sestavljalo 7 pacientov, od tega 3 ženske in 4 moški, povprečne starosti 61,7 $\pm$ 6,6 let.

2.2 Merilni sistem

Strojno opremo merilnega sistema predstavlja haptični robot HapticMaster, proizvajalca »FSC Control System«. Ima tri prostostne stopnje − rotacijo in translacijo v osi z ter translacijo v osi x. Je admitančno voden − ko na vrh robota delujemo s silo, se robot odzove s premikom. Na vrh robota smo pritrdili dvoročno držalo, razvito v okviru magistrske naloge [**3**], ki sistemu doda še dve prostostni stopnji − rotacijo vrha robota okoli vertikalne in horizontalne osi. Držalo ima na vsaki ročki tudi senzor sil in navorov − 50M31, proizvajalca JR3 Inc., ki omogoča merjenje sile v interakciji uporabnik-robot.

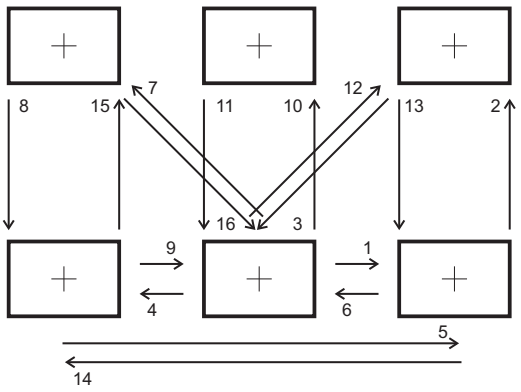
Kljub temu, da z držalom razširjen sistem omogoča gibanje v 5ih prostostnih stopnjah, smo za potrebe naše študije aktivno uporabljali le 3 − translacijo ter rotacijo v

osi z ter vrtenje držala okoli horizontalne osi. Gibanje v preostalih prostostnih stopnjah smo v primeru translacije v osi x onemogočili s pomočjo navidezne stene, gibanje držala okoli vertikalne osi pa smo omejili z mehansko blokado.

S pomočjo programskih paketov Matlab in Simulink ter okolja xPC Target, proizvajalca The MathWorks Inc., je bila razvita programska oprema, potrebna za izvedbo vodenja robota. Za vizualizacijo navideznega okolja smo uporabili programski paket Unity 3D, proizvajalca Unity Technologies. Uporabljeni sistem omogoča avtomatski zajem in obdelavo podatkov meritev, uporablja pa tudi navidezno okolje z vizualno povratno zanko, ki uporabniku posreduje informacijo o poteku izvedbe naloge. Programska oprema in navidezno okolje sta bila razvita v sklopu magistrske naloge [**4**], tako, da je bilo potrebno za našo študijo definirati le novo gibalno nalogo, določiti število ponovitev ter vrednosti navideznega dušenja b in v sistemu vodenja dodati navidezno steno za zamejitev gibanja vzdolž osi x.

2.3 Opis naloge in navideznega okolja

Nalogo, ki so jo preiskovanci izvajali na robotskem sistemu, smo zasnovali za koordinirano gibanje rok v frontalni ravnini. Cilj naloge je bil voditi vrh robota tako, da se s sledilnim objektom prekrije referenčni objekt, katerega položaji so predhodno določeni. Orientacija referenčnega objekta se ni spreminjala. Za postopno oteževanje izvajanja naloge smo na podlagi empiričnih opazovanj določili 3 stopnje navideznega dušenja b, pri katerih se je naloga opravljala – 0, 20 in 40 Ns/m.



Slika 1: Prikaz postavitve tarč in zaporedja gibov med njimi.

Spreminjanje lege referenčnega objekta prikazuje Slika 1, pri čemer zaporedje sprememb prikazujejo številke ob puščicah. Lega referenčnega objekta se spremeni, ko ga s sledilnim objektom primerno pokrijemo ter pokritost ohranjamo 3 s – kot primerna pokritost se smatra interval ujemanja položajev sledilnega in referenčnega objekta, z orientacijsko napako manjšo od 5° ter napako lege manj kot 2 cm. Naloga je bila sestavljena iz gibanj gor-dol, levo-desno ter diagonalno. Ker so bili gibi sestavljeni iz referenčnih začetnih in končnih položajev, smo lahko ocenili ponovljivost in primerjali različne parametre sil interakcije, tako pri posameznih vrednostih navideznega dušenja b, kot pri posameznem gibu.

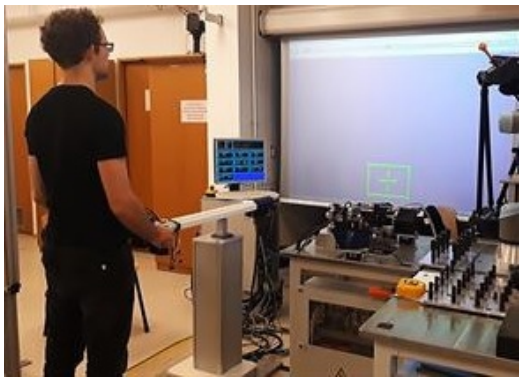
Navidezno okolje omogoča vizualno povratno informacijo in s tem olajša izvajanje ter samo razumevanje naloge, hkrati pa jo naredi tudi bolj zabavno, kar pomembno vpliva na povečanje ter vzdrževanje motivacije med izvajanjem naloge. V navideznem okolju je bil referenčni objekt prikazan kot belo, sledilni pa kot rumeno obarvan pravokotnik. Diagram prehajanja stanj virtualnega simulatorja pokrivanja tarč je prikazan na Sliki 3. Imamo seznam objektov, z določenim številom objektov O in ponovitev N. Ko objekt dosežemo, se sledilni objekt obarva zeleno, po vzdrževanju pokritosti 3 s pa se referenčni objekt preslika v novo lego, določeno na seznamu. Če smo že dosegli vse lege, program poveča število ponovitev za 1 ter znova začne s prikazovanjem lege referenčnega objekta od začetka seznama. Ko opravimo predvideno število ponovitev N, program zaključi z izvajanjem.

Pomembna spremenljivka diagrama prehajanja stanj je indikator pokritosti objekta – ko se referenčni objekt pokrije s sledilnim, se vrednost indikatorja v programu iz 0 spremeni v 1 (stanje 2 v diagramu), kar pri analizi meritev uporabimo kot pomoč pri segmentaciji signalov.

2.4 Potek meritev

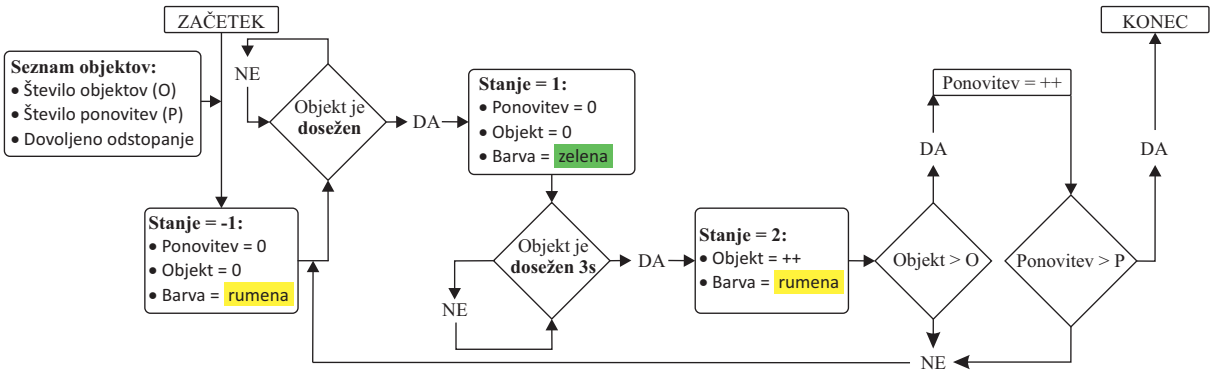
Preiskovanci so meritve opravljali v Laboratoriju za robotiko na Fakulteti za elektrotehniko, Univerze v Ljubljani. Pred začetkom meritev so vsi preiskovanci podpisali pristopno izjavo. Vsi preiskovanci, tako zdravi kot pacienti po kapi, so nalogo izvajali v stoječem položaju. Postavljeni so bili pred platno, velikosti 1,4 m x 1,4 m, na katerega smo projecirali navidezno okolje, ter pred robota, tako, da je bil center držala v liniji vertikalne telesne osi.

Razdalja med držalom in telesom je bila tolikšna, da so bile nadlakti merjencev ob telesu, ko je bil robot postavljen v začetni položaj (Slika 2). Pacientom, ki so imeli težave s prijemanjem ročke z okvarjenim udom, smo na dlan namestili opornico, ki je preprečevala zdrs roke z držala.



Slika 2: Prikaz merilnega sistema, navideznega okolja ter začetnega položaja merjenca.

Protokol meritev je vseboval 3 ponovitve naloge pri 3 različnih vrednostih navideznega dušenja b – v zaporedju 0, 20 in 40 Ns/m, kar pomeni 48 gibov pri posameznem dušenju b in 144 gibov tekom celotne meritve. Po izvedbi naloge pri posameznem dušenju b, so preiskovanci, po potrebi, naredili nekaj minutni premor. Izvedba



Slika 3: Diagram prehajanja stanj.

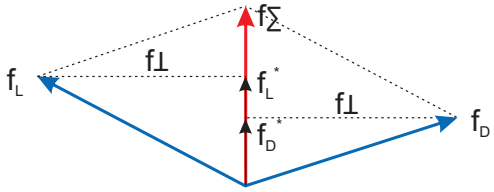
celotnega protokola za posameznega preiskovanca je trajala najmanj 15-20 minut.

2.5 Analiza meritev

Pri analizi meritev smo se zaradi omejitev študije omejili le na skupini zdravih preiskovancev ter pacientov s hujše okvarjenim desnim zgornjim udom.

Da bi lahko ovrednotili parametre gibanja, smo zajete signale najprej primerno obdelali – da bi jih lahko med seboj primerjali, smo trajektorije gibanja parametrizirali z normalizirano dolžino loka ter, s pomočjo informacije o stanju indikatorja pokritosti tarče, posamezno trajektorijo razdelili na intervale med tarčami. S tem smo pridobili enako dolge vektorje, ki omogočajo enostavno primerjavo med različnimi ponovitvami in preiskovanci.

Nazadnje smo izvedli še dekompozicijo sil interakcije z robotom, saj smo za nadaljni izračun parametrov gibanja potrebovali vrednosti sil levega in desnega uda ter vrednosti pravokotnih komponent in aktivnega dela sil (Slika 4).



Slika 4: Dekompozicija vsote sil $\mathbf{f_\Sigma}$ na aktivne, v smeri gibanja, $\mathbf{f^*_{L,D}}$ ter pravokotne $\mathbf{f}_\perp$.

Za statistično analizo razlik v parametrih ocenjevanja gibanja, smo uporabili Mann-Whitneyev test, s katerim smo analizirali razlike med različnimi skupinami ter znotraj posameznih skupin preiskovancev. Podatke smo razdelili v vzorec zdravih preiskovancev ter vzorec podatkov pacientov z okvarjenim desnim zgornjim udom. Primerjave med skupinami preiskovancev ter primerjave med različnimi stopnjami dušenja b smo obravnavali kot neodvisne vzorce. Mejo statistično značilnih vrednosti pri Mann-Whitneyevem testu smo določili pri $p \leq 0{,}05$.

2.6 Parametri ocenjevanja gibanja

Za potrebe našega članka smo izbrali sledeče tri parametre za ocenjevanje gibanja: kooperativnost, delo ter gladkost gibanja.

Kooperativnost M_{ko} (1) predstavlja mero, ki pove, koliko si uda pri izvedbi naloge pomagata med seboj in jo izračunamo kot:

$$M_{ko} = \frac{\sum_{k=1}^{N} \cdot \delta_k \cdot \|f_{\Sigma k}\|}{\sum_{k=1}^{N} \cdot \|f_{\Sigma k}\|}, \tag{1}$$

kjer je δ_k parameter kooperativnosti. Ko je $\delta = 0$, je $\mathbf{f^*_D} = \mathbf{f^*_L} = \frac{1}{2}\mathbf{f_\Sigma}$. Če je $|\delta| \leq \frac{1}{2}$, oba uda prispevata k izvedbi naloge. Če je $|\delta| > \frac{1}{2}$, en ud deluje v smeri, ki je nasprotna $\mathbf{f_\Sigma}$, zato mora drugi ud to silo kompenzirati in izvajati delo, ki je večje od dela, potrebnega samo za izvedbo naloge [**5**]. Pri dvoročnih gibih zgornja uda izvajata različni sili na objekt. Pri izračunu dela A_D (2) smo upoštevali le aktivne sile in je določeno kot:

$$A_D = \int_{t_0}^{t_t} \mathbf{f^*_D} \cdot \dot{p_D} \cdot dt = \int \mathbf{f^*_D} \cdot dp_D, \tag{2}$$

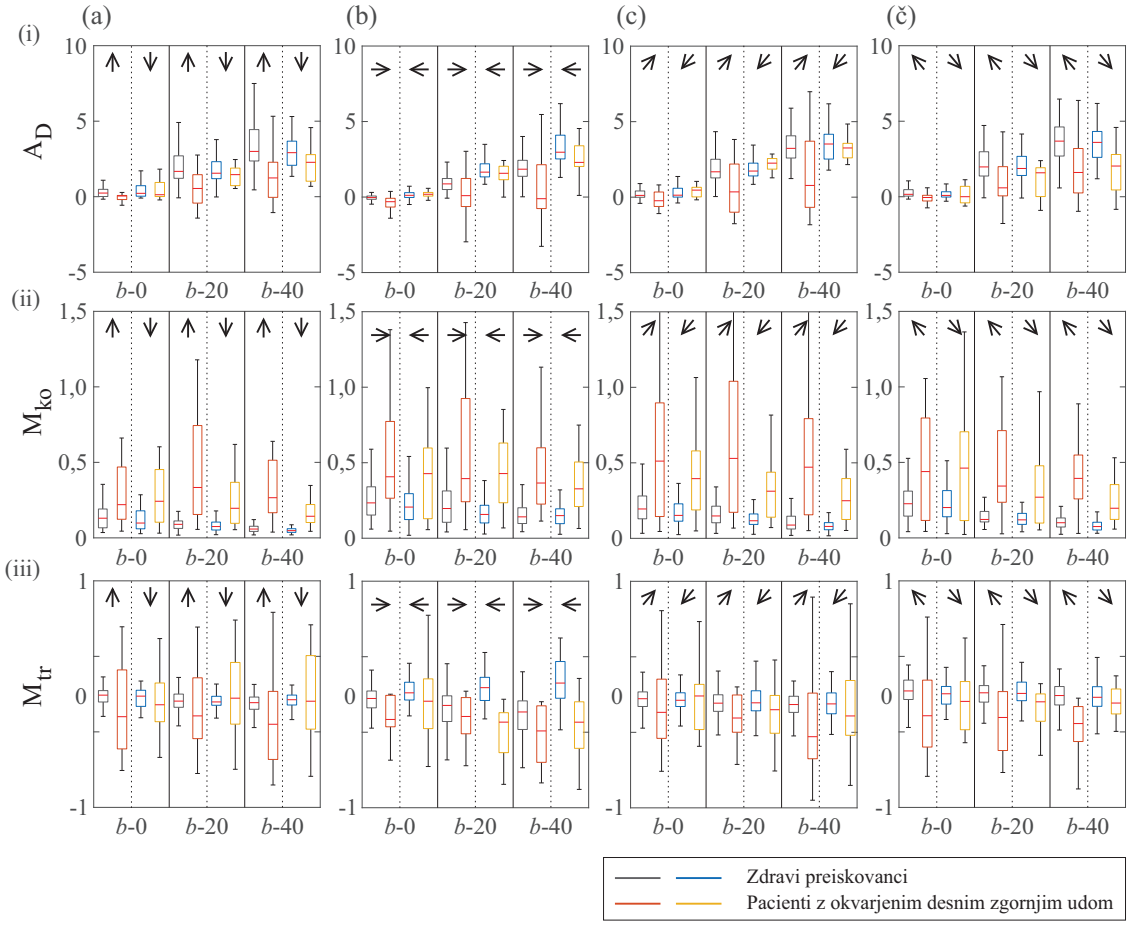
pri čemer sta t_0 in t_t začetni ter končni čas giba [**6**]. Gladkost giba moremo ocenjevati na osnovi trzaja, ki je določen kot tretji odvod položaja po času, oziroma prvi odvod pospeška po času. Sila in pospešek sta linearno povezana, zato lahko kot približek ocene trzaja uporabimo prvi odvod sile po času. Za izračun mere trzaja M_{tr} (3) smo izračunali razmerje trzajev [**6**]:

$$M_{tr} = \frac{1}{2} \log \left(\frac{\sum_{k=1}^{N} \|\dot{\mathbf{f}}_{L_k}\|^2}{\sum_{k=1}^{N} \|\dot{\mathbf{f}}_{D_k}\|^2} \right). \tag{3}$$

3 Rezultati in diskusija

Slika 5 na naslednji strani prikazuje rezultate analize meritev za izbrane parametre ocenjevanja gibanja pri določenih gibih – smer gibanja je označena s puščico na vrhu vsakega škatlastega diagrama. Vrednosti vsakega ocenjevanega parametra (i) so prikazane za vsako od skupin preiskovancev za posamezno gibanje (a-č) in stopnjo dušenja b. Opravljena je bila tudi statistična analiza, pri kateri smo ugotavljali statitično pomembne razlike v vrednostih parametrov med skupinami preiskovancev ter znotraj posamezne skupine preiskovancev za posamezna dušenja b, ki pa je, zaradi omejitve dožine, nismo vključili v članek.

V prvi vrstici so prikazane vrednosti parametra opravljenega dela A_D. Te se pri zdravih preiskovancih pri



Slika 5: Rezultati analize izbranih parametrov - dela A_D, mere kooperativnosti M_{ko} ter gladkosti giba M_{tr}.

vseh gibanjih s povečanjem stopnje dušenja b povečujejo, pri pacientih pa je opazen upad vrednosti. To nakazuje, da so pacienti pri težji nalogi v manjši meri uporabljali okvarjeni, desni ud ter v večji meri nalogo izvajali z zdravim, levim udom.

Pri vrednostih parametra mere kooperativnosti M_{ko} v drugi vrstici, vrednosti tem bližje 0 pomenijo tem boljši rezultat in enakomerno sodelovanje udov. Vrednosti parametra so pri skupini pacientov višje kot pri zdravih preiskovancih, pri vseh gibanjih, kar kaže na to, da so pacienti premik držala v večji meri opravili s pomočjo neokvarjenega, levega uda.

Zadnja vrstica prikazuje vrednosti parametra gladkosti gibanja M_{tr}. Vrednosti bližje 0 pomenijo bolj gladek, nadzorovan gib. Vrednosti zdravih preiskovancev se tekom vseh nalog gibljejo blizu 0, neodvisno od dušenja b in vrste giba, so pa zaradi hitrejšega izvajanja gibanja in posledično trzaja pri preklopu smeri gibanja iz leve v desno, vrednosti povišane pri gibanju (b). Pri pacientih opazimo s povečanjem dušenja b večji odmik od vrednosti 0, kar pomeni, da je bilo pri težji nalogi gibanje manj gladko.

4 Zaključek

S pomočjo robotskega sistema smo uspešno objektivno ovrednotili parametre ocenjevanja gibanja – kooperativnost, delo ter gladkost giba, pri zdravih osebah ter pacientih po kapi in jih primerjali med seboj. To za sistem pomeni veliko uporabnost ter, da ga v praksi lahko uporabljamo za rehabilitacijsko vadbo, hkrati pa tudi za ocenjevanje ter sledenje napredka rehabilitacije pacientov po možganski kapi.

Zahvala

Študija je bila opravljena v sklopu raziskovalnega programa št. P2-0228, ki ga je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

[1] M. Strgar-Hladnik, "Možganska kap", *Združenje zdravnikov družinske medicine Slovenije, vol. 4*, 2014.

[2] S. Waller in J. Whitall, "Bilateral arm training: Why and who benefits?", *NeuroRehabilitation*, vol. 23(1), str. 29-41, 2008.

[3] S. Lokar, *"Rehabilitacijski robot za dvoročno vadbo"*, diplomsko delo visokošolskega študija, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2018.

[4] A. Stražar, *"Analiza gibanja pri dvoročni vadbi z robotskim sistemom HapticMaster"*, magistrsko delo, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2019.

[5] E. Noohi in M. Žefran, "A model for human-human collaborative object manipulation and its application to human-robot interaction", *IEEE Transactions on robotics*, vol. 32(4), str. 880-896, 2016.

[6] E. Čebašek, *"Personalizirana senzorno in robotsko podprta vadba za zgornje ude"*, doktorska disertacija, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2019.

Merilni sistem za analizo dinamičnega trka sodelujočega robota v skladu z ISO/DIS 21260

Sebastjan Šlajpah, Mario Klenovšek, Marko Munih, Matjaž Mihelj

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: {sebastjan.slajpah, marko.munih, matjaz.mihelj}@fe.uni-lj.si, mk8933@student.uni-lj.si

Measurement system for collaborative robot's collision analysis according to ISO/DIS 21260

In this paper, we present the design of a measurement system for analysing dynamic contact of the collaborative robot. Proper assessments of impact forces are utmost importance for providing a safe robotic system that can collaborate with the human worker. The measurement system comprises a force/torque sensor and compression spring for mimicking dynamic of human tissue. Weights are added for changing the inertia of the system and thus modelling different parts of the human body. The whole system is placed on a custom-designed gantry that enables nearly frictionless motion in the direction of impact to mimic the human reaction to the collision. Experimental tests were performed with the robot Universal Robot UR5e at four TCP velocities and with different sensor's inertia. Results show that the maximal impact force depends on the velocity of the robot and the inertia of the measurement system and less on the configuration of the robot. Analysis of the collisions show that contacts with the robot with TCP velocity below 250 mm/s are classified as G2 – occasional contacts; contacts with the robot with velocity 750 mm/s are classified as G3 – rare contacts; contact with the robot with TCP velocity 1000 mm/s are not permitted.

1 Uvod

Sodelujoči roboti predstavljajo pomemben gradnik avtomatizacije v Industriji 4.0, saj so fleksiblni in enostavni za uporabo obenem pa omogočajo izvajanje operacij skupaj z ljudmi v skupnem delovnem prostoru [1]. Sodelujoči roboti imajo lahko več strategij zagotavljanja varnosti uporabnika. Najpogostejša je omejevanje moči in sil, ki poskrbi, da ob neželenem kontaktu z uporabnikom ne pride do prevelikega prenosa energije ter posledično poškodbe uporabnika [2].

Področja varnosti sodelujočih robotov se dotika več ISO standardov [3, 4, 5]. Med bolj pomembne spada tehnična specifikacia ISO/TS 15066, kjer so definirane varnostne meje za sile in pritiske na določene dele človeškega telesa. Trenutno je v pripravi tudi posodobljena različica standarda ISO/DIS 21260 [6], ki podaja metodologijo klasifikacije trkov, te razvrsti v skupine in zanje določi ustrezne meje.

Za zagotavljanje varnosti uporabnika je potrebna detajlna obravnava vsake robotske celice. Pri tem je pomembna analiza tveganja, ki identificira posamezne potencialne nevarne situacije. Za te je potrebno uvesti dodatne varnostne ukrepe ter z meritvami preveriti skladnost s standardi. Merilni sistem za ocenjevanje trka je sestavljen iz merilne celice ter kombinacije vzmeti in gume, ki modelirata človeško telo [7]. Pri meritvi se senzor togo vpne, sam trk pa se deli na dva dela: dinamični del od trka do časa 0,5 s ter kvazi-statičen po času 0,5 s. Največja pomankljivost komercialnih sistemo je, da ne upoštevajo dinamike človeškega telesa, kot je na primer nihaj roke ob trku.

V tem delu bomo predstavili razviti merilni sistem za merjenje trkov robota v dinamičnih razmerah ter ga eksperimentalno preiskusili s sodelujočim robotom UR5e.

2 Standard ISO/DIS 21260

Standard ISO/DIS 21260 [6] podaja omejitve za fizične stike med delom stroja in človekom, ki so posledica gibanja stroja ali njegove zlorabe. Dokument zajema vse vrste strojev, ki so zasnovani tako, da se lahko človek nahaja v njihovem delovnem prostoru ali pa lahko pride do fizičnega stika pri opravljanju nalog.

Inherentno varna zasnova po definiciji vsebuje zaščitne ukrepe a) omejevanje pogonske sile, da gnan del ne povzroča nevarnosti in b) omejevanje mase in/ali hitrosti gibajočih se delov in s tem omejevanje njihove kinetične energije.

Za uporabo tega dokumenta se upošteva postopek za ocenjevanje in zmanjševanje tveganja predstavljen v ISO 12100 [3]. Upoštevajo se vsi pogoji uporabe in predvidljive zlorabe. Priporočeni koraki za analizo kontaktov so sledeči.

- Identifikacija kontaktov med strojem in človekom.
- Kontakt v območju 50 mm od očesa je potrebno preprečiti. Za ostale kontakte je potrebno določiti klasifikacijsko skupino.
- Določiti je potrebno dovoljene pogoje kontakta.
- Primerjava dejanskih in dovoljenih pogojev stika. Če dejanski pogoji presežejo dovoljene, je potrebno spremeniti zasnovo varnostnih ukrepov, dokler niso vrednosti pod dovoljeno mejo.
- Ponovitev postopka za vsak prepoznan stik.

2.1 Klasifikacija kontaktov in meje

V ISO/DIS 21260 so kontakti razdeljeni v štiri skupine glede na čas trajanja. Skupina G1 (visokofrekvenčni kontakti) obsega dinamične kontakte enega posameznika, ki se v obdobju osmih ur pojavljajo več kot enkrat na uro. Skupina G2 (občasni kontakti) obsega dinamične kontakte enega posameznika, ki se v obdobju osmih ur pojavljajo manj kot enkrat na uro, ampak več kot enkrat na teden. Skupina G3 (redki kontakti) obsega dinamične kontakte enega posameznika, ki se pojavljajo manj kot enkrat na teden. V tabeli 1 so podane meje za kontaktne skupine G1, G2 in G3.

	G1	G2	G3
Najmanjša površina kontakta / cm^2	0,5	0,5	0,5
Največji prenos energije / J	1	2	4
Največja dinamična sila / N	75	150	400

Tabela 1: Limite za kontakte skupine G1, G2 in G3

Skupina G4 opisuje statične ali kvazi-statične kontakte in obsega vsak kontakt, kjer je prenos energije manjši od 2 J/s ali pa je sila prisotna za več kot 500 ms. V tabeli 2 so podane meje za kontakte skupine G4.

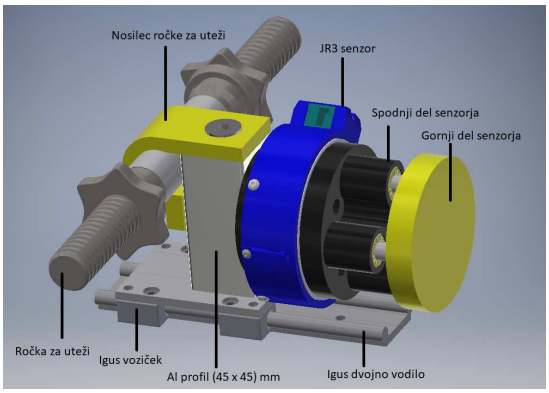
Najmanjša površina kontakta	0,5 cm^2
Največja sila na enoto površine	35 N/cm^2
Največja stalna kontaktna sila	100 N

Tabela 2: Limite za kontakte skupine G4

3 Merilni sistem

ISO/DIS 21260 posebej obravnava dinamične kontakte. Za merjenje dinamičnih trkov je potreben merilni sistem, ki je sposoben posnemati dinamiko človeškega telesa.

Merilni sistem je bil sestavljen iz senzorja sil in navorov JR3 45E15A4 ter vzmeti s togostjo 70 N/mm. Z vzmetjo smo posnemali elastičnost človeškega tkiva - uporabljena vzmet ustreza parametrom dlani in prstov. Vztrajnost segmentov smo modelirali s spreminjajočo maso ter tako dobili različen dinamični odziv sistema.



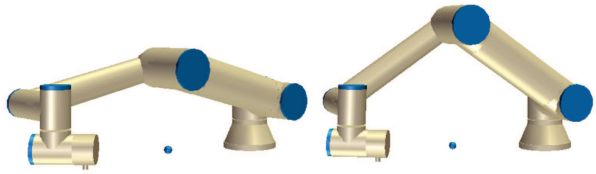
Slika 1: Model sestavljenega merilnega sistema

Merilni sistem smo pritrdili na Igus DryLin WW voziček, ki je bil nameščen na vodilo Igus DryLin WS. Z drsnimi ležaji smo zagotovili minimalen vpliv trenja. Ta pasivna linearna stopnja je ponazarjala odmik človeškega telesa pri trku z robotom. Na sliki 1 je predstavljen celoten sestav merilnega sistema.

Za trk smo uporabili robota UR5e proizvajalca Universal Robots (Danska) z nastavljeno varnostno omejitev sile na 250 N. Robot je bil opremljen s prijemalom Robotiq 2F85 (Kanada). Kot kontaktno površino smo izbrali vrh prstov prijemala.

3.1 Merilni protokol

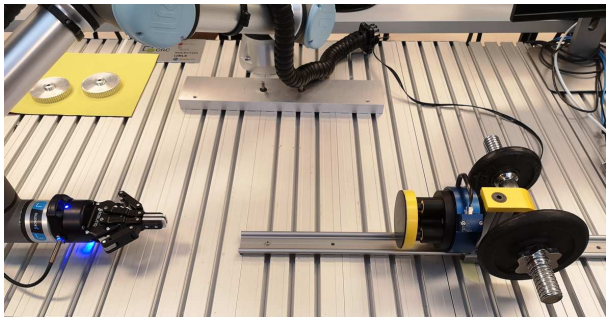
Meritve smo izvedli pri dveh različnih konfiguracijah robota. S tem smo preverili, kako sama postavitev vpliva na pravilno delovanje varnostne funkcije manipulatorja, ter če prihaja do razlik pri izmerjeni sili trka med konfiguracijama. Prva konfiguracija je bila pri iztegnjeni roki robota pri $\mathbf{p} = [-130, -647, -306]$ mm, druga pa pri pokrčenem komolčnem sklepu $\mathbf{p} = [-253, -444, -309]$ mm (slika 2).



Slika 2: Konfiguracija 1 (levo) in konfiguracija 2 (desno)

Najprej smo naredili primerjavo med predstavljenim merilnim sistemom in komercialnim merilnim sistemom KMG-500 (GTE, Nemčija). KMG-500 ima merilno območje od 20 N do 500 N z maksimalno napako 3 %. Analizirali smo trke pri dveh masah (4 kg in 9 kg) ter pri dveh hitrostih vrha robota (250 mm/s in 300 mm/s) za prvo konfiguracijo. Vsaka meritev je bila ponovljena petkrat.

Z gibajočim merilnim sistemom smo analizirali trke pri obeh konfiguracijah. Pri tem smo spreminjali tako hitrost s korakom 250 mm/s v mejah od 250 mm/s do 1000 mm/s, kot tudi maso s korakom 2,5 kg v mejah od 2 kg do 11,5 kg. Pri vsaki kombinaciji smo zajeli pet meritev trka. Na sliki 3 je predstavljena postavitev me-



Slika 3: Postavitev merilnega sistema

rilnega sistema z robotom. Dobljene rezultate smo tudi statistično primerjali med posameznima konfiguracijama robota z uporabo *t*-testa.

Analizirali smo tudi primer, ko je na merilnemu sistemu neskončna vztrajnost (togo vpet merilni sistem). Izkazalo se je, da varnostna funkcija robota ne samo ustavi gibanje, temveč tudi odmakne vrh robota od mesta trka, zato statičnih sil trka nismo mogli primerjati.

Merilni sistem smo povezali preko vmesnika Simulink RealTime, ki deluje v realnem času. Zajem in obdelava podatkov sta potekala preko programskega paketa Matlab R2019b. Vzorčna frekvenca merilnega sistema je bila 1 kHz.

4 Rezultati

Rezultati zajemajo primerjavo predstavljenega merilnega sistema s komercialnim merilnim sistemom ter analizo trkov pri gibajočim senzorjem z različnimi masami ter stacionarnim senzorjem.

4.1 Primerjava z referenčnim merilnikom

Naredili smo primerjavo meritev trka zajetimi tako s predstavljenim kot tudi s komercialnim merilnim sistemom. Komercialni senzor je imel nameščeno vzmet s togostjo 75 N/mm. Pritrdili smo ga na voziček, da se je lahko prosto gibal po vodilu.

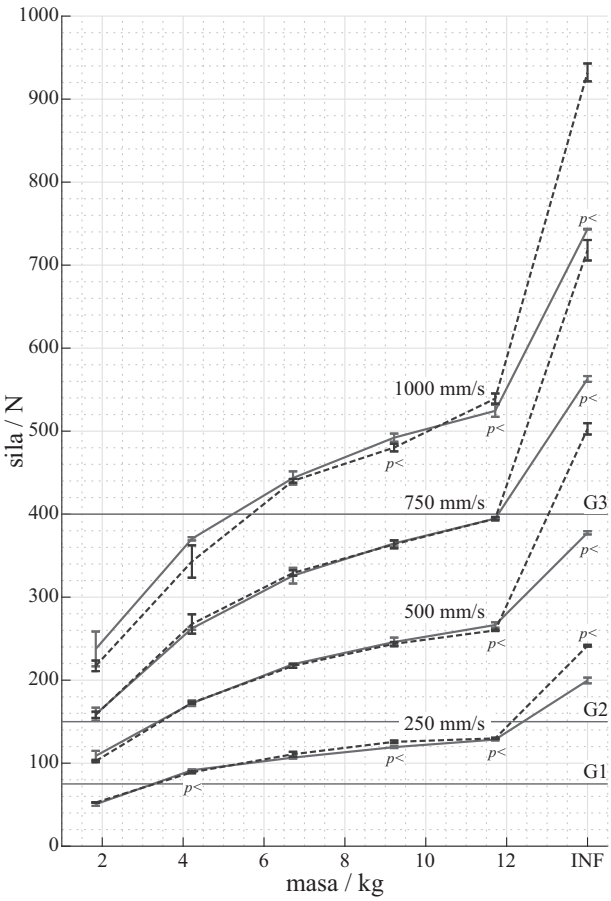
Analizirali smo trk pri hitrostih vrha robota 250 mm/s in 300 mm/s. Ker je bila masa komercialnega senzorja večja od našega sistema, smo za primerjavo uporabili samo kombinacije uteži 4 kg in 9 kg (naš sistem: 4,22 kg in 9,22 kg; komercialni: 4,30 kg in 9,30 kg). V tabeli 3 so prikazane povprečne sile $\bar{F}$ in standardne deviacije SD petih meritev dinamičnega trka za prvo konfiguracijo. Podane so tudi p vrednosti statistične primerjave dobljenih rezultatov z obema merilnima sistemoma.

Masa	**Sistem**	**250 mm/s**		**300 mm/s**	
		$\bar{F}$ / N	SD / N	$\bar{F}$ / N	SD / N
4 kg	S1	87,9	1,8	106,2	4,9
	S2	89,3	1,5	108,3	0,6
		$p = 0,24$		$p = 0,18$	
9 kg	S1	121,4	1,1	143,7	3,1
	S2	117,7	1,2	142,0	1,7
		$p = 0,19$		$p = 0,93$	

Tabela 3: Primerjava rezultatov predstavljenega merilnega sistema (S1) in komercialnega merilnega sistema (S2)

4.2 Premikajoči senzor

Na sliki 4 so predstavljene srednje vrednosti maksimalnih izmerjenih sil trka pri obeh konfiguracijah z ustreznim raztrosom ($\pm$ ena standardna deviacija). Merive so predstavljene za linearne hitrosti vrha robota 250 mm/s, 500 mm/s, 750 mm/s in 1000 mm/s. Maso senzornega sistema smo povečevali s korakom 2,5 kg (1,84 kg, 4,22 kg, 6,72 kg, 9,22 kg, 11,72 kg). Meritve, ki se statistično signifikantno razlikujejo med prvo in drugo konfiguracijo, so označene s $p<$. Na grafu so označene tudi mejne vrednosti največje dinamične sile za kontaktne skupine G1, G2 in G3.



Slika 4: Srednje vrednosti maksimalne sile pri trku v prvi konfiguraciji (polna črta) in drugi konfiguraciji robota (črtkana črta) z razstrosom pri različnih hitrostih in masah merilnega sistema. S $p<$ so označene meritve, ki se med obema konfiguracijama statistično signifikantno razlikujejo ($p < 0,05$).

4.3 Mirujoči senzor

Mirujoči senzor predstavlja senzor z neskončno veliko vztrajnostjo. Na tak način se merijo tudi dinamične sile pri komercialnih sistemih glede na standard ISO/TS 15066. Senzor smo privili na fiksno mesto na vodilu, konfiguraciji in hitrosti robota pa so pri tem ostale nespremenjene. Tabela 4 prikazuje izmerjeno dinamično silo pri trku s togo vpetim senzorjem (povprečna sila $\bar{F}$ in standardna deviacija SD petih meritev).

Hitrost	**Konfiguracija 1**		**Konfiguracija 2**	
	$\bar{F}$ / N	SD / N	$\bar{F}$ / N	SD / N
250 mm/s	199,6	3,4	241,5	1,1
500 mm/s	377,5	2,0	502,8	6,8
750 mm/s	562,9	3,5	717,9	12,3
1000 mm/s	743,3	0,4	932,2	10,8

Tabela 4: Izmerjene sile pri mirujočem senzorju

Naredili smo statistično analizo, ki je pokazala, da

so vse meritve pri isti hitrosti in različnih konfiguracijah statistično signifikantno različne ($p < 0,002$).

5 Diskusija

Iz tabele 3 vidimo, da je naš merilni sistem primerljiv s komercialnim. Razlike pri izmerjenih maksimalnih vrednostih sile trka so manjše od 5 N oziroma 3 %. Statistična analiza je pokazala, da ni statistično signifikantnih razlik med meritvami s komercialnim in meritvami s predstavljenim merilnim sistemom.

Slika 4 prikazuje rezultate trka robota s premikajočim senzorjem pri spreminjajoči vztrajnosti merilnega sistema. Iz grafov vidimo, da na maksimalno silo trka vplivata tako hitrost robota kot tudi masa merilnega sistema, ne pa tudi konfiguracija robotskega mehanizma. Za 14 merilnih točk od 20 ne moremo trditi, da so statistično signifikantno različne, če primernjamo konfiguraciji robota. Za ostalih šest bi se verjetno rezultat statistične analize spremenil s povečanjem števila meritev.

Dobljeni rezultati so v nasprotju z delom [7], kjer so imeli senzor togo vpet. Njihov eksperiment je pokazal močno korelacijo med maksimalno silo trka in oddaljenostjo vrha robota od baze robota. V skladu z njihovimi rezultati pa so rezultati trka pri merilnem sistemu z neskončno vztrajnostjo, ki so prikazani v tabeli 4. Tu vidimo, da je razlika pri isti hitrosti in različnih konfiguracijah lahko tudi 200 N. Zavedati se je potrebno, da mora robotski krmilnik zagotoviti enako hitrost ob trku pri vseh konfiguracijah, če želimo imeti primerljive rezultate. Zaradi mehanskih in kinematičnih omejitev se namreč lahko naredi, da robotski krmilnik, v primeru, da je vrh robota v bližini baze, prilagodi trapezni hitrostni profil gibanja vrha na tak način, da se zagotovi linearnost giba. Ob tem seveda hitrost premikanja robota ustrezno pade.

Tabela 4 nakazuje, da se glede na standard ISO/DIS 21260 trki v mirujoče telo z omejenim gibanjem (na primer roka na mizi) s hitrostjo 250 mm/s uvrščajo v skupino G3, kar pomeni, da se lahko ponavljajo manj kot enkrat na teden. Ostale hitrosti prekoračijo mejo 400 N, tako da ti kontakti niso dovoljeni.

Situacijo, ko je telo prosto gibljivo v prostoru (na primer roka v zraku), opisuje slika 4. Trk z robotom pri hitrosti 250 mm/s spada v skupino G2 – občasni kontakti. Pri tem velja, da je trk v spodnji del roke (masa dlani s podlaktjo odraslega moškega z maso 80 kg znaša približno 1,8 kg, celotna roka pa 4 kg [8]) klasificiran kot G1 – pogosti kontakti. Pogoj pri tem je, da je roka prosto gibljiva, kontakt pa pride iz lateralne smeri; s tem se zagotovi vrtišče v komolcu. Pri izvajanju naloge, kjer si robot in človek delita delovni prostor, bi se, glede na meje podane v standardu [6], lahko v obdobju osmih ur pojavljali trki v podlakt ali dlan več kot enkrat na uro (skupina G1), v nadlakt pa manj kot enkrat na uro (skupina G2).

Trki s hitrostjo robota do 750 mm/s spadajo v skupino G3. V skupino G3 spadajo tudi trki s hitrostjo 1000 mm/s v telo z maso 4 kg ali manj (na primer, človeška roka). Ostali trki robota s hitrostjo 1000 mm/s niso dovoljeni.

6 Zaključek

Primerjava razvitega sistema s komercialnim je pokazala, da sta merilna sistema primerljiva. Meritve trka pri gibljivem senzorju so pokazale, da je maksimalna sila trka odvisna od hitrosti robota in vztrajnosti senzorja. Sama konfiguracija robota ima manjši vpliv. Pri meritvah s stacionarnim senzorjem pa ima ravno konfiguracija robota velik vpliv na maksimalno silo: bližje kot je vrh robota svoji bazi, večja je maksmimalna sila trka.

Rezultati so pokazali, da trki z robotom s hitrostjo do 250 mm/s spadajo v kategorijo G2. Trke v roko pri tej hitrosti lahko uvrstimo v kategorijo G1 – visokofrekvenčni kontakti. Kontakti pri hitrosti robota do 750 mm/s spadajo v kategorijo redkih kontaktov, medtem ko kontakti pri hitrosti robota 1000 mm/s niso dovoljeni (izjema je trk v roko, ki spada v skupino G3).

Merjenje trka z gibljivim senzorjem s spremenljivo vztrajnostjo je pokazalo zanimive ugotovitve. Bolj poglobljena študija lahko služi kot osnova za nadgradnjo obstoječega standarda ISO/TS 15066, kjer so podane mejne vrednosti sil za dinamične in kvazi-statične trke, pri čemer bi se seveda omejitve ustrezno zmanjšale s tem pa povečala fleksibilnost in uporabnost sodelujočih robotov.

Zahvala

Delo je nastalo v okviru raziskovalnega programa številka P2-0228, ki ga sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

[1] D. Surdilovic, G. Schreck, and U. Schmidt, "Development of collaborative robots (cobots) for flexible human-integrated assembly automation," in *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, June 2010, pp. 1–8.

[2] M. Mihelj, T. Bajd, A. Ude, J. Lenarčič, A. Stanovnik, M. Munih, J. Rejc, and S. Šlajpah, *Robotics*. Springer, 2019.

[3] ISO, "ISO 12100:2010: Safety of machinery – General principles for design – Risk assessment and risk reduction," Ženeva, Švica, 2010.

[4] ——, "ISO 10218-1:2011: Robots and robotic devices – Safety requirements for industrial robots – Part 1: Robots," Ženeva, Švica, 2011.

[5] ——, "ISO/TS 15066:2016: Robots and robotic devices – Collaborative robots," Ženeva, Švica, 2016.

[6] ——, "ISO/DIS 21260:2018: Safety of machinery – Mechanical safety data for physical contacts between moving machinery or moving parts of machinery and persons," Ženeva, Švica, 2018.

[7] A. Schlotzhauer, L. Kaiser, J. Wachter, M. Brandstötter, and M. Hofbaur, "On the trustability of the safety measures of collaborative robots: 2D Collision-force-map of a sensitive manipulator for safe HRC," in *2019 IEEE 15th international conference on automation science and engineering (CASE)*. IEEE, 2019, pp. 1676–1683.

[8] P. De Leva, "Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters," *Journal of biomechanics*, vol. 29, no. 9, pp. 1223–1230, 1996.

Vodenje sodelujoče mobilne robotske celice

Peter Kmecl, Matjaž Mihelj, Marko Munih, Janez Podobnik

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana
E-pošta: peter.kmecl@robo.fe.uni-lj.si

Control of a colaborative mobile robot cell

In recent years we have seen a steady growth in the use of mobile robots in industrial scenarios. The growth can be attributed to the improvement of battery and computer technologies as well as new autonomous control algorithms. This paper describes a mobile robot, developed in the Laboratory of robotics at the Faculty of Electrical Engineering, that serves as an autonomous collaborative mobile robot cell and was created as a response to the aforementioned trend. Its purpose is to allow research and development of collaborative mobile applications, as well as enable students to have access to a modern and open mobile robotics system. The paper provides a basic overview of the cell's electrical, communication and information systems as well as the software used to control and plan its movement.

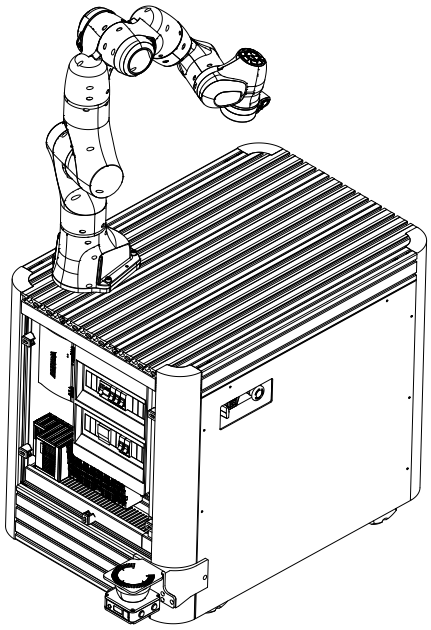
1 Uvod

Mobilna robotika je danes v porastu tako v vsakdanjem življenju, kot tudi v industrijskem okolju. Ključni napredki na področju baterijske tehnologije, procesorske moči in algoritmov vodenja nam omogočajo avtonomijo sistema in natančno vodenje, napredni senzorski sistemi pa zagotavljajo varnost in dovoljujejo deljenje robotovega delovnega prostora z ljudmi [1].

Uporabnost mobilnih robotov se še bistveno poveča, če platformo kombiniramo z robotskim manipulatorjem. Takemu sistemu rečemo tudi mobilni manipulator. Mobilni manipulatorji razširijo funkcionalnost obeh komponent v sistemu. Premična platforma manipulatorju poveča delovni prostor, platforma pa pridobi sposobnost samostojnega opravljanja nalog, ki zahtevajo fleksibilnost manipulatorja (pobiranje in odlaganje, izpolnjevanje naročil, oskrbovanje in vzdrževanje naprav itd.) [1, 2].

Zaradi vse večje uporabe tovrstnih sistemov v tovarnah in skladiščih, smo v Laboratoriju za robotiko na Fakulteti za elektrotehniko, UL razvili sodelujočo mobilno robotsko celico za raziskovalne in pedagoške namene.

V članku bomo predstavili napajalne, komunikacijske in računalniške sisteme robotske celice, ter visokonivojsko in nizkonivojsko programsko opremo, ki nam omogočata avtonomno vožnjo. Na kratko bomo omenili tudi njene varnostne sisteme in algoritme za lokalno načrtovanje poti, ki smo jih preizkusili.



Slika 1: Skica sodelujoče mobilne robotske celice s pritrjenim manipulatorjem Franka Emika Panda.

2 Strojna oprema

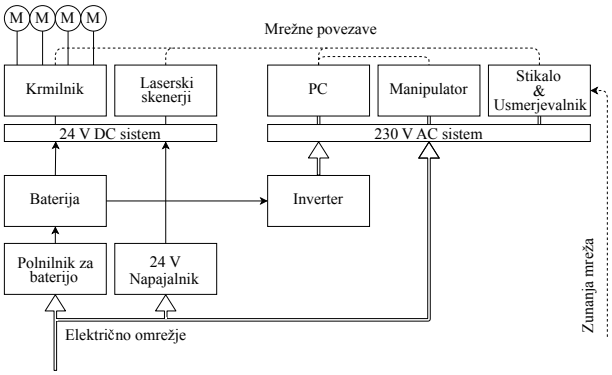
Osnovne funkcije povezane z mobilnostjo naše sodelujoče robotske celice nadzira Beckhoffov industrijski krmilnik oziroma PLK. Zasnovan je modularno in tako omogoča preprosto razširitev ter nadgradnjo funkcionalnosti, ki bi jo pri nadaljnjem razvoju morda potrebovali. Krmilnik vsebuje štirijedrni procesor in poganja operacijski sistem *Windows 10 embedded*. Nameščeno programsko okolje *TwinCAT3* omogoča dodelitev jeder specifičnim nalogam ter komunikacijo z moduli preko protokola *EtherCAT*. Taka izvedba krmilnika pripomore k preprostejši integraciji nizkonivojske strojne opreme v visokonivojski sistem preko standardnih protokolov in komunikacijskih poti. S posebnimi varnostnimi moduli krmilnik poskrbi tudi za osnovne varnostne funkcije. Poleg vodenja motorjev in varnosti krmilnik podaja osnoven uporabniški vmesnik, ki operaterju omogoča ročno premikanje platforme preko krmilne palice s tremi prostostnimi stopnjami. Skupaj z napajalniki in distribucijskimi sistemi je pritrjen na zadnji del platforme, ki je prikazan na sliki 2.

Na celico smo za povečanje mobilnosti vgradili 24 V



Slika 2: Napajalnik (modra), električni varnostni ter distribucijski sistemi (rdeča) in industrijski krmilnik (zelena).

litij-železov polimerno baterijo s kapaciteto 50 Ah in 230 V inverter z maksimalno močjo 1 kW, ki skrbi za napajanje računalnika in robotskega manipulatorja. Električni sistemi so napeljani v nadometno omarico za električno napeljavo, kjer s pomočjo varovalk in FID stikala poskrbimo za varnost (rdeč okvir na sliki 2). Za namene nadalnjega razvoja so v celico vgrajeni vsi potrebni napajalni sistemi za napajanje celice preko električnega omrežja. Shema celotnega električnega sistema je prikazana na sliki 3.

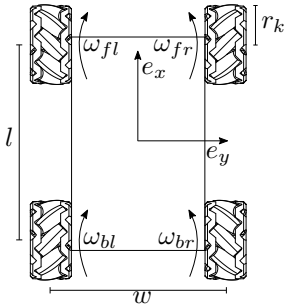


Slika 3: Blokovna shema električnih sistemov na celici.

Na krmilnik so povezani štirje servo motorji, na katere so preko reduktorjev nameščena *Mecanum* kolesa. Ta celici omogočajo premike po osi x in y, rotacijo okoli osi z, ter njihove poljubne kombinacije. Koordinatni sistem mobilne platforme celice, ter njegove osnovne oznake so prikazane na sliki 4.

Naprednejše funkcije vodenja se izvajajo na zmogljivem strežniškem računalniku, na katerega je nameščen operacijski sistem *ESXi*. Ta omogoča namestitev več virtualnih sistemov na en računalnik, kar pripomore h kompaktnosti in varčnosti robotske celice. Nizkonivojsko in visokonivojsko programsko opremo povezuje omrežje Ethernet.

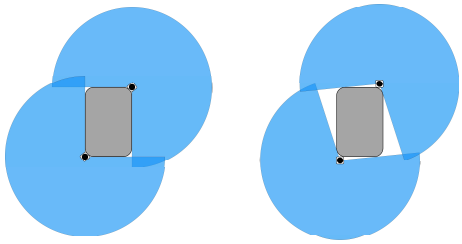
V komunikacijsko omrežje sta povezana tudi dva SICK laserska skenerja. Senzorja omogočata branje podatkov preko omrežja Ethernet in bosta uporabljena za zaznavanje ovir in izgradnjo zemljevida delovnega prostora. Podpirata varnostne digitalne izhode, ki se sprožijo



Slika 4: Koordinatni sistem in osnovne oznake podvozja celice.

v primeru zaznave ovire v vnaprej nastavljenem območju. Digitalni izhodi so povezani na varnostni krmilnik, ki poskrbi za varen izklop motorjev.

Pri laserskih skenerjih smo se morali odločiti za pozicijo in orientacijo pritrditve. Postavljeni morajo biti dovolj nizko, da zaznajo vse ovire na tleh, hkrati pa dovolj visoko, da človeka med hojo oziroma tekom še vedno zaznajo. Poleg višine smo se morali odločiti med dvema montažnima pristopoma - diagonalnim in kotnim. Diagonalni s svojimi varnostnimi polji pokrije celo okolico platforme, vendar znatno poveča profil platforme. Povečani profil oteži navigacijo po tesnih prehodih in približevanje statičnim delovnim površinam (angl. docking). Kotna pritrditev olajša prej omenjeno navigacijo, žrtvuje pa trikotne dele varnostnega polja tik ob platformi. Varnostna polja obeh pristopov so prikazana na sliki 5 [1].



Slika 5: Možni konfiguraciji laserskih skenerjev.

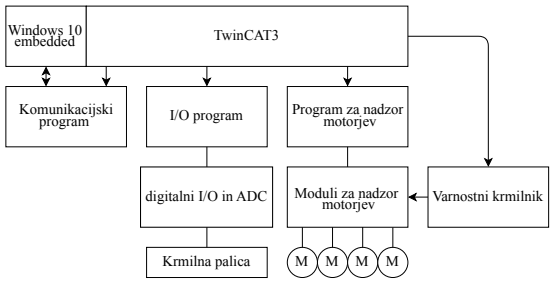
Celica ima na svojo delovno površino pritrjen tudi sodelujoč robotski manipulator Franka Emika Panda s 7 prostostnimi stopnjami, ki je viden na slikah 1 in 6.



Slika 6: Robotski manipulator, ki je pritrjen na celico.

3 Nizkonivojska programska oprema

Nizkonivojska programska oprema teče na Beckhoffovem industrijskem krmilniku. Njena naloga je omogočiti povezavo med strojno opremo, ter višjenivojskimi komponentami, ki skrbijo za pravilno delovanje robotske celice na nivoju aplikacije. V našem primeru jo sestavljajo štiri komponente: vodenje motorjev, obravnava vhodnih signalov, komunikacijski vmesnik ter obravnava varnostnih vhodov in izhodov. Blokovna shema sistema je prikazana na sliki 7.



Slika 7: Blokovna shema nizkonivojskega sistema vodenja.

Za nizkonivojsko programsko opremo je ključnega pomena izvajanje v realnem času s fiksnim časovnim intervalom, saj potrebujemo znano frekvenco izvajanja krmilne programske opreme za natančno računanje kinematike in odometrije mobilne platforme. To zagotovimo z dodelitvijo vsake programske komponente na lastno krmilniško jedro, kar nam omogoča izbiro fiksnega koraka izvajanja posamezne komponente v nadzornem vmesniku *TwinCAT3*. Varnostna logika teče na lastnem varnostnem krmilniku (angl. safety PLC) in je od ostalih komponent neodvisna [3].

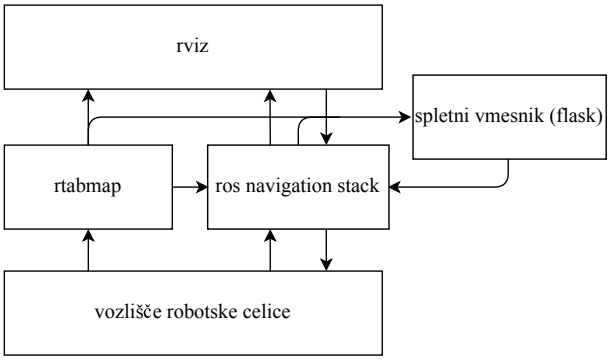
Vodenje motorjev je procesorsko najzahtevnejša komponenta nizkonivojske programske opreme. Vključuje matrično računanje kinematičnega modela in odometrije, za katerega krmilnik ni prilagojen, hkrati pa se mora za natančno delovanje izvajati s čim krajšim časovnim korakom. V našem primeru dela s frekvenco 1 kHz oziroma s časovnim korakom 1 ms [3].

Nizkonivojski programski modul za obdelavo signalov je trenutno uporabljen le kot preprost integriran uporabniški vmesnik za ročno premikanje mobilne platforme preko krmilne palice. Za delovanje uporablja analogni vhodni modul in digitalni vhodno/izhodni modul. V nadaljnjem razvoju bi se nanj lahko priklopili dodatni senzorji, potrebni za robustnejše in varnejše delovanje sodelujoče mobilne robotske celice.

Do nizkonivojskih funkcij mobilne celice, ki jih ponuja njena nizkonivojska programska oprema (odometrija, hitrostno krmiljenje itd.), si želimo dostopati preko visokonivojske programske opreme, zato med njima potrebujemo hiter komunikacijski vmesnik. S stališča hitrosti in preprostosti implementacije je najprimernejša tehnologija komunikacijskega vmesnika Ethernet, preko katerega lahko s protokolom UDP komuniciramo z zelo majhnimi zamiki. Komunkacijski vmesnik skrbi za pošiljanje informacij in odometrije ter za prejemanje in izvajanje ukazov, prejetih preko vmesnika UDP.

4 Visokonivojska programska oprema

Naloga visokonivojske programske opreme mobilne robotske celice je zagotovitev vmesnika za preprostejšo uporabo njenih ključnih sistemov, ter opravljanje in izvrševanje odločitev na nivoju aplikacije. Zasnovana je modularno na podlagi virtualizacije in odprtokodnega okolja *ROS* oziroma *Robot Operating System*. Okolje *ROS* zagotavlja standardizacijo komunikacijskega vmesnika med moduli, virtualizacija pa nam omogoča uporabo sicer nekompatibilnih operacijskih sistemov na enem zmogljivem strežniškem računalniku brez potrebe po dodatnem prostoru na platformi in z minimalnim povišanjem njene energijske porabe. S stališča vodenja premikov celice je najpomembnejši visokonivojski sistem ROS, katerega shema je prikazana na sliki 8 [1].



Slika 8: Poenostavljena blokovna shema visokonivojskega vodenja v sistemu ROS.

Za lokalizacijo smo uporabili *ROSov* paket *rtabmap*, odprtokodno implementacijo algoritma za VSLAM (angl. Visual Simultanious Localization And Mapping). Prednost paketa *rtabmap* pred ostalimi sorodnimi paketi je podpora različnih vrst 2D in 3D senzorjev. To nam omogoča dodajanje lidarjev, laserskih skenerjev in RGB-D kamer glede na dodatne potrebe brez spremembe strukture navigacijskega sistema. Hkrati podpira "večsejno" mapiranje (angl. multi-session mapping), lokalizacijo po ugrabitvi (angl. kidnapping) ter lokalizacijo na predhodno shranjenem zemljevidu, kar omogoča raznoliko delovanje v vseh okoliščinah [4, 5, 6, 7].

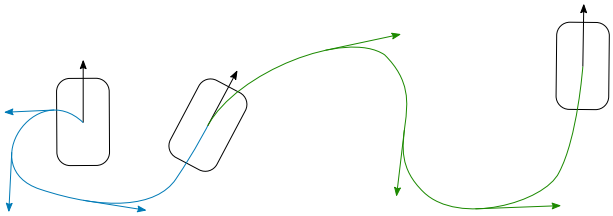
Pot po zemljevidu do željene lokacije načtruje ROSov vgrajeni paket imenovan *move_base*, ki je del sistema *ROS navigation stack*. Razdeljen je na globalni in lokalni algoritem, na podlagi katerih izračuna optimalno pot okoli statičnih ovir na zemljevidu in dinamičnih ovir zaznanih na senzorjih.

Paket za komunikacijo z nizkonivojsko programsko opremo robotske celice smo napisali sami v programskem jeziku Python.

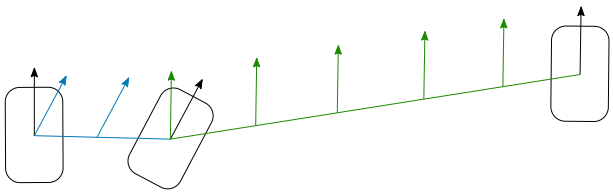
Lokalna navigacija v paketu *move_base* je zasnovana na podlagi vtičnikov. Ti vsebujejo implementacije različnih algoritmov, ki se razlikujejo po sposobnosti izogibanja dinamičnim oviram in podpori za določene pogonske sisteme [7]. V sklopu razvoja našega sistema smo v simulacijskem okolju preizkusili štiri vtičnike, ki so implementirali algoritme *Trajectory rollout*, *Dynamic window*

approach [8], *Elastic band* [9] in *Timed elastic band* [9]. Skice trajektorij, pridobljenih z omenjenimi vtičniki so prikazane na slikah 9, 10 in 11.

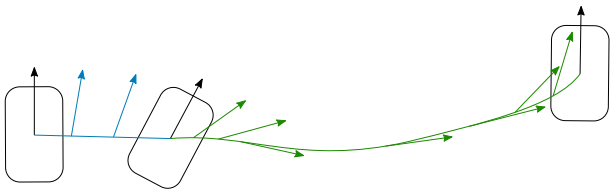
Kot najprimernejši algoritem za naš sistem se je izkazal *Timed elastic band*, ki za razliko od algoritmov *Trajectory rollout* in *Dynamic window approach* primerno izkoristi vsesmernost platforme. Trajektorije, pridobljene z algoritmoma *Elastic band* in *Timed elastic band* se bistveno ne razlikujeta, a se je slednji izkazal za bolj robustno v dinamičnih situacijah (predvsem pri izogibanju dinamičnim oviram) [9].



Slika 9: Skica trajektorije celice, pridobljene z algoritmom *Trajectory rollout*.



Slika 10: Skica trajektorije celice, pridobljene z algoritmom *Dynamic window approach*.



Slika 11: Skica trajektorije celice, pridobljene z algoritmom *Timed elastic band*.

5 Zaključek

Razvili smo sodelujočo mobilno robotsko celico za namen izvajanja raziskav s področja sodelovanja človeka z mobilnimi roboti in razvoj mobilnih robotskih aplikacij. Sistem smo zasnovali modularno, kar nam omogoča hiter nadaljnji razvoj in nadgradnjo sistema.

Nizkonivojsko vodenje temelji na modularnem industrijskem krmilniku Beckhoff. Ta skrbi za pravilno delovanje motorjev in osnovne varnostne funkcije, naprednejše varnostne funkcije in načrtovanje poti pa so zasnovane na odprtokodnem sistemu ROS, ki teče na zmogljivem strežniškem računalniku. Poleg visokonivojskega vodenja platforme s pomočjo virtualizacije nudi sistem za visokonivojsko vodenje robotskega manipulatorja, ki je pritrjen na celico, ter možnost namestitve poljubnih drugih visokonivojskih sistemov.

Celica je opremljena z baterijo, ki celotnemu sistemu omogoča avtonomno delovanje nekaj ur. Za namen razvoja je opremljena tudi z vsemi potrebnimi napajalniki, ki omogočajo napajanje celice iz električnega omrežja. Za napajanje 230 V sistemov, ki so pritrjeni na celico imamo v platformo vgrajen tudi zmogljiv inverter.

Celica omogoča razvoj naprednih sodelujočih aplikacij, varnostnih funkcij in sodelovanja med mobilno celico ter pritrjenim robotskim manipulatorjem. Zaradi odprtosti sistemov je primerna tako za raziskovalno kot tudi pedagoško dejavnost v laboratorijskem okolju.

Zahvala

Študija je bila opravljena v sklopu raziskovalnega programa št. P2-0228, ki ga je sofinancirala Javna agencijaza raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Industrijski krmilnik Beckhoff, ter pripadajoče module in motorje je prispevalo podjetje Beckhoff Avtomatizacija d.o.o.

Literatura

[1] P. Kmecl, "Vodenje sodelujoče mobilne robotske celice", *Magistrsko delo*, Ljubljana: Univerza v Ljubljani, Fakulteta za elektrotehniko, 2020.

[2] Biao Zhang, C. Martinez, JianjunWang, T. Fuhlbrigge, W. Eakins in Heping Chen, "The challenges of integrating an industrial robot on a mobile platform," v *2010 IEEE International Conference on Automation and Logistics*, str. 255–260, 2010.

[3] H. Taheri, B. Qiao in N. Ghaeminezhad, "Kinematic model of a four mecanum wheeled mobile robot," *International Journal of Computer Applications*, vol. 113, no. 3, str. 6–9, 2015.

[4] M. Labbé in F. Michaud, "Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, str. 416–446, 2019.

[5] C. Rösmann, F. Hoffmann in T. Bertram, "Online trajectory planning in ros under kinodynamic constraints with timed-elastic-bands," v *Robot Operating System (ROS). Studies in Computational Intelligence*, vol. 707, Springer.

[6] C. Rösmann, W. Feiten, T. Wösch, F. Hoffmann in T. Bertram, "Efficient trajectory optimization using a sparse model," v 2013 *European Conference on Mobile Robots*, str. 138–143, 2013.

[7] P. Marin-Plaza et al., "Global and local path planning study in a ros-based research platform for autonomous vehicles," *Journal of Advanced Transportation*, vol. 2018, Feb 2018.

[8] D. Fox, W. Burgard in S. Thrun, "The dynamic window approach to collision avoidance," IEEE Robotics Automation Magazine, vol. 4, no. 1, str. 23–33, 1997.

[9] S. Quinlan in O. Khatib, "Elastic bands: connecting path planning and control," v *[1993] Proceedings IEEE International Conference on Robotics and Automation*, str. 802–807 vol.2, 1993.

Simulacije in modeliranje

Simulations and Modeling

Analytical Modelling of Achieving Energy Efficiency with Green Walls

Dejan Tasić[1], Amor Chowdhury[1,2], Dalibor Igrec[1,2]

[1]University of Maribor, Faculty of Energy Technology, Hočevarjev trg 1, 8270 Krško
[2]Margento R&D d.o.o., Turnerjeva ulica 17, 2000 Maribor
E-mail: dalibor.igrec@um.si

Abstract. *In the last years, energy efficiency has become an important issue for the European Union (EU) and Slovenia, who both strive for more sustainable use of energy in buildings. As an innovative approach to such matter, green walls are vertical structures that can serve both as a façade or an insulation layer as well as an aesthetic design for building applications.*

This paper is an overview of author's master's thesis [1] and, therefore, contains calculated annual heat (transmission) losses through different wall assemblies with and without a constructed green wall in front of them. This paper presents a comparison of values of heat losses between those wall assemblies and the explanation of such results, which can determine if green walls present a viable energy-efficient structure for buildings.

1 Introduction

Population growth and an ever-increasing demand for energy are redirecting people to start using energy more wisely, and efficiently. To reduce greenhouse gas emissions, people have become aware that not only is it essential to reduce the usage of fossil fuels and switch to renewable energy sources but also to change the way how people use energy. Changing our homes to energy-efficient habitat's that immensely reduce the usage of energy, thus becoming energy efficient, has attracted the EU's attention which now greatly invests in such sector. Slovenia is also bounded to increase the percentage of energy-efficient buildings in the upcoming future.

This paper will focus on green walls as an element, a structure that could serve as a potential energy-efficient layer in front of a regular (main) wall assembly that could possibly reduce heat losses through such walls. In this respect, this paper contains the results of analytical models of the green wall, placed in front of three different main wall assemblies and analytical models of main wall assemblies, without the green wall. All simulations (models) were developed in MATLAB by MathWorks.

The goal was to determine the difference of heat (transmission) losses between wall assemblies that contain the green wall structure in front of them and wall assemblies that do not contain such structure. Losses are accounted for and compared by/with each model on an annual basis for the year 2018. This paper will also elaborate on potential energy savings and CO_2 emissions reduction and the results that were given based on the input data.

2 Green walls

Green walls are described as vertical structures (gardens) that have different types of plants or other greenery attached to them. Plants are planted in a growth medium that consists of stone, water, and soil. A built-in irrigation system is featured in a green wall for sustaining living greenery [2].



Figure 1. Green wall system – Kyoto railway station, Japan [3]

Green walls can improve the energy efficiency of buildings in cities, where it is vital to reduce energy consumption so they account for being a sustainable technology that acts as a natural insulation system. They also improve the urban climate (reducing buildings' surface temperature, urban air temperature, regulating CO_2 concentration), and improving air and water quality. Besides good thermal and acoustic performances, they also serve as a great aesthetic structure, that helps to diversify the city landscape. There are also disadvantages such as intensive maintenance due to the vegetation, and a need for a durable supporting construction [4]. While green walls are almost nonexistent in Slovenia, there are already significant strives in countries with big urban areas, such as Japan and Singapore, especially because of large heat island phenomena and high humidity.

Green wall systems include pre-cultivated panels (that provide support) with greenery, vertical modules (planter boxes, filled with a substrate - soil) which are vertically fixed to a support structure or on the wall and an irrigation system to sustain plant life [5]. Plants are appropriately suited to specific climate conditions. During their growth, plants become rooted and a system acts similar to a garden foam [6]. Plants on a wall absorb a large amount of solar radiation for their growth and their biological functions (photosynthesis,

evaporation, transpiration, and respiration), and act also as a solar barrier that reduces the absorption of solar energy by reflecting the incident solar radiation. In return, the greenery needs to be carefully maintained with an irrigation system that provides water and nutrient supplements. The structure also restricts the wind effect around the building, traps fine dust, and manages the humidity of the building's environment. Thus, these walls regulate the temperature of the building, and with that, they are reducing energy consumption [5].

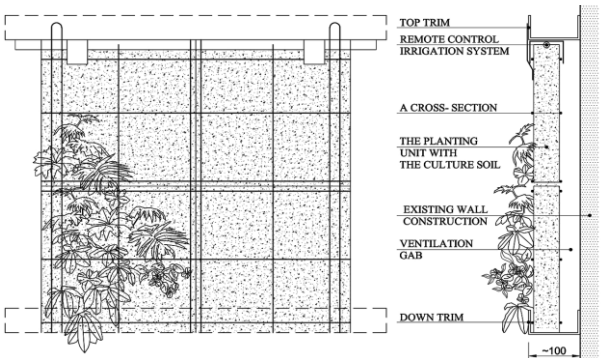


Figure 2. A design example of a green wall system [6]

3 Modeling of the proposed model

The proposed model (building wall) was a theoretical approach with the same dimensions (surface area of 25 m^2) in each cardinal direction (north, south, east, and west) and was not based on a structure, taken from a real, present-time case. The location of the "structure" was set in Maribor, Slovenia (the location is important for assessing values such as solar radiation and external air temperature). Different types of main wall assemblies were set as examples: an energy-efficient wall (highly energy-efficient masonry with thermal insulation – Type A), a highly energy-inefficient wall (old masonry with no thermal insulation – Type B) and an energy-inefficient wall (concrete with some thermal insulation – Type C). The whole process of determining thermal resistivities in our model is explained in [1] and based on those, the density and thickness of the green wall have different impact on heat losses.

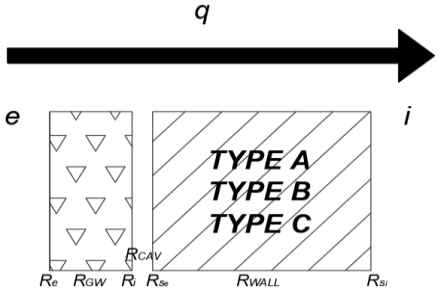


Figure 3. Thermal resistivities for the whole structure [1]

Before we began simulations on the proposed model, we observed a theoretical approach to general building physics, from general terms to thermal resistances for a wall assembly (Figure 3), green wall parameters, normal wall parameters [1], and equations that explain the entire heat flow through a wall: a building structure-green wall [7] and later: a complete structure. Boundary conditions were also applied afterward and a mathematical model was created for each wall assembly respectively. First, by establishing energy balance, plants can perform their functions (photosynthesis), therefore their attempt to maintain an energy balanced leaf temperature attributes to maximizing plants' usefulness. This means that the supplied energy E_{IN} equals the exiled energy E_{OUT} . At equilibrium state:

$$q_{RAD,IN} + q_{SOL,IN} = q_{IR,OUT} + q_{CV} + q_{TR} \quad (1)$$

where $q_{RAD,IN}$ is the heat flow density of absorbed long-wave radiation, $q_{SOL,IN}$ is the heat flow density of absorbed solar radiation, $q_{IR,OUT}$ is the heat flow density of irradiated long-wave radiation, q_{CV} is the heat flow density of convection and q_{TR} is the heat flow density of transpiration. If leaves are not at equilibrium state with its environment, the temperature of the leaves (on its surface) will naturally increase or decrease until such equilibrium is achieved [7]. As the amount of solar radiation depends not only on the power contained in the sunlight but also on the angle between the wall's surface and the Sun (which is continually changing throughout the year), we needed to add some moderations to some equations of the model, when calculating solar radiation given our proposed wall:

$$q_{MODULE} = q_{SOL,DIR} \cdot \cos(180 - \theta_I) \ \left[\frac{\mathrm{W}}{\mathrm{m}^2}\right] \quad (2)$$

where q_{MODULE} is the solar radiation given the wall, $q_{SOL,DIR}$ is the solar radiation from the Sun, and θ_I is the solar incidence angle, comprised of the declination angle, latitude angle, solar azimuth angle, and declination angle [1]. Equation 2 is the final form for calculations from vernal to the autumnal equinox, while:

$$q_{MODULE} = q_{SOL,DIR} \cdot \left(-\cos(180 - \theta_I)\right) \left[\frac{\mathrm{W}}{\mathrm{m}^2}\right] \quad (3)$$

is a final form equation, regarding the period from autumnal to the vernal equinox, Equations 2 and 3 are for north and south cardinal directions while Equation 4 is for east and west cardinal directions [1]:

$$q_{MODULE} = q_{SOL,DIR} \cdot \left(-\cos(\theta_I)\right) \ \left[\frac{\mathrm{W}}{\mathrm{m}^2}\right] \quad (4)$$

4 Results of the simulation

Transmission losses through walls $Q_{T,l}$ were assessed with:

$$Q_{T,l} = \Phi_{T,l} \cdot t_M \cdot 10^{-3} \ [\mathrm{kWh}] \quad (5)$$

where $\Phi_{T,l}$ are transmission losses and t_M is time in hours based on a certain month (for example: January lasts 744 hours) [1]. Wall assemblies (A, B, and C) were compared with the green wall and without it. The results (comparisons) are seen in Figures 4, 5, and 6.

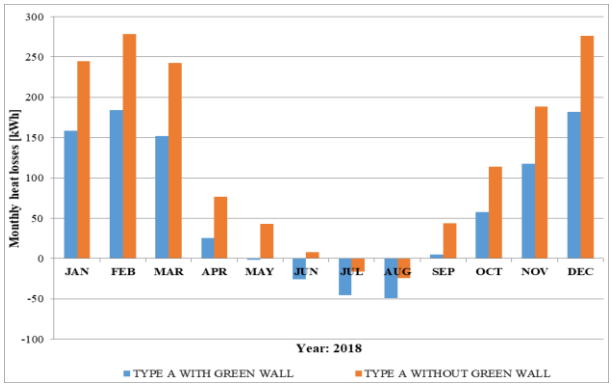


Figure 4. Comparison of monthly heat losses between Type A wall structure with the green wall and without it

If a Type A wall would have a green wall installed, annual heating energy needs (heat losses) would reduce for 41,75 %, yet, annual cooling energy needs would increase. Combined, there would be 550,8 kWh of annual energy savings with Type A wall that would have a green wall installed in front of it (Figure 4).

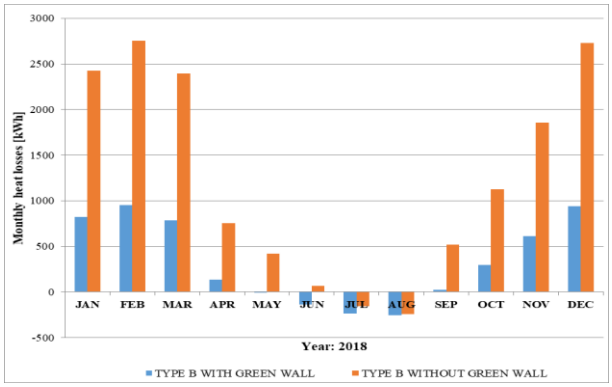


Figure 5. Comparison of monthly heat losses between Type B wall structure with the green wall and without it

If a Type B wall would have a green wall installed, the annual heating energy need (heat losses) would reduce for staggering 69,7 % (from 15.055,5 kWh to 4572,1 kWh) in total respectively. Annual cooling energy need would increase for a bit, however, combined, there would be 10.253,5 kWh of annual energy savings with Type B wall that would have a green wall installed in front of it (Figure 5).

With regards to the third type: if a Type C wall would have a green wall installed, the annual heating energy need (heat losses) would reduce for 59,2 %, annual cooling energy need would increase only from 148,9 kWh to 319,2 kWh, however, combined, there would be 3.193,9 kWh of annual energy savings with Type C wall that would have a green wall installed in front of it (Figure 6).

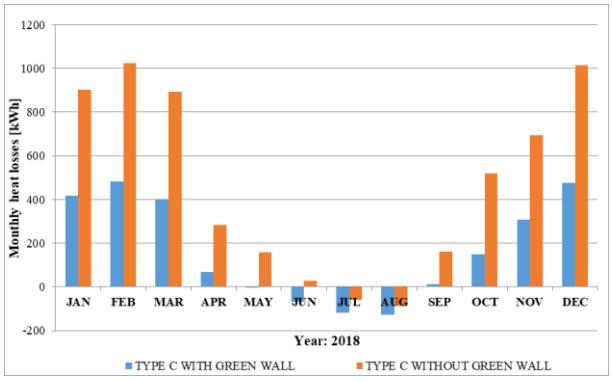


Figure 6. Comparison of monthly heat losses between Type C wall structure with the green wall and without it

For the calculation of CO_2 emissions reduction, we calculated how many CO_2 emissions are produced »in our pilot building« before and after the installment of a green wall. The more energy that a building requires to achieve thermal comfort, the more CO_2 emissions are produced. Not only that green walls make the structure more energy-efficient (also reducing greenhouse gases) but also, due to having greenery as part of the structure, green walls filter particulate matter from the air and convert CO_2 into oxygen. This essentially means an »extra production« of purified air, and the reduction of CO_2 emissions, while serving other essential functions as a wall assembly. According to [8], 1 m^2 of living green wall extracts 2,3 kg of CO_2 from the air per year while roughly producing 1,7 kg of oxygen at the same time. Our pilot building (with an area of 100 m^2) can, therefore, produce 170 kg of oxygen and extract 230 kg of CO_2. With that, we could assess the benchmark if the (possible) reduction of greenhouse gas emissions could contribute to a decision of installing a living green wall as a structure, even if they do not perform well as an insulative layer Yearly produced CO_2 emissions GHG_{kgCO2} are for each energy source, calculated [9]:

$$GHG_{(\text{kg CO2})} = Q_{T,l} \cdot EF \text{ [kg]} \quad (6)$$

where EF (Table 1) represents the emission factor (in kg/kWh). An emission factor essentially tells us, how many kilograms of CO_2 emissions will be produced with one kWh of input energy, respectively, later transformed to energy value E_{VALUE} (Table 1) for assessing the amount/supply of each energy source, needed for heating (in their assigned units).

Table 1. EF and E_{VALUE} values, used for heating [10]

ENERGY SOURCE	UNIT	E_{VALUE} [KWH/UNIT]	EF [KG/KWH]
Heating oil	[litre]	10,31	0,267
Natural gas	[Sm^3]	9,44	0,199
Liquefied natural gas	[kg]	12,78	0,227
Biomass	[kg]	4,17	0,000
District heating	[kWh]	1,00	0,356
Electricity	[kWh]	1,00	0,509

Equation 6 is presented for walls without the installed green wall. With the installed green wall, we need to retract 230 kg from the equation.

Table 2. Annual CO_2 emissions reduction (in kg) for each energy source for each wall type assembly

TYPE A	**Type A without green wall**	**Type A with green wall**
Energy source	GHG_{kgCO2} [kg]	GHG_{kgCO2_GW} [kg]
Heating oil	415,2	38,1
Natural gas	309,4	**-30,2**
Liquefied natural gas	353,0	**-3,0**
Biomass	0,0	**-230,0**
District heating	553,6	127,5
Electricity	791,5	281,1
TYPE B	**Type B without green wall**	**Type B with green wall**
Heating oil	4.126,8	1.159,1
Natural gas	3.075,7	805,3
Liquefied natural gas	3.508,5	951,0
Biomass	0,0	**-230,0**
District heating	5.502,3	1.622,1
Electricity	7.867,1	2.218,1
TYPE C	**Type C without green wall**	**Type C with green wall**
Heating oil	1.554,9	472,1
Natural gas	1.158,9	293,3
Liquefied natural gas	1.322,0	366,9
Biomass	0,0	**-230,0**
District heating	2.073,2	706,2
Electricity	2.964,2	1.108,5

The potential installment of a green wall in front of the main Type A wall assembly would drastically reduce CO_2 emissions, also where the presented values are negative, this means that the reduction of CO_2 emissions goes so beyond, that it is an indication that with the green wall, more CO_2 emissions are extracted from the air than produced (the structure becomes carbon negative). Green walls would then effectively contribute to the reduction of fuel consumption, needed for heating the pilot case building. Naturally, energy savings are the biggest in the case of Type B wall assembly, regarding their quantity. Quite opposite is with Type A wall assembly (the most energy-efficient wall assembly in our research) where consumption of energy sources, regarding their quantity, would be the smallest.

5 Conclusion

Simulation models should be treated only as a simple interpretation of calculating heat losses through such wall assemblies, more complex calculations would require computations for a real-time practical case. The calculation of CO_2 emissions reduction was connected with the mentioned annual heat losses. The assessment of models was based on mathematical, predetermined terms only, nonetheless, these models serve as good analytical research for exploring such matter in an area of innovative approach in energy efficiency.

With simulations, we concluded that, indeed, heat losses for each wall type respectively, would decrease on an annual level, if each type would have a green wall installed in front of the main wall. We can also assume due to results that, the more energy-inefficient the wall is, the better is to install a vertical green wall in front of it. However, the green wall cannot be a »layer« that could single-handedly solve the problem of energy inefficiency of the main wall. Regarding CO_2 emissions: such emissions would decrease drastically.

The values of a real-time »pilot building« would most likely differ to some extent, which would present us with a more explicit way of seeing such matters with a better understanding of its process of functioning. Regardless of results, the development and demand of such innovative structures are rising so there will sure be more research-based, complex models of calculating heat losses through wall assemblies with living green walls and even assessments of CO_2 emissions reduction as well. This paper serves as a good analytical tool with an explored background of such matter, yet surely requires further research in the upcoming future.

References

[1] D. Tasić: Analytical modelling of achieving energy efficiency with green walls. Master's thesis. University of Maribor, Faculty of Energy Technology, 2019.

[2] https://www.naava.io/editorial/what-are-green-walls [26.6.2020]

[3] JFS, https://www.japanfs.org/en/news/archives/news_id030614.html [26.6.2020]

[4] F. Convertino, G. Vox, E. Schettini: Heat transfer mechanisms in vertical green systems and energy balance equations. International Journal of Design & Nature and Ecodynamics, vol. 14, No. 1, January 2019, p.p. 7-18.

[5] C. M. Hui, Z. S. Zheng: Thermal regulation performance of green living walls in buildings. Joint Symposium 2013: Innovation and Technology for Built Environment, 12. November 2013. Tsim Sha Tsui East, Kowloon, Hong Kong. p.p. 1-12.

[6] M. Kmieć: Green wall technology. Cracow University of Technology, Faculty of Architecture, Institute of Building Design. Published in: Czasopismo Techniczne, Architektura Zeszyt 10 A (24), 2014, p.p. 47-60.

[7] J.A. Oosterlee: Green Walls and building energy consumption (Part 1: Building energy simulation). Master's thesis. Eindhoven: Eindhoven University of Technology, Faculty of Architecture, Building and Planning, 2013.

[8] https://www.sempergreen.com/en/solutions/living-wall/living-wall-benefits [26.6.2020]

[9] https://coolaustralia.org/wp-content/uploads/2013/12/Calculating-GHG-emissions.pdf [26.6.2020]

[10] https://ceu.ijs.si/izpusti-co2-tgp-na-enoto-elektricne-energije/ [26.6.2020]

Vpliv viskoznosti nosilne tekočine na izgube magnetne tekočine v izmeničnem magnetnem polju

Jakob Vizjak, Miloš Beković in Anton Hamler

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, 2000 Maribor, Slovenija
E-pošta: jakob.vizjak1@um.si

Influence of carrier fluid viscosity on magnetic fluid losses in an alternating magnetic field

Abstract. *In magnetic fluids, different carrier fluids have different viscosities. The result is differences in magnetic fluid losses when exposed to an alternating magnetic field. Differences in losses due to various viscosities are theoretically evaluated using a mathematical model that is based on relaxation losses, which are the result of Brownian and Néel relaxation. It is found that fluids with lower viscosity show higher losses compared to fluids with higher viscosities. Furthermore, the change in losses is small in comparison to change in losses when other parameters are changed, such as magnetic field amplitude or volume fraction of magnetic nanoparticles.*

1 Uvod

Magnetne tekočine so suspenzija[1] magnetnih nanodelcev v nosilni tekočini. Pod vplivom magnetnega polja, takšna tekočina izkazuje magnetne lastnosti.

Magnetna tekočina je sestavljena iz nosilne tekočine, kot je na primer voda, kerozin ali pa kakšno olje. V nosilni tekočini se nahajajo magnetni nanodelci velikosti 3 – 15 nm [1]. Običajno se za nanodelce uporablja železove okside, kot so na primer magnetit (Fe_3O_4), hematit (α-Fe_2O_3) ali maghemit (γ-Fe_2O_3), uporabljajo se pa tudi druge magnetne spojine [2]. Pomembno za te delce je, da so dovolj majhni, da imajo samo eno magnetno domeno, saj takrat izkazujejo superparamagnetno obnašanje. Za boljšo stabilnost proti aglomeraciji[2] se delce prevleče s surfaktantom[3], ki mora biti izbran tako, da se ujema z dielektričnimi lastnostmi nosilne tekočine. Za to je pogosta uporaba oleinske kisline [3].

Magnetne tekočine se uporabljajo v tehniki, na primer za dinamično tesnjenje rotirajočih teles (trdi diski). Prav tako se uporabljajo v medicini, za ciljno doziranje zdravil ali pa za hipertermično segrevanje rakavega tkiva.

V primeru hipertermične aplikacije so pomembne grelne lastnosti magnetne tekočine. Te so povezane z močjo izgub v časovno odvisnem magnetnem polju, saj je stopnja gretja višja, kadar so višje izgube. Na izgube vplivajo različni parametri, kot so na primer jakost magnetnega polja, frekvenca, volumska koncentracija magnetnih delcev. Med te parametre spada tudi viskoznost. Zato smo teoretično ovrednotili vpliv nosilnih tekočin, ki imajo različne viskoznosti na moč izgub [4].

2 Metode in materiali

2.1 Matematični model izračuna izgub

Izhajali smo iz relaksacijskih[4] izgub magnetne tekočine v časovno spreminjajočem se magnetnem polju. Za izračun izgub, smo uporabili enak model, kot je bil uporabljen v [5], kjer so ugotovili dobro ujemanje tega modela z meritvami po kalorimetrični metodi.

Izhajali smo iz skupnega relaksacijskega časa τ:

$$\tau^{-1} = \tau_B^{-1} + \tau_N^{-1} \tag{1}$$

Ta čas je sestavljen iz dveh delov. Prvi je Brownov relaksacijski čas τ_B, ki ga opisuje enačba (2), drugi je Néelov relaksacijski čas, ki ga opisuje enačba (3) [5].

$$\tau_B = \frac{3V_h\eta}{k_BT}, \tag{2}$$

$$\tau_N = \tau_0 e^{\frac{KaVp}{k_BT}}. \tag{3}$$

Pri tem je η viskoznost, V_h je hidrodinamični volumen delca s surfaktantom, k_B je Boltzmannova konstanta, T je temperatura, K_a je konstanta anizotropije, V_p je volumen magnetnega delca in $\tau_0 = 10^{-9}$ s je eksponentni faktor.

Skupni relaksacijski čas nastopa v magnetizaciji tekočine **M**, ki se za posamično komponento x in y zapiše kot:

$$\hat{M}_i = \chi_i \frac{\hat{h}_i}{j\omega\tau+1}, \tag{4}$$

kjer je i indeks, ki označuje posamično komponento, $\omega = 2\pi f$ je frekvenca magnetnega polja, $\hat{h}_x = H_x$ ter $\hat{h}_y = H_y$ pa sta kompleksni komponenti vzbujenega magnetnega polja. Susceptibilnost χ_i se zapiše s pomočjo Langevinove funkcije kot:

$$\chi_i = \frac{\phi M_d}{H_i}\left(\coth(\alpha_i) - \frac{1}{\alpha_i}\right), \tag{5}$$

[1] Zmes tekočine in v njej netopne trdne snovi.
[2] Sprijemanje delcev.
[3] Spojina, ki znižuje površinsko napetost med dvema snovema.
[4] Relaksacija predstavlja povratek sistema v ravnovesje in se jo karakterizira z relaksacijskim časom.

kjer je ϕ volumska koncentracija magnetnih delcev v tekočini, M_d je magnetizacija nasičenja nakopičenega materiala in α_i je Langevinov parameter, izražen kot:

$$\alpha_i = \frac{\mu_0 V_p M_d H_i}{k_B T}. \quad (6)$$

Tukaj je μ_0 permeabilnost praznega prostora. Časovno povprečena moč na enoto prostornine P se na koncu dobi kot:

$$P = \frac{\mu_0}{2}\mathrm{Re}\left(j\omega\hat{M}_x\hat{h}_x^* + j\omega\hat{M}_y\hat{h}_y^*\right). \quad (7)$$

To omogoča računanje izgub v rotacijskem magnetnem polju, ki je razdeljeno na x in y komponetni. Mi smo računali le za izmenično polje v x smeri. Za določanje grelnih zmožnosti magnetne tekočine se pogosto uporablja parameter SAR (specific absorption rate), ki pomeni specifična stopnja absorpcije. Je tudi najpogostejši parameter za eksperimentalno določitev izgub [5]. Izračunano vrednost izgub po enačbi (7), lahko preračunamo v parameter SAR kot [5]:

$$SAR = \frac{P}{\rho_m}, \quad (8)$$

kjer je ρ_m masna gostota magnetnih nanodelcev. Eksperimentalno se lahko SAR določi na več načinov. Najpogostejša je kalorimetrična metoda, kjer se po meritvi krivulje segrevanja SAR izračuna po enačbi (9) [5].

$$SAR = \frac{c\rho}{m_{Fe}}\left(\frac{\mathrm{d}T}{\mathrm{d}t}\right)_{max}, \quad (9)$$

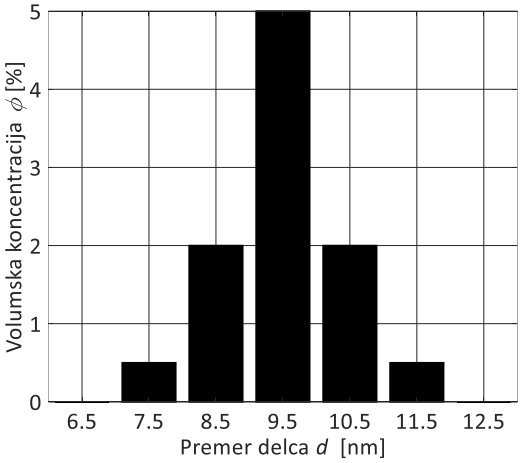
kjer je c specifična toplota, ρ je gostota vzorca, m_{Fe} masa magnetnih nanodelcev na enoto volumna in $(\mathrm{d}T/\mathrm{d}t)_{\max}$ je maksimalna vrednost časovnega odvoda merjene temperature.

Poračunali smo odvisnosti izgubne moči od temperature, volumske koncentracije in jakosti magnetnega polja. Pri izračunih smo upoštevali parametre, ki jih prikazuje tabela 1.

Tabela 1. Parametri za izračun moči izgub. Parameter d_s predstavlja debelino surfaktanta na površini magnetnega delca.

Parameter	Vrednost
K_a	46,8 kJ/m^3
M_d	400 kA/m
d_s	0,5 nm
f	400 kHz

Pri izračunih odvisnosti od temperature in odvisnosti od volumske koncentracije smo uporabili amplitudo jakosti magnetnega polja $H = 3$ kA/m. Pri odvisnosti od volumske koncentracije in odvisnosti od jakosti magnetnega polja smo uporabili konstantno temperaturo $T = 20$ °C. Pri odvisnosti od volumske koncentracije smo predpostavili velikost delcev $d = 9{,}5$ nm. Pri preostalih dveh odvisnostih smo uporabili porazdelitev delcev, kot jo prikazuje slika 1, pri čemer je bila skupna volumska koncentracija delcev postavljena na 10 %.



Slika 1. Porazdelitev delcev za izračune odvisnosti izgub magnetne tekočine od temperature in magnetnega polja.

Porazdelitev delcev smo upoštevali tako, da smo, kot so v literaturi [5], izračunali vsoto posameznih prispevkov posamezne skupine velikosti. To je prikazano z enačbo (10).

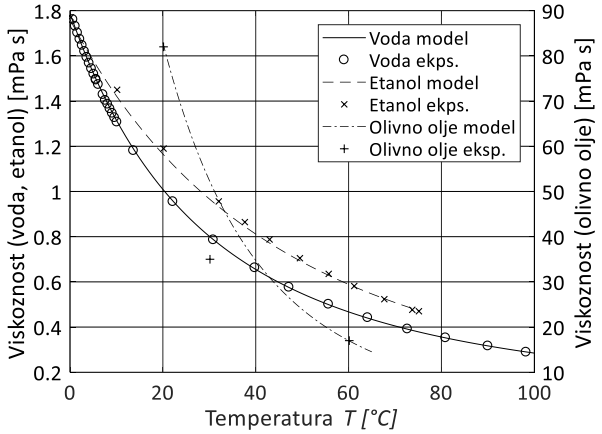
$$P = \frac{\mu_0}{2}\left(\sum_{k=1}^{n}\mathrm{Re}\left(j\omega\hat{M}_{xk}\hat{h}_x^* + j\omega\hat{M}_{yk}\hat{h}_y^*\right)\right). \quad (10)$$

Pri tem je n število skupin, ki je bilo v našem primeru $n = 5$.

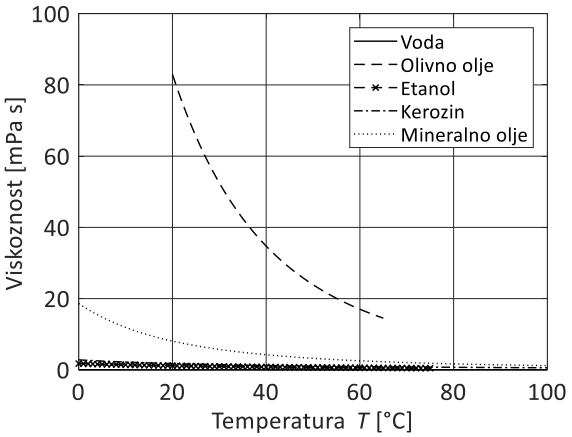
Analizirali smo pet možnih nosilnih tekočin: voda , olivno olje, etanol, kerozin RP1 [6] in mineralno olje 23D [7]. Za vodo, olivno olje in etanol smo uporabili model (11) iz literature [8], kjer so bili parametri A, B in C prilagojeni tako, da se čim bolje ujemajo z meritvami. To ujemanje prikazuje slika 2, kjer so podatki meritev vzeti iz literature [8].

$$\log\eta = A + \frac{B}{C-T}. \quad (11)$$

Tukaj je η viskoznost tekočine. Parametre za tri tekočine pa prikazuje tabela 2.



Slika 2. Primerjava modela (11) z izmerjenimi vrednostmi.



Slika 3. Primerjava temperaturnih odvisnosti obravnavanih tekočin.

Tabela 2. Parametri za enačbo (11) za vodo, olivno olje in etanol, pridobljeni iz [8].

Tekočina	*A*	*B*	*C*
Voda	- 4,5318	- 220,57	149,39
Olivno olje	- 4,9110	- 699,70	110,30
Etanol	- 2,52	- 799,10	- 15,9

Za kerozin smo uporabili kombinacijo Vogelove relacije temperature in viskoznosti, ter Basusove relacije tlaka in viskoznosti [6]:

$$\eta(T,p) = \eta_0 e^{\frac{A}{T+D_V}+\frac{pB}{T+D_p}}. \qquad (12)$$

Kjer so konstante, pridobljene iz [6]:
$\eta_0 = 6{,}33\cdot10^{-5}$ Pa s,
A = 542,4 °C/GPa,
B = 544 °C,
D_V = 143 °C,
D_p = 25,6 °C.
Za tlak smo pri tem upoštevali normalni zračni tlak: p = 101325 Pa. Za mineralno olje pa smo upoštevali relacijo Vogel, Tammann in Fulcher [7]:

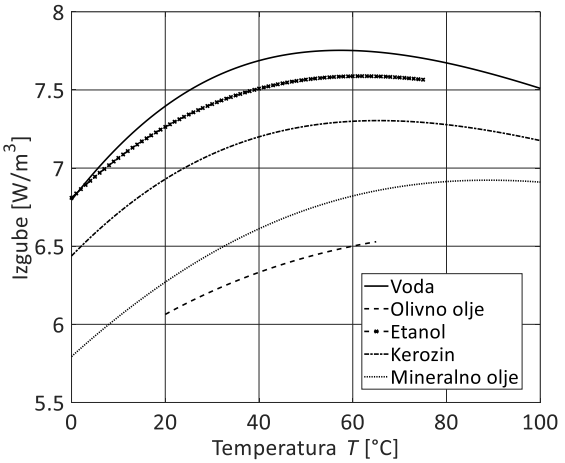
$$\eta = \eta_\infty e^{\frac{D_F T_\infty}{T-T_\infty}}, \qquad (13)$$

kjer pa so konstante [7]:
η_∞ = 0,0300 mPa s,
T_∞ = 139 K,

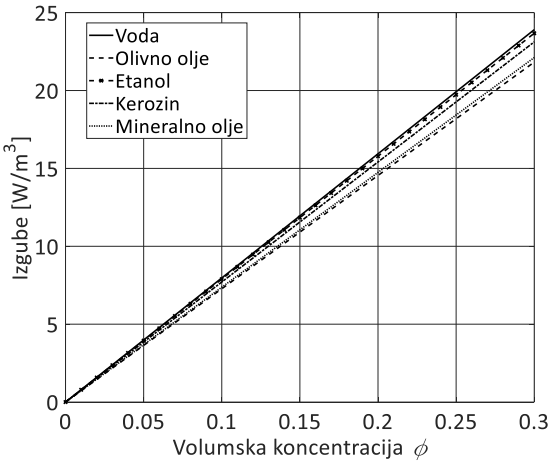
3 Rezultati

Temperaturne odvisnosti viskoznosti petih tekočin izračunane po modelih (11), (12) in (13) prikazuje slika 3.

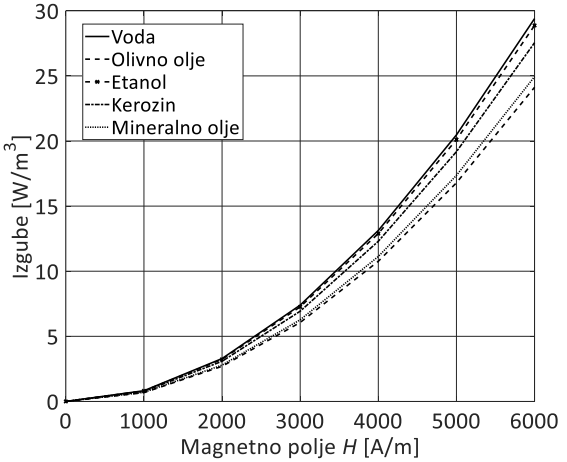
Iz tega je vidno, da ima najvišjo viskoznost olivno olje, potem mineralno olje, nato pa še kerozin. Voda in etanol imata podobno viskoznost, pri čemer je viskoznost etanola nekoliko višja.
Rezultat izračuna izgub po enačbi (7) prikazujejo slika 4, slika 5 in slika 6.



Slika 4. Rezultati izračuna izgub v odvisnosti od temperature.

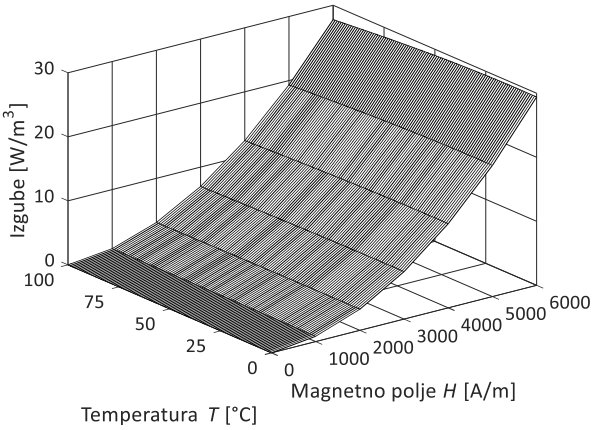


Slika 5. Rezultati izračuna izgub v odvisnosti od volumske koncentracije magnetnih nanodelcev.



Slika 6. Rezultati izračuna izgub v odvisnosti od amplitude jakosti magnetnega polja.

Iz teh treh grafov je razvidno, da so izgube višje, kadar je viskoznost nosilne tekočine nižja, vendar pa je sprememba izgub višja pri spremembi volumske koncentracije ali magnetnega polja, kot pri spremembi nosilne tekočine. Pri temperaturi $T = 20$ °C je, glede na izračunane temperaturne odvisnosti viskoznosti, viskoznost olivnega olja približno 82-krat toliko, kot viskoznost vode. Pri izračunih odvisnosti izgub od volumske koncentracije (slika 5) to pomeni, da so pri $\phi = 0{,}3$, zaradi različne viskoznosti, izgube pri vodi približno 10 % višje kot izgube pri olivnem olju. Pri spremembi ϕ iz 0,2 na 0,3 pa se izgube povečajo za približno 50 %. Podobno so pri izračunih odvisnosti izgub od magnetnega polja (slika 6), pri $H = 6$ kA/m, zaradi različne viskoznosti, izgube pri vodi za približno 22 % višje kot izgube pri olivnem olju, medtem ko se izgube pri spremembi jakosti magnetnega polja H iz 4 kA/m na 6 kA/m izgube povečajo za približno 124 %. Slika 7 prikazuje še 3-D prikaz odvisnosti izgub od temperature in magnetnega polja za vodo, pri porazdelitvi delcev, ki jih kaže slika 1.



Slika 7. 3-D prikaz odvisnosti izgub od temperature in magnetnega polja za primer vode.

4 Zaključek

Z matematičnim modelom za izračun izgub magnetne tekočine v magnetnem polju smo preverili, kakšen vpliv ima viskoznost nosilne tekočine na rezultat izračuna. Ugotovili smo, da viskoznost ima vpliv na izračun in sicer so izgube večje v primeru, ko je viskoznost nosilne tekočine nižja. Tako so bile v naši obravnavi petih tekočin izgube največje v primeru vode in najnižje v primeru olivnega olja. So pa razlike v izgubah zaradi viskoznosti manjše v primerjavi z razlikami v izgubah zaradi spremembe drugih parametrov, kot sta volumska koncentracija in pa amplituda jakosti magnetnega polja. Zato sklepamo, da je pri izbiri nosilne tekočine, sploh za namene magnetne hipertermije, najprej bolj smiselno upoštevati njene druge lastnosti, na primer biokompatibilnost, kot pa viskoznost. Če pa je cilj pridobiti čim večje izgube, da je gretje magnetne tekočine čim višje in učinkovito, pa je izbira nosilne tekočine zaradi njene viskoznosti vseeno pomembna.

Literatura

[1] R. E. Rosensweig, *Ferrohydrodynamics*. Mineola, NY: Dover, 2014.

[2] I. Sharifi, H. Shokrollahi, and S. Amiri, "Ferrite-based magnetic nanofluids used in hyperthermia applications," *J. Magn. Magn. Mater.*, vol. 324, no. 6, pp. 903–915, 2012.

[3] S. Genc and B. Derin, "Synthesis and rheology of ferrofluids: a review," *Curr. Opin. Chem. Eng.*, vol. 3, pp. 118–124, 2014.

[4] J. Vizjak, "Analiza vpliva magnetnega polja na viskoznost in grelno moč magnetne tekočine," University of Maribor, 2019.

[5] M. Beković *et al.*, "Numerical Model for Determining the Magnetic Loss of Magnetic Fluids," *Materials (Basel).*, vol. 12, p. 591, 2019.

[6] J. L. Tevaarwerk, *The Measurement, Modeling, and Prediction of Traction for Rocket Propellant 1*. NASA, 1989.

[7] S. Bair, "Temperature and Pressure Dependence of Density and Thermal Conductivity of Liquids BT - Encyclopedia of Tribology," Q. J. Wang and Y.-W. Chung, Eds. Boston, MA: Springer US, 2013, pp. 3527–3533.

[8] D. S. Viswanath, T. K. Ghosh, D. H. L. Prasad, N. V. K. Dutt, and K. Y. Rani, *Viscosity of Liquids: Theory, Estimation, Experiment, and Data*. Springer Netherlands, 2007.

Model globokih gaussovskih procesov za identifikacijo procesa Bouc-Wen

Tadej Krivec[1,2], **Juš Kocijan**[1,3]

[1]*Institut Jožef Stefan, Jamova cesta 39, Ljubljana, Slovenija*
[2]*Mednarodna podiplomska šola Jožefa Stefana, Jamova cesta 39, Ljubljana, Slovenija*
[3]*Univerza v Novi Gorici, Vipavska 13, Nova gorica, Slovenija*
E-pošta: tadej.krivec@ijs.si, jus.kocijan@ijs.si

Deep Gaussian process model for the identification of the Bouc-Wen process

Gaussian process regression is a well know probabilistic machine learning method that can be used for dynamic system identification where the goal is to find a simulation model. Full Gaussian process regression has a cubic complexity with respect to the input data. Different approximations are used which make a compromise between the flexibility of the model and the computational demand. This article presents the use of a deep Gaussian process, a hierarchical Gaussian process that provides a very flexible structure on the cost of increasing the computational complexity, to model a dynamic system. Since this kind of architecture is analytically intractable, doubly stochastic variational inference is used for approximation. Prediction and simulation results are compared between different number of hidden layers for modeling the Bouc-Wen process.

1 Uvod

Idejo o modeliranju z gaussovskimi procesi je področju strojnega učenja približal C.E. Rasmussen s svojo doktorsko dizertacijo [1]. Gaussovski procesi ponujajo alternativo bolj priljubljenemu pristopu k modeliranju z nevronskimi mrežami. Monografija s tega področja je [2], modeliranje dinamičnih sistemov z gaussovskimi procesi pa je opisano v [3]. Glavni problem metode predstavlja kubična računska kompleksnost glede na število vhodnih podatkov. Pregled približkov za znižanje računske kompleksnosti lahko najdemo v [4]. Med približki, ki se lahko zelo dobro skalirajo na veliko količino podatkov je stohastični variacijski gaussovski proces [5]. Ta približek je osnova za širjenje ideje modeliranja na hierarhično strukturo, podobno nevronskim mrežam, kjer so nivoji nevronov zamenjani z gaussovskimi procesi. Takemu modelu pravimo globok gaussovski process in je predstavljen v [6]. V tem prispevku predstavljamo uporabo globokih gaussovskih procesov za identifikacijo dinamičnih sistemov na primeru procesa Bouc-Wen v kombinaciji z avtoregresijskim modelom.

2 Nelinearna identifikacija dinamičnih sistemov

S hitrim razvojem informacijske tehnologije, predvsem z znižanjem stroškov shranjevanja podatkov in računalniškega obdelovanja podatkov, se popularnost in uporabnost metod za identifikacijo dinamičnih sistemov hitro zvišuje. V tem poglavju je predstavljen možen podatkovni način kako pristopiti k reševanju identifikacije dinamičnih sistemov z metodo gaussovskih procesov.

2.1 Avtoregresijski model

Nelinearni dinamični sistem lahko modeliramo kot

$$\boldsymbol{y}(k) = f(\boldsymbol{z}(k), \boldsymbol{\theta}) + \boldsymbol{\nu}(k), \tag{1}$$

kjer k predstavlja zaporedno število vzorca, f nelinearno preslikavo, $\boldsymbol{\theta}$ vektor parametrov, $\boldsymbol{\nu}(k)$ gaussovski beli šum in $\boldsymbol{z}(k), \boldsymbol{y}(k)$ predstavljata vektor vhodov in izhodov v odvisnosti od k. Predpostavimo, da interne strukture sistema, ki ga identificiramo, ne poznamo (ang. blackbox). Manipuliramo lahko z vhodnim signalom in merimo izhod. V našem primeru predpostavimo, da ima model avtoregresijsko strukturo (ang. Nonlinear Autoregressive Exogenous Model). V takem primeru je vhodni vektor definiran kot

$$\begin{aligned}\boldsymbol{z}(k) = [&\boldsymbol{y}(k-n_p), \ldots, \boldsymbol{y}(k-n_p-n_a+1)\\ &\boldsymbol{u}(k-n_k), \ldots, \boldsymbol{u}(k-n_k-n_b+1)],\end{aligned} \tag{2}$$

kjer k predstavlja zaporedno število vzorca, n_p, n_a, n_k, n_b pa parametre zakasnitev vzorcev. Identifikacija dinamičnih sistemov je proces, kjer v našem primeru iščemo ustrezne parametre avtoregresijskega modela in modeliramo nelinearno preslikavo med vhodnim in izhodnim vektorjem. Tak model lahko uporabljamo za napovedovanje časovnih vrst, vodenje dinamičnih sistemov, detekcijo napak itd. Uporablja se lahko za napovedovanje ali simulacijo, kjer pri napovedi napovedujemo za en časovni korak vnaprej, pri simulaciji pa iterativno izračunamo bodoče vrednosti izhoda za poljubno število korakov vnaprej po enačbi

$$\hat{\boldsymbol{y}}(k) = f(\hat{\boldsymbol{z}}(k), \boldsymbol{\theta}), \tag{3}$$

kjer velja

$$\begin{aligned}\hat{\boldsymbol{z}}(k) = [&\hat{\boldsymbol{y}}(k-n_p), \ldots, \hat{\boldsymbol{y}}(k-n_p-n_a+1)\\ &\boldsymbol{u}(k-n_k), \ldots, \boldsymbol{u}(k-n_k-n_b+1)]\end{aligned} \tag{4}$$

in $\hat{\boldsymbol{z}}$ in $\hat{\boldsymbol{y}}$ predstavljata ocenjene (napovedane) vrednosti iz prejšnih korakov.

2.2 Model gaussovskih procesov za dinamične sisteme

Model gaussovskih procesov je bayesovska metoda strojnega učenja, ki posplošuje gaussovsko porazdelitev naključnih spremenljivk ali vektorjev funkcijskih vrednosti. Predpostavimo, da imamo znane meritve vhodov in izhodov v obliki

$$\boldsymbol{y} = f(\boldsymbol{z}) + \boldsymbol{\epsilon}, \tag{5}$$

kjer $\boldsymbol{z}$ predstavlja vhodni vektor v obliki avtoregresijskega modela, f latentno funkcijo, $\boldsymbol{y}$ opazovane izhodne podatke popačene s šumom $\boldsymbol{\epsilon}$. Predpostavimo, da je šum neodvisno in identično porazdeljen z gaussovsko porazdelitvijo

$$\epsilon_n \sim \mathcal{N}(0,\, \sigma_n^2). \tag{6}$$

Vektor $\boldsymbol{f} = [f_1, f_2, ..., f_n]^T$ predstavlja vrednosti latentne funkcije kjer velja $f_i = f(\boldsymbol{z_i})$. Stolpični vektor vhodov dolžine n sestavlja matriko $\mathbf{Z} \in \mathcal{R}^{D\times n}$, kjer je poljubna vrstica predstavljena z vektorjem $\boldsymbol{z}$. Množica podatkov je potem označena kot $\mathcal{D} = (\mathbf{Z}, \boldsymbol{y})$. Cilj je poiskati porazdelitev napovedi neznanega vektorja latentnih funkcijskih vrednosti $\boldsymbol{f}_*$ pri znanih testnih lokacijah $\boldsymbol{z_*}$.

Gaussovski proces definira porazdelitev funkcij, bolj natančno porazdelitev vektorjev funkcijskih vrednosti latentne funkcije f. Definiran je s funkcijo srednje vrednosti $m(\boldsymbol{z})$ in s kovariančno funkcijo $k(\boldsymbol{z}, \boldsymbol{z\prime})$

$$m(\boldsymbol{z}) = \mathbb{E}[f(\boldsymbol{z})], \tag{7a}$$

$$k(\boldsymbol{z}, \boldsymbol{z\prime}) = \mathbb{E}[(f(\boldsymbol{z}) - m(\boldsymbol{z}))(f(\boldsymbol{z\prime}) - m(\boldsymbol{z\prime}))] \tag{7b}$$

in je označen z $f(\boldsymbol{z}) \sim \mathcal{GP}(m(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z\prime}))$. Parametri kovariančne funkcije so poimenovani kot hiperparametri [3]. V modelu gaussovskih procesov je apriori porazdelitev nad vektorjem latentnih funkcijskih vrednosti transformirana z Bayesovim izrekom v aposteriorno porazdelitev

$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f}|\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})}, \tag{8}$$

kjer $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\theta})$ prestavlja pogojno verjetnost pri definiranih hiperparametrih in opazovanim vektorjem $\boldsymbol{f}$, $p(\boldsymbol{f}|\boldsymbol{\theta})$ apriori porazdelitev nad vektorjem latentnih funkcijskih vrednosti, $p(\boldsymbol{y}|\boldsymbol{\theta})$ marginalno pogojno porazdelitev pri znanih hiperparametrih in $p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\theta})$ predstavlja aposteriorno porazdelitev vektorja latentnih funkcijskih vrednosti. V nadaljevanju predpostavimo pogojenost porazdelitve glede na hiperparametre $\boldsymbol{\theta}$ in jih v notaciji izpustimo.

Kovariančna funkcija definira apriori porazdelitev vektorja latentnih funkcijskih vrednosti. Kovariančna funkcija je lahko katerakoli funkcija, ki tvori simetrično, semipozitivno in kvadratno matriko. Popularna izbira je kovariančna funkcija RBF (ang. Radial Basis Function)

$$k(\boldsymbol{z}, \boldsymbol{z\prime}) = \sigma_f^2 e^{-\frac{||\boldsymbol{z}-\boldsymbol{z\prime}||^2}{2l^2}}. \tag{9}$$

Po definiciji gaussovskega procesa ima skupna porazdelitev med vektorjem latentnih funkcijskih vrednosti $\boldsymbol{f}$ in testnim vektorjem latentnih funkcijskih vrednosti $\boldsymbol{f_*}$ tudi skupno gaussovsko porazdelitev. Apriori skupna porazdelitev je potem transformirana z Bayesovim izrekom ob znanem vektorju vrednosti izhodov na aposteriorno skupno porazdelitev definirano z

$$p(\boldsymbol{f}, \boldsymbol{f}_*|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{f}_*)}{p(\boldsymbol{y})}, \tag{10}$$

kjer je apriori skupna porazdelitev definirana z

$$p(\boldsymbol{f}, \boldsymbol{f}_*) \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{nn} & \mathbf{K}_{n*} \\ \mathbf{K}_{*n} & \mathbf{K}_{**} \end{bmatrix}) \tag{11}$$

in $\mathbf{K}_{nn}$ predstavlja kovariančno matriko opazovanih učnih podatkov, $\mathbf{K}_*$ kovariančno matriko učnih in testnih podatkov and $\mathbf{K}_{**}$ kovariančno matriko testnih podatkov. Pogojna porazdelitev $p(\boldsymbol{y}|\boldsymbol{f})$ je definirana kot

$$p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}, \sigma_n^2\mathbf{I}), \tag{12}$$

kjer $\mathbf{I}$ predstavlja matriko identitete in σ_n^2 standardni odklon šuma. Hiperparametre učimo z maksimizacijo logaritma marginalne pogojne porazdelitve $p(\boldsymbol{y})$. Porazdelitev napovedi je izračunana z marginalizacijo $\boldsymbol{f}$ iz skupne pogojne porazdelitve

$$p(\boldsymbol{f}_*|\boldsymbol{y}) = \int p(\boldsymbol{f}, \boldsymbol{f}_*|\boldsymbol{y})df. \tag{13}$$

Pri iterativni simulaciji je vhodni vektor v poljubnem koraku stohastičen in odziv nima analitične rešitve. V tem primeru se odziv simulacije v poljubnem koraku izračuna kot

$$p(\boldsymbol{f}_*|\boldsymbol{y}) \approx \frac{1}{S}\sum_{i=1}^{S} \int p(\boldsymbol{f}, \boldsymbol{f}_*|\boldsymbol{y}, \boldsymbol{z}_*)df, \tag{14}$$

kjer S predstavlja število vzorcev iz stohastičnega vhodnega vektorja $\boldsymbol{z}_*$. Model gaussovskih procesov ima kubično računsko kompleksnost, ki se lahko izboljša z aproksimacijskimi metodami.

2.3 Stohastični variacijski gaussovski proces

Stohastični variacijski gaussovski proces je aproksimacijska metoda za znižanje računske kompleksnosti modela na podlagi gaussovskih procesov. Predpostavimo, da obstaja vektor vhodnih psevdo lokacij $\boldsymbol{z}_u$ in pripadajoči psevdo vektor latentnih spremenljivk $\boldsymbol{u} = [u_1, \ldots, u_m]$, ki je skupno gaussovsko porazdeljen z naključnima spremenljivkama $\boldsymbol{f}$ in $\boldsymbol{f}_*$. Predpostavimo, da sta $\boldsymbol{f}$ in $\boldsymbol{f}_*$ pogojno neodvisna ob znanem $\boldsymbol{u}$. Aposteriorna skupna porazdelitev je potem definirana z

$$p(\boldsymbol{f}, \boldsymbol{f}_*, \boldsymbol{u}, |\boldsymbol{y}) = p(\boldsymbol{f}|\boldsymbol{u}, \boldsymbol{y})p(\boldsymbol{f}_*|\boldsymbol{u})p(\boldsymbol{u}). \tag{15}$$

Hensman [5] definira variacijski približek kot

$$q(\boldsymbol{f}, \boldsymbol{f}_*, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{f}_*|\boldsymbol{u})q(\boldsymbol{u}), \tag{16}$$

kjer je $q(\boldsymbol{u})$ gaussovska porazdelitev definirana s $q(\boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Lambda})$. Spodnja meja marginalne pogojne porazdelitve je potem izračunana z minimizacijo divergence Kullback Leibler (KL) med aposteriorno porazdelitvijo definirano z enačbo (15) in variacijskim približkom definiranim z enačbo (16) in je prikazana z enačbo (17).

$$\begin{aligned} \log p(\boldsymbol{y}) \geq & \int [\log p(\boldsymbol{y}|\boldsymbol{f})]q(\boldsymbol{f})d\boldsymbol{f} \\ & - \mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})]. \end{aligned} \tag{17}$$

Spodnja meja marginalne pogojne porazdelitve se lahko poenostavi. Izraz $p(\boldsymbol{y}|\boldsymbol{f})$ se faktorizira, $q(\boldsymbol{f}_i)$ pa je odvisen le od pripadajočih vhodov $\boldsymbol{z}_i$. Poenostavljena spodnja meja marginalne pogojne porazdelitve je potem definirana kot

$$\begin{aligned}\log p(\boldsymbol{y}) \geq \sum_{n=1}^{N} \int [\log p(y_n|f_n)] q(f_n) df \\ - \mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})].\end{aligned} \tag{18}$$

Spodnja meja marginalne pogojne porazdelitve se lahko izračuna stohastično, kjer naključno vzamemo le podmnožico originalnih podatkov za izračun ene iteracije optimizacije (ang. batch optimisation). Porazdelitev vektorja napovedi testnih latentnih funkcijskih vrednosti ob znanem $\boldsymbol{y}$ je izračunana kot

$$q(\boldsymbol{f}_*) = \int p(\boldsymbol{f}_*|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{u}. \tag{19}$$

Simulacijski odziv se izračuna z enačbo (14).

3 Globoki gaussovski proces

Fleksibilnost stohastičnega gaussovskega procesa je lahko omejena z izbiro kovariančne funkcije. Kovariančna funkcija definira družino funkcij, ki jih lahko modeliramo. Na ta način vnesemo apriori znanje o modelirani preslikavi, posledično pa dodamo tudi omejitve, če o procesu ne vemo nič. Globoki gaussovski procesi podobno kot globoke nevronske mreže definirajo hierarhično strukturo in povečajo fleksibilnost gaussovskiega procesa. V [7] definirajo rekurzivno apriori porazdelitev nad matriko latentnih vrednosti funkcije $\mathbf{F}$ in opazovano matriko meritev $\mathbf{Y}$, ki sta lahko poljubnih dimenzij. Tak proces ima skupno porazdelitev definirano kot

$$\begin{aligned}p(\mathbf{Y}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L) = \\ = p(\boldsymbol{y}_i|\boldsymbol{f}_i^L) \prod_{l=1}^{L} p(\mathbf{F}^l|\mathbf{F}^{l-1}, \mathbf{U}^l) p(\mathbf{U}^l),\end{aligned} \tag{20}$$

kjer $\mathbf{F}^l$ in $\mathbf{U}^l$ predstavljata vektorja latentnih in psevdo funkcijskih vrednosti v nivoju l in velja $\mathbf{F}^0 = \mathbf{Z}$. Ker ta model nima analitične rešitve, v [6] uporabijo variacijski približek skupne porazdelitve definiran kot

$$q(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L) = \prod_{l=1}^{L} p(\mathbf{F}^l|\mathbf{F}^{l-1}, \mathbf{U}^l) q(\mathbf{U}^l), \tag{21}$$

kjer $q(\mathbf{U}^l) \sim \mathcal{N}(\boldsymbol{m}^l, \boldsymbol{\Lambda}^{\mathbf{l}})$ predstavlja gaussovsko porazdelitev v nivoju l. Spodnja meja marginalne pogojne porazdelitve je izračunana z minimizacijo divergence KL med $p(\mathbf{Y}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)$ in $q(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)$ in je definirana z

$$\begin{aligned}\log p(\boldsymbol{Y}) \geq \sum_{i=1}^{N} \int [\log p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)] q(\boldsymbol{f}_i^L) \\ - \sum_{l=1}^{L} \mathrm{KL}[q(\mathbf{U}^l)||p(\mathbf{U}^l)].\end{aligned} \tag{22}$$

Spodnja meja ima podobne lastnosti kot meja v enačbi (18), kjer so i-te marginalne porazdelitve v zadnjem nivoju odvisne le od i-tih marginalnih porazdelitev v spodnjih nivojih. Zaradi tega lahko pri optimizaciji uporabimo le naključno podmnožico celotnega učnega vzorca v vsaki iteraciji. Napoved se izračuna z gaussovsko mešanico in je definirana z

$$q(\boldsymbol{f}_*^L) = \frac{1}{S} \sum_{s=1}^{S} q(\boldsymbol{f}_*^L|\boldsymbol{f}_*^{(s)L-1}), \tag{23}$$

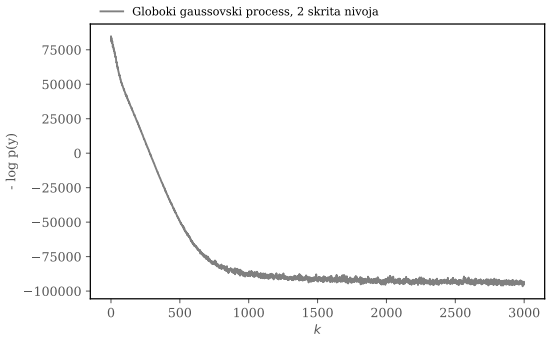
kjer vzamemo S vzorcev naključne spremenljivke $\boldsymbol{f}_*^{L-1}$. Simulacijski odziv se podobno kot pri navadnem gaussovskem procesu izračuna z enačbo (14).

4 Globok gaussovski proces za identifikacijo procesa Bouc-Wen

V tem poglavju so predstavljeni rezultati modeliranja procesa Bouc-Wen. Primerjali smo delovanje globokih gaussovskih procesov pri različnih vrednostih števila skritih nivojev in preizkusili ali hierarhična struktura izboljša rezultate napovedovanja in simulacije na testnih podatkih.

4.1 Proces Bouc-Wen

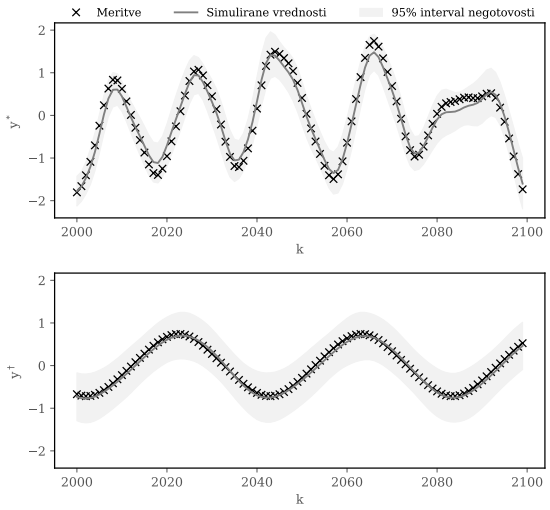
Podatkovni set procesa Bouc-Wen predstavlja vibracije Bouc-Wen sistema, ki vsebuje histerezo. Podrobnejše informacije o problemu so prestavljene v [8]. Učna množica je sestavljena iz vzbujanja s sinusnim vektorjem različnih frekvenc (ang. multi-sine) $u(k)$ in izhodnim vektorjem $y(k)$, kjer je $k = 1, \ldots, N$ in $N = 40,960$. Prva testna množica je dobljena iz vzbujanja s sinusnim vektorjem različnih frekvenc (ang. multi-sine) in vsebuje 8,192 vzorcev. Druga testna množica je dobljena iz sinusnega signala s spreminjajočo se frekvenco (ang. sine-sweep) in vsebuje 153,000 vzorcev. Parametri avtoregresijskega modela so bili izbrani kot $n_k = 0, n_b = 5, n_p = 1, n_a = 5$. Za kovariančno funkcijo smo izbrali funkcijo RBF in 100 psevdo vhodov, ki so bili inicializirani z rojenjem kot je predlagano v [9]. V vsakem koraku optimizacije smo uporabili samo naključnih 1000 točk iz celotne množice podatkov.



Slika 1: Potek marginalne pogojne porazdelitve v odvisnosti od koraka optimizacije za globok gaussovski proces z dvema skritima nivojema.

Tabela 1: Koren srednje vrednosti kvadrata napake za napoved (ang. rooted mean squared error [RMSE]) in simulacijo za proces Bouc-Wen, kjer RMSE* predstavlja napako pri sinusnem vzbujanju različnih frekvenc, RMSE† pa napako pri sinusnem vzbujanju spreminjajočih se frekvenc. Model z enim skritim nivojem je ekvivalenten stohastičnemu variacijskemu gaussovskemu procesu. Rezultati za primerjavo z drugimi metodami identifikacije procesa Bouc-Wen so na voljo v [11].

Št. nivojev	**Napovedni RMSE*** $[\times 10^{-6}]$	**Simulacijski RMSE*** $[\times 10^{-5}]$	**Napovedni RMSE†** $[\times 10^{-6}]$	**Simulacijski RMSE†** $[\times 10^{-5}]$
1	7.06 ± 0.18	13.38 ± 0.67	4.66 ± 0.27	8.75 ± 0.91
2	$\mathbf{6.10 \pm 0.15}$	$\mathbf{12.18 \pm 0.53}$	$\mathbf{3.23 \pm 0.23}$	$\mathbf{6.77 \pm 1.16}$
3	6.46 ± 0.33	16.1 ± 1.82	4.44 ± 0.58	15.21 ± 2.94
4	6.99 ± 0.32	19.5 ± 2.13	5.41 ± 0.63	19.54 ± 3.08



Slika 2: Simulacija procesa Bouc-Wen za prvo testno množico, ki je vzbujana s sinusnim signalom različnih frekvenc (y^*) in simulacija procesa Bouc-Wen za drugo testno množico, ki je vzbujana s sinusnim signalom s spreminjajočo se frekvenco ($y^\dagger$).

Rezultati

Slika 1 prikazuje logaritem marginalne pogojne porazdelitve v odvisnosti od koraka optimizacije za najboljšo rešitev, kjer smo za optimizacijski algoritem uporabili Adam [10]. Tabela 1 predstavlja rezultate korena srednje vrednosti kvadrata napake in standardni odklon dotične mere napake za obe testni množici podatkov pri čemer so bili rezultati pridobljeni z učenjem modela s 30 naključnimi ponovitvami. Vidimo lahko, da globoki gaussovski proces z 2 skritima nivojema deluje najbolje tako za napoved kot za simulacijo na prvi testni množici podatkov s sinusnim vzbujanjem različnih frekvenc. Tudi za drugo testno množico podatkov (vzbujane s sinusnim signalom s spreminjajočo se frekvenco) najbolje deluje globoki gaussovski proces z dvema skritima nivojema. Časovna odziva simulacije sta prikazana na sliki 2. Zanimiva opazka je, da se s povečevanjem števila skritih nivojev (večjega od 2) napaka povečuje. Razlog za to je lahko velika kompleksnost modela, kjer zaradi približka analitične rešitve potencialno ne dosežemo teoretične zmožnosti modela. Drugi problem lahko predstavlja občutljivost modela na parametre optimizacije zaradi katerih lahko pristanemo v lokanem minimumu cenilke, ki jo optimiziramo.

5 Zaključek

Globoki gaussovski proces predstavlja feksibilno strukturo modela za modeliranje z gaussovskimi procesi. Posledično lahko modelira bolj kompleksne procese na račun večje računalniške zahtevnosti. Rezulati vrednotenja globokih gaussovskih procesov za modeliranje dinamičnih sistemov so pozitivni. Spoznali smo, da globoki gaussovski procesi v kombinaciji z avtoregresijskim modelom izboljšajo sposobnost modeliranja v primerjavi s stohastičnim variacijskim gaussovskim procesom, ki predstavlja ekvivalent globokega gaussovskega procesa z enim skritim nivojem. V prihodnosti želimo testirati avtoregresijske modele globokih gaussovskih procesov na bolj kompleksnih primerih. Zanimiva je tudi ideja o bolj prilagojenih strukturah globokih gaussovskih procesov za modeliranje dinamičnih sistemov.

Literatura

[1] C. E. Rasmussen: Evaluation of Gaussian Processes and other Methods for Non-Linear Regression, University of Toronto, 1999.

[2] C. K. Williams, C. E. Rasmussen, Gaussian Processes for Machine Learning, no. 3, MIT press Cambridge, MA, 2006.

[3] J. Kocijan, Modelling and Control of Dynamic Systems using Gaussian Process Models, Springer, 2016.

[4] J. Quiñonero-Candela, C. E. Rasmussen, A Unifying View of Sparse Approximate Gaussian Process Regression, Journal of Machine Learning Research 6 (Dec) (2005) 1939–1959.

[5] J. Hensman, A. Matthews, Z. Ghahramani, Scalable Variational Gaussian Process Classification, JMLR, 2015.

[6] H. Salimbeni, M. Deisenroth. Doubly stochastic variational inference for Deep Gaussian Processes, Advances in Neural Information Processing Systems, 2017.

[7] A. Damianou, N. Lawrence. Deep gaussian processes, Artificial Intelligence and Statistics. 2013.

[8] J. Noël, M. Schoukens, Hysteretic Benchmark with a Dynamic Nonlinearity, in: Workshop on nonlinear system identification benchmarks, 2016, pp. 7–14.

[9] M. Bauer, M. van der Wilk, C. E. Rasmussen, Understanding Probabilistic Sparse Gaussian Process Approximations, in: Advances in neural information processing systems, 2016, pp. 1533–1541.

[10] D. P. Kingma, J. Ba, Adam, A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980.

[11] Nonlinear System Identification Benchmarks, http://www.nonlinearbenchmark.org/

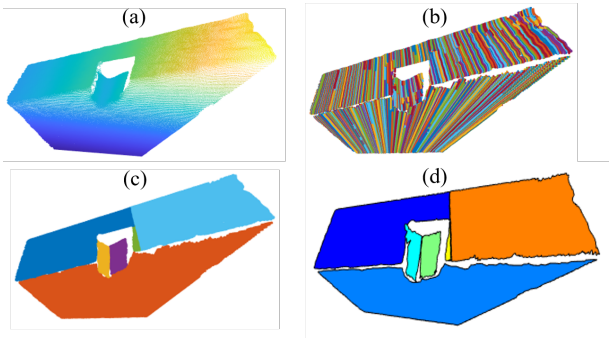
Samorazvijajoča se segmentacija globinske slike stereo kamere za notranja strukturirana okolja

Miloš Antić, Andrej Zdešar, Igor Škrjanc

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija
E-pošta: {milos.antic, andrej.zdesar, igor.skrjanc}@fe.uni-lj.si

Evolving segmentation of a stereo-camera depth map for structured indoor environments

In this paper, we present an adaptive approach to point cloud segmentation based on the EPCC (Evolving Principal Component Clustering) method. The method exploits the properties of the ordered data and their neighbor relations, which are the result of the way the data of certain sensors are captured. Adaptability refers to the use of a recursive estimate of the parameters of linear prototypes, which are used to describe clusters of similar data. The main contribution of this work is the extension and application of the EPCC method in 3D space for recursive and real-time prototype detection, such as lines and planes describing similar clusters. The developed algorithm was compared with known methods for point cloud segmentation, such as RANSAC and region growing.



Slika 1: Prikaz oblaka točk (a). Z metodo EPCC naredimo segmentacijo oblaka na daljice (b), kar uporabimo v naslednjem koraku pri določevanju ravnih površin (c). Rekonstrukcija prizora z ravnimi površinami (d).

1 Uvod

Na področju 3D-segmentacije oblakov točk je bil dosežen velik napredek. Eden glavnih dejavnikov hitrega razvoja tega področja je dostopnost nizkocenovnih senzorjev, kot so laserski merilniki razdalje, stereo kamere itd. Pojavila se je visoka potreba po 3D-razumevanju okolja, oz. rekonstrukciji in klasifikaciji tridimenzionalnih nizov točk v robotiki, za namene lokalizacije robotskih ali mobilnih sistemov in avtonomne gradnje zemljevidov oz. SLAM [1], [2], ki je lahko podlaga za planiranje poti ali sledenju trajektorijam [3], [4] . Velik pomen ima tudi 3D-modeliranje objektov in klasifikacija le-teh, kar je lahko uporabno v aplikacijah kot so manipulacija objektov z robotskimi rokami [5] in hoja četveronožnega robota po stopnicah [6].

Veliko truda je bilo vloženega v razvoj algoritmov za detekcijo ravnin v oblaku točk. Z detekcijo in predstavitvijo oblaka točk z modeli ravnin ali planarnimi območji dosežemo bolj kompaktno predstavitev tridimenzionalnega prostora (slika 1). Predlagana je bila 3D-segmentacija v zunanjih [7] in notranjih okoljih [8]. Notranja okolja so posebej prikladna za planarno segmentacijo oblaka točk, saj se v takih okoljih nahaja ogromno predmetov in površin, ki jih je naredil človek in so primerna za planarno rekonstrukcijo. Notranji prostori, katere lahko opišemo z ravninami ali pa z območji točk, ki jih pridobimo s pomočjo modelov prototipa ravnin, so torej bolj strukturirani.

Grilli in drugi so v [9] naredili analizo metod in pristopov, ki se uporabljajo pri segmentaciji oblaka točk. Izpostavili so, da je 3D-segmentacija oblaka točk zahtevna naloga, saj imajo oblaki točk običajno neenakomerno gostoto vzorčenja točk, so pogosto neurejeni in vsebujejo veliko šuma. Ena izmed prednosti našega sistema za obdelavo oblaka točk je v naravi pridobivanja podatkov z globinskega senzorja, ki na izhodu vrača matriko urejenih točk, kar omogoča bolj intuitivno, lažjo in hitrejšo implementacijo algoritma za ekstrakcijo ravnin oz. planarnih območij oblaka točk. Še vedno pa ostaja problem neenakomerne gostote točk in prisotnosti šuma, ki se pri meritvi globine veča z razdaljo.

Evolucijsko računanje pridobiva, na področju računalniške inteligence in strojnega učenja, na pomenu, saj izkazuje velik potencial za uporabo v stvarnem času in sistemih ter okoljih, ki se skozi čas spreminjajo [10]. Veliko metod na omenjem področju temelji na predpostavki, da imamo na voljo velik nabor zgodovinskih podatkov, ki jih uporabijo za generacijo modelov pri regresiji ali identifikaciji. Avtorji v [10] so izpostavili, da je ta predpostavka v realnih aplikacijah omajana, saj je v teh sistemih nabor prejšnjih podatkov lahko nezadosten, pri čemer imamo običajno opravka z dinamičnem okoljem ali sistemom. Poudarili so tudi neučinkovitost iterativnih algoritmov pri obdelavi naraščajočih podatkov, saj običajno zahtevajo

več prehodov čez enake dele podatkov. Opisano problematiko naslavlja področje samorazvijajočega se modeliranja [11–14], ki se je pokazalo za nov okvir, kjer obdelavo podatkov iz pretoka podatkov rešuje s pristopom samorazvoja (*angl. self-adaption*), učenjem z enim prehodom čez podatke in evolucijo, kot tudi kontrakcijo, parametrov modelov na zahtevo in spotoma.

V tem prispevku predstavljamo razširitev metode EPCC [15] v 3D-prostor za detekcijo ravnih površin v oblaku točk. Segmentacija poteka na podatkih z globinske kamere, kjer smo 3D-predstavitev okolja pridobili s triangulacijo razdalj. Modificirana rekurzivna metoda EPCC ekstrahira daljice s prototipi premic, kjer je glavni poudarek na rekurzivni oceni parametrov le-teh, s čimer pohitrimo algoritem. Algoritem je dodatno pohitren z iskanjem daljic v 2D-projekcijah oblaka točk z eliminacijo ene izmed koordinat. Daljice uporabimo za ekstrakcijo planarnih območij oz. prototipov ravnin. Ker je oblak točk urejen, smo omenjeni metodi implementirali v algoritem tako, da sproti odkriva daljice in ekstrahira prototipe ravnin. To pomeni, da lahko niz podatkov obdelamo v enem koraku. Prednost uporabljenega algoritma je tudi možnost upoštevanja urejenosti podatkov oblaka točk, kar pohitri in olajša razvoj metode za segmentacijo.

Preostanek prispevka ima naslednjo strukturo. Poglavje 2 podaja kratek pregled sorodnih del. Poglavje 3 opisuje ozadje planarne segmentacije z metodo EPCC. V poglavju 5 so prikazani rezultati eksperimentov na bazi podatkov, ki so prikazani v poglavju 4. V poglavju 6 so predstavljeni zaključki ob diskusiji rezultatov.

2 Sorodna dela

Prispevek [9] opisuje metodologije, ki se uporabljajo za segmentacijo oblaka točk. V našem prispevku smo se osredotočili na literaturo, ki je povezana z našim delom in izpostavili tri pogosto uporabljene metode za detekcijo ravnin v oblaku točk, in sicer pristope, ki temeljijo na metodi RANSAC (*angl. RANdom SAmple Consensus*), metodi rastočega območja (*angl. region growing*) in Houghovi transformaciji.

2.1 Pristopi z metodo RANSAC

Metodo RANSAC sta predstavila Fischler in Bolles v [16] za gradnjo modela oz. oceno parametrov le-tega iz eksperimentalnih podatkov.

Xu in drugi so v [8] predstavili planarno segmentacijo, kjer so z naučenimi SVM-modeli ter SVM-predikcijami klasificirali točke oblaka v različne ravninske površine in rezultate primerjali z navadno in NDT-RANSAC metodo.

Poz in Ywata sta v [17] predstavila adaptivni pristop k segmentaciji streh zgradb, ki vključuje tako predprocesiranje oblaka točk za ločevanje točk pripadajočim zgradbam, kot tudi segmentacijo ravnin z metodo RANSAC in postopek agregacije nadsegmentiranih območij.

Nguyen in drugi so v [18] predstavili segmentacijo delno urejenih MLS-oblakov točk, kjer najprej segmentirajo profile skeniranja z laserskim merilnikom na podlagi smernih vektorjev, nato te profile grupirajo in detektirajo ravnine glede na planarne vrednosti različnih sosedskih profilov skeniranja.

2.2 Pristopi s Houghovo transformacijo

Houghova transformacija, ki jo je prvič predstavil Hough v [19], je poleg metode RANSAC standardna metoda za detekcijo parametrov modelov ravnin ali krivulj v 2D- ali 3D-prostoru.

Vera in drugi so v [20] predstavili detekcijo ravnin v globinskih slikah za delovanje v realnem času, ki uporablja implicitno 'quadtree' drevesno strukturo za identifikacijo rojev približno koplanarnih točk v 2,5-D prostoru.

Khanh in drugi so v [21] predstavili izboljšavo segmentacije tal v oblaku točk za mobilne robote, ki temelji na metodi RHT (*angl. Randomized Hough Transform*) v kombinaciji z omejitvami razdalj ter kotov in jo primerjali z metodo RANSAC.

Tian in drugi so v [7] predstavila novo metodo za segmentacijo planarnih lastnosti v neurejenih oblakih točk na podlagi ekstrahiranih daljic z uporabo 2D Houghove transformacije in drevesne strukture (*angl. octree*).

2.3 Pristopi z metodo rastočega območja

Prednost metode rastočega območja je, da izkorišča lastnosti urejenih oblakov točk in sosedske strukture točk.

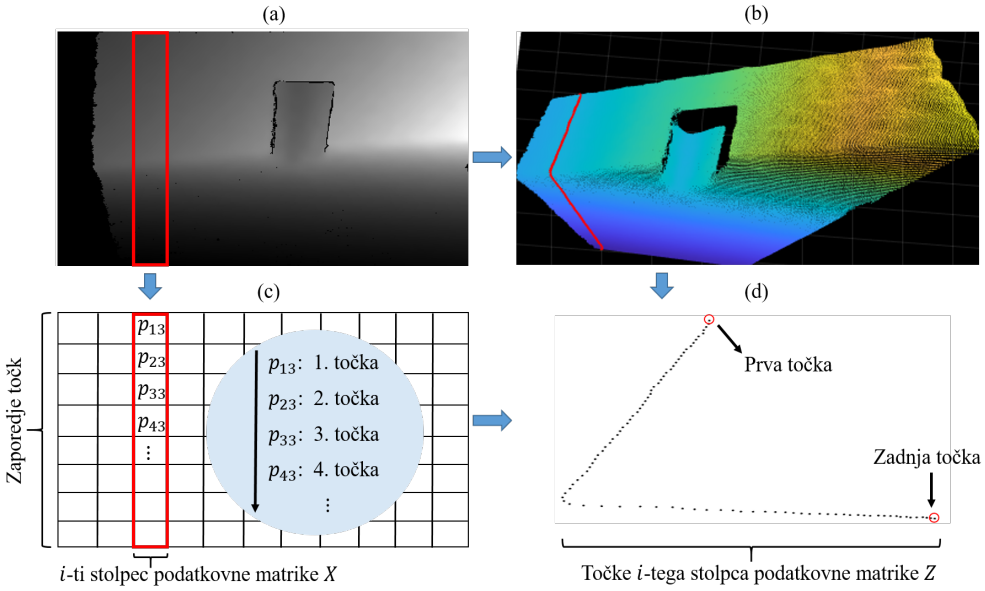
Dong in drugi so v [22] predstavili novo hibridno metodo rastočega območja, kjer so problem segmentacije oblaka točk predstavili kot robustno globalno energijsko optimizacijo. Konvergenco algoritma oz. energijske funkcije so zagotovili s pristopom simuliranega ohlajanja.

Wu in drugi so v [23] predstavili planarno segmentacijo TLS-oblakov točk z metodo MSTVM za boljšo določitev točke, ki predstavlja seme algoritma. Predstavili so novo lastnost — t. i. indikator planarne jakosti (*angl. plane strength indicator*) — za bolj intuitivno določitev točke semena.

Huang in drugi so v [24] predstavili algoritem EVBS (*angl. encoding voxel-based segmentation*), ki temelji na metodi rastočega območja s preiskovanjem strukture kvadrov (*angl. voxels*) in upoštevanju omejitev zveznosti ter gladkosti.

3 Detekcija daljic

Slika 2 prikazuje način pojavljanja podatkov v oblaku točk oz. globinski sliki. Algoritem za detekcijo daljic izkorišča vzorec zaporednosti točk in naravo rekurzivne obdelave podatkov metode EPCC. Struktura podatkov in narava zajemanja točk prostora znotraj (stožčastega) vidnega kota kamere omogoča ekstrakcijo daljic v 2D-projekcijah oblaka točk oz. podatkovnih matrikah X. Prostor iskanja daljic je razdeljen na tri dele (osrednje območje in dva zunanja) glede na horizontalni vidni kot α globinske kamere, ki znaša približno $90°$. Pri delitvi prostora upoštevamo koordinatni sistem kamere, kjer os z kaže iz kamere, x-os je v horizontalni smeri in y-os je v vertikalni smeri. 3D-točke matrike Z, ki se nahajajo v osrednjem območju oz. za α na intervalu $-30° - 30°$, projiciramo na yz-ravnino. Točke, ki se nahajajo v zunanjema



Slika 2: Globinska slika (a) in pripadajoč oblak točk (b). Pod (c) je prikazana matrika globinske slike, kjer so točke po posameznih stolpcih urejene po zaporedju od prve do zadnje vrstice. Prikaz točk označenega stolpca 3D-matrike oz. oblaka točk (d).

območjema oz. za kote, ki so po absolutni vrednosti večji od $\pm 30°$, pa projiciramo na xy-ravnino.

V vsaki projekciji poteka rojenje točk z rekurzivno oceno statističnih lastnosti 2D-podatkov in parametrov modelov premic, ki so določeni direktno iz kovariančne matrike podatkov $\boldsymbol{\Sigma}_j$. j-ti model premice opisuje normalni oz. lastni vektor $\boldsymbol{p}_j$, ki pripada najmanjši lastni vrednosti matrike $\boldsymbol{\Sigma}_j$

$$\boldsymbol{\lambda}_j = \boldsymbol{p}_j^T \boldsymbol{\Sigma}_j \boldsymbol{p}_j \tag{1}$$

Lastni vektor j-tega roja je določen iz $\boldsymbol{\Sigma}_j$ kot

$$\boldsymbol{p}_j^T = \left[-\frac{\theta_1}{\sqrt{\theta_1^2+1}} \quad \frac{1}{\sqrt{\theta_1^2+1}} \right] \tag{2}$$

kjer je

$$\theta_1 = \frac{\sigma_{12}}{\sigma_{11}} \tag{3}$$

in kjer so $\sigma_{io}, i, o \in \{1, 2\}$ elementi matrike $\boldsymbol{\Sigma}_j$. Psevdokoda detekcije daljic je predstavljena v Algoritmu 1. Na začetku algoritma je potrebno nastaviti parametre k_{min}, σ_{init} in n_{buf}. k_{min} pomeni minimalno število točk, ki jih potrebujemo za zanesljivo gradnjo novega prototipa roja, σ_{init} pa začetna varianca razdalje roja. S parametrom n_{buf} nastavimo največje število točk, ki jih še obravnavamo v postopku odkrivanja rojev. Algoritem roji podatke po stolpcih podatkovne matrike Z tako, da inicializira nov roj in ga postavi na aktivni seznam ter ga širi glede na določen kriterij. Širjenje roja je ustavljeno, ko podatek ne izpolnjuje kriterija in je odstranjen z aktivnega seznama. Tako širimo vedno tiste roje, ki so bili nazadnje odkriti. Kriterij za dodajanje v aktivni oz. j-ti roj je odvisen od ortogonalne razdalje do j-tega modela premice

$$d_j(k) = |(\boldsymbol{z}(k) - \boldsymbol{\mu}_j)^T \cdot \boldsymbol{p}_j| \tag{4}$$

Algoritem 1 Odkrivanje daljic z metodo EPCC

1: Definicija parametrov k_{min}, k_{max}, n_{buf}=3k_{min}, σ_{init}, n (število podatkov matrike $\boldsymbol{Z}$).
2:for $k = 1 : n$
3: switch $state$
4: case 0
5: $\boldsymbol{z}(k)$ shranjujemo v medpomnilnik buf.
6: if length(buf)>= k_{min}
7: Poiščemo kandidata za nov roj v buf in ocenimo $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma_j}$, $\boldsymbol{p_j}$ in d_j naslednje točke v buf ter postavimo $\sigma_j = \sigma_{init}$.
8: if $d_j < k_{max}\sqrt{\sigma_j}$
9: Ustvarimo roj in posodobimo $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma_j}$, $\boldsymbol{p_j}$ in σ_j. Postavimo $state = 1$ in počistimo medpomnilnik buf.
10: end
11: end
12: if length(buf)>= n_{buf}
13: Odstranimo najstarejši podatek iz buf.
14: end
15: case 1
16: Izračunamo $d_j(k)$ vzorca $\boldsymbol{z}(k)$.
17: if $d_j(k) <= k_{max}\sqrt{\sigma_j}$
18: Dodamo $\boldsymbol{z}(k)$ v j-ti roj in posodobimo $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma_j}$, $\boldsymbol{p_j}$ in σ_j.
19: else
20: $\boldsymbol{z}(k)$ damo v buf in postavimo $state = 0$
21: end
22: end
23:end

kjer je $\boldsymbol{z}(k)$ točka novega vzorca in $\boldsymbol{\mu}_j$ povprečje j-tega roja. V zajetih podatkih je prisoten šum, ki se veča s kvadratom razdalje. Da omogočimo robustno rojenje podat-

kov, je potrebno upoštevati prisotnost šuma. V ta namen vpeljemo v kriterij rojenja (5) definicijo variance ortogonalnih razdalj vseh točk v j-tem roju σ_j

$$d_j(k) < k_{max}\sqrt{\sigma_j} \tag{5}$$

kjer je k_{max} pozitivna konstanta, ki določa občutljivost kriterija rojenja v primeru normalne porazdelitve šuma.

Vsaki širitvi roja z novim vzorcem, sledi posodobitev normalnega vektorja in statističnih lastnosti po rekurzivnih enačbah (6)

$$\begin{aligned}
\sigma_j(k_j) &= \sigma_j(k_j-1)\frac{k_j-2}{k_j-1} + \frac{1}{k_j}d_j(k)^2 \\
\boldsymbol{\mu}_j(k_j) &= \frac{k_j-1}{k_j}\boldsymbol{\mu}_j(k_j-1) + \frac{1}{k_j}\boldsymbol{z}(k_j) \\
\boldsymbol{\Sigma}_j(k_j) &= \frac{k_j-2}{k_j-1}\boldsymbol{\Sigma}_j(k_j-1) \\
&+ \frac{1}{k_j}(\boldsymbol{z}(k_j)-\boldsymbol{\mu}_j(k_j-1))(\boldsymbol{z}(k_j)-\boldsymbol{\mu}_j(k_j-1))^T
\end{aligned} \tag{6}$$

kjer je k_j trenutni indeks podatkov in $\boldsymbol{z}(k_j)$ je trenutni podatek j-tega roja.

3.1 Detekcija ravnih površin

Izhod algortima 1 je urejena matrika rojev točk M, ki predstavljajo daljice. Urejenost matrike zopet izkoristimo z uporabo metode EPCC, tokrat v 3D-prostoru. V tem primeru poteka rojenje podatkov s prototipi ravnin (7), ki jih pridobimo direktno iz matrike $\boldsymbol{\Sigma}_j$ za primer 3D-podatkov

$$\boldsymbol{p}_j^T = \begin{bmatrix} -\frac{\theta_1}{\sqrt{\theta_1^2+\theta_2^2+1}} & -\frac{\theta_2}{\sqrt{\theta_1^2+\theta_2^2+1}} & \frac{1}{\sqrt{\theta_1^2+\theta_2^2+1}} \end{bmatrix} \tag{7}$$

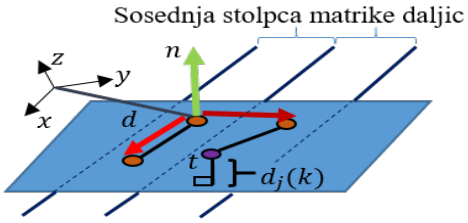
kjer sta

$$\theta_1 = \frac{\sigma_{13}\sigma_{22}-\sigma_{23}\sigma_{12}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2} \quad \theta_2 = \frac{\sigma_{13}\sigma_{12}-\sigma_{23}\sigma_{11}}{\sigma_{12}^2-\sigma_{11}\sigma_{22}} \tag{8}$$

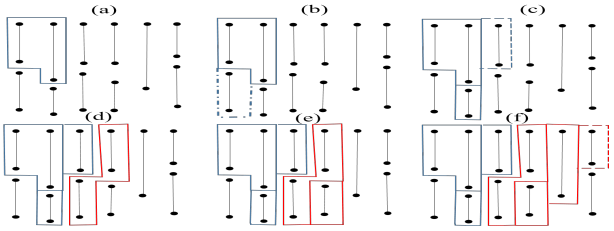
Algoritem za detekcijo ravnin roji podatke s primerjanjem daljic med sosednjmi stolpci matrike M. Postopek se začne z inicializacijo roja ravne površine s primerjanjem začetnih in končnih točk daljic med sosednjima stolpcema (slika 3). Pri vsaki inicializaciji pripišemo razvrščenim daljicam oznako pripadnosti j-temu roju (na sliki 4 so roji označeni z različno barvo). S pomočjo oznak lahko širimo in sledimo rojem skozi stolpce matrike M. Po vsaki inicializaciji sledi postopek širjenje roja, ki je pogojeno s kriterijem (5) oz. se preverja ustreznost ortogonalne razdalje začetne točke d_{j1} in končne točke d_{j2} opazovane daljice do j-tega prototipa ravnine. Da bi posamezno daljico razvrstili v pravi roj, je potrebno preveriti pripadnost le-te do vseh rojev. Daljice, ki jih ni možno razvrstiti, gredo v naslednjem koraku v postopek incializacije novega roja.

Zaradi podvrženosti podatkov šumu, se napaka meritev prenese tudi na odkrite daljice. Ta problem lahko omilimo z modeliranjem šuma, ki so ga predstavili Sung in drugi v [25]. Tako v kriterij rojenja (5) vpeljemo model šuma

$$\sigma_z(z) = k_1 z^2 + k_2 z + k_3 \tag{9}$$



Slika 3: Inicializacija roja ravne površine, glede na ortogonalno razdaljo $d_j(k)$ do le-te.



Slika 4: Tvorimo roj ravne površine (a). Kandidat za širjenje (b), ki ne ustreza kriterijem razvrstitve. Črtkan kandidat, ki se preverja v naslednjem koraku (c). Postopek se ponovi za vsak novi roj.

kjer je $k_1 = 0.003949$, $k_2 = 0.0007278$ in $k_3 = 0.001063$ ter dobimo

$$d_j(k) < k_{max}\sqrt{\sigma_j} + \sigma_z \tag{10}$$

V oblaku točk je prisotnih veliko lukenj, ki lahko predstavljajo ločnice med površinami. Zveznost ravnih površin ohranimo z uvedbo kriterija (11), ki upošteva najmanjšo evklidsko razdaljo d_{evk}, začetne ali končne točke daljice, do opazovanega roja ravne površine,

$$d_{evk} <= d_1 \tag{11}$$

kjer je d_1 konstanta in določa zgornjo mejo d_{evk}. V Algoritmu 2 je predstavljena psevdokoda detekcije ravnin.

4 Eksperimenti

Eksperimenti so izvedeni na podatkovni bazi, ki smo jo pridobili z globinskim senzorjem Intel RealSense D435i in predstavlja štiri notranje prizore (slika 5). Posamezni prizor vsebuje 129600 točk objektov z ravnimi površinami. Modificirano metodo EPCC smo primerjali z metodo rastočega območja in RANSAC, ki sta del knjižnice za obdelavo oblakov točk Point Cloud Library [26]. V tabeli 1 so opisani nastavljeni parametri algoritmov, ki so določeni s preizkušanjem za posamezen notranji prizor.

5 Rezultati

Na sliki 6 so prikazani rezultati planarne segmentacije po posameznih prizorih podatkovne baze. V primeru tretjega in četrtega prizora smo za vse metode upoštevali uspešno detekcijo tistih ravnih površin, ki so po zajemu z globinske kamere ostale dovolj nepopačene. Površini, ki sta pri tem izvzeti sta označeni na sliki 5. Omenimo naj,

Tabela 1: Opis vhodnih parametrov

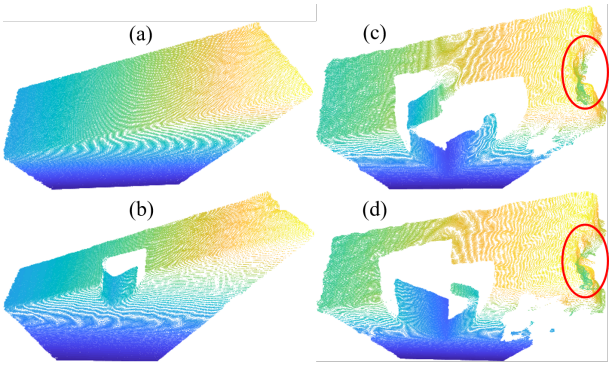
Param.	Vrednost				Opis	Metoda
c_{th}	1.0				Prag ukrivljenosti	Rastoče območje
Θ_{th}	2°	3°	2°	2°	Prag kota med dvema normalama	
num_k	30				Število upoštevanih sosednjih točk pri izračunu normale izbrane točke	
min_c	50				Minimalna velikost roja	
max_c	10^6				Maksimalna velikost roja	
k_n	50				Upoštevanje k-najbližjih sosedov v postopku segmentacije	
d_{th}	0.5 cm	1 cm	1.5 cm	1.5 cm	Prag oddaljenosti do modela ravnine	RANSAC
it_{max}	100				Največje število iteracij	
σ_{initP}	$10^{-6}m^2$	$10^{-6}m^2$	$5 \cdot 10^{-6}m^2$	$5 \cdot 10^{-6}m^2$	Začetna varianca razdalj pri detekciji ravnin	EPCC
σ_{init}	$2.5 \cdot 10^{-7}m^2$				Glej poglavje 3	
k_{min}	9					
k_{max}	3					
d_1	2 cm	1 cm	3 cm	4 cm		
Prizor	(a)	(b)	(c)	(d)		

Algoritem 2 Odkrivanje ravnin z metodo EPCC

1: Definicija parametrov σ_{initP}, d_1, k_{max}, c (št. stolpcev matrike M) in matrika M.
2: for $k = 1 : c$
3: k-ti stolpec matrike M shranimo v st_k, $(k-1)$-ti pa v st_{k-1}.
4: for $i = 1 : length(st_k)$
5: i-to daljico shranimo v seg_i
6: for $h = 1 : length(st_{k-1})$
7: h-to daljico shranimo v seg_h
8: if (seg_h pripada j-temu roju)
9: if $(d_{j1},\ d_{j2}) <= k_{max}\sqrt{\sigma_j} + \sigma_z$ & & $d_{evk} <= d_1$
10: seg_i dodamo v j-ti roj in posodobimo $\boldsymbol{\mu_j}$, $\boldsymbol{\Sigma_j}$, $\boldsymbol{p_j}$, σ_j ter prekinemo izvajanje trenutne zanke.
11: end
12: end
13: if (Klasifikacija neuspešna)
14: if $d_j(k) <= k_{max}\sqrt{\sigma_{initP}} + \sigma_z$ & & $d_{evk} <= d_1$
15: Tvorimo nov prototip in ocenimo: $\boldsymbol{\mu_j}$, $\boldsymbol{\Sigma_j}$, $\boldsymbol{p_j}$, σ_j.
16: end
17: end
18: end
19: end
20: end

da rdeče točke, v oblaku točk, predstavljajo nerazvrščene točke v primeru segmentacije z metodo rastočega območja.

V tabeli 2 so prikazani rezultati segmentacij v smislu deleža pravilno detektiranih ravnih površin $\frac{D}{N}$, kjer je D število pravilno detektiranih površin in N število vseh ravnih površin v določenem prizoru. Ocenjujemo tudi delež nadsegmentiranih $\frac{N_n}{N}$ in podsegmentiranih $\frac{N_p}{N}$



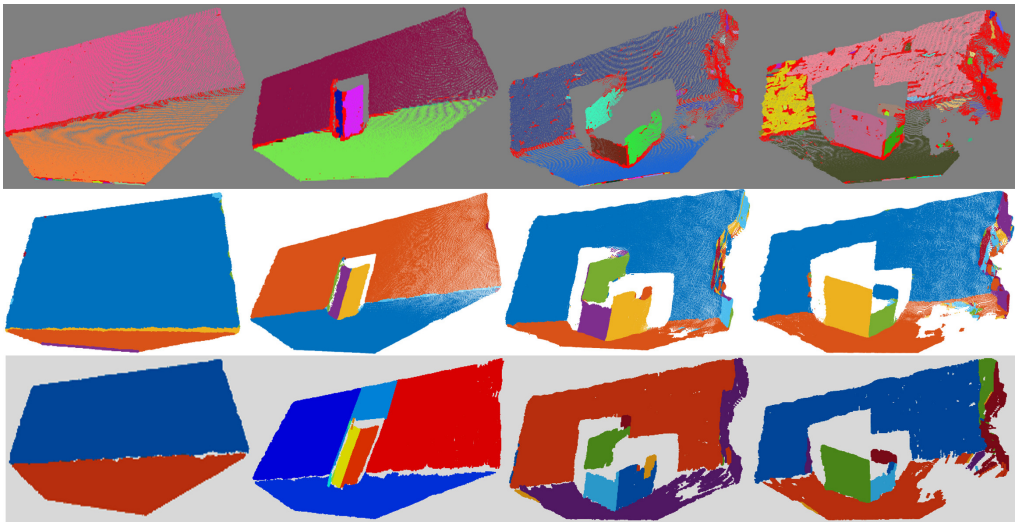
Slika 5: Podatkovna baza: prvi (a), drugi (b), tretji (c) in četrti prizor (d)

površin, kjer je N_n število nadsegmentiranih in N_p število podsegmentiranih površin. Glede na rezultate vidimo, da lahko metodo EPCC primerjamo z ustaljenimi metodami za planarno segmentacijo oblakov točk na podani podatkovni bazi. Za razliko od metode RANSAC, pa metoda EPCC, poleg rastočega območja, izloči neplanarne točke v postopku ekstrakcije ravnih površin. Slednje, poleg izbire vhodnih parametrov, pripomore tudi k manjši nadsegmentaciji površin, kot je razvidno iz tabele 2.

Pomankljivost uporabljene metode RANSAC je, da ne ohranja zveznost površin, kar lahko vodi do prekrivanja določenih območij, v kolikor niso bila v prejšnjih korakih detekirana kompletna območja. Prednost osnovne metode RANSAC pa je, da ima manjše število vhodnih parametrov, ki jih moramo empirično določiti.

6 Zaključek

Predstavili smo rekurziven pristop k segmentaciji oblaka točk, ki temelji na metodi EPCC. Razvit algoritem lahko predstavlja sestavni del avtonomnega mobilnega sistema



Slika 6: Rezultati planarne segmentacije prizorov z metodo rastočega območja (zgoraj), RANSAC (sredina) in EPCC (spodaj).

Tabela 2: Rezultati segmentacije notranjih prizorov

Metoda	D ($\frac{D}{N}$)	$\frac{N_n}{N}$	$\frac{N_p}{N}$	N
Prvi Prizor				
Rastoče območje	2 (100%)	0.5	0.0	2
RANSAC	2 (100%)	1	0.0	
EPCC	2 (100%)	0.0	0.0	
Drugi Prizor				
Rastoče območje	4 (100%)	0.0	0.5	4
RANSAC	4 (100%)	0.5	0.0	
EPCC	4 (100%)	0.25	0.0	
Tretji Prizor				
Rastoče območje	5 (50%)	0.2	0.3	10
RANSAC	8 (80%)	0.5	0.0	
EPCC	8 (80%)	0.2	0.0	
Četrti Prizor				
Rastoče območje	5 (62.5%)	0.38	0.5	8
RANSAC	8 (100%)	0.38	0.0	
EPCC	6 (75%)	0.25	0.13	

za potrebe vizualne navigacije ali gradnje zemljevida. Predlagana metoda izkorišča urejenost podatkov, kar omogoča hitrejšo in lažjo implementacijo ter sprotno obdelavo podatkov. Metoda EPCC je torej dobra izbira za odkrivanje linearnih prototipov v urejenih podatkih, saj daje dobre rezultate, kljub navzočnosti šuma. Algoritem tudi omogoča delovanje z drugimi senzorskimi sistemi, ki imajo podoben princip zajemanja podatkov (npr. laserski merilniki razdalje). Rezultati uporabljenih metod so pokazali, da EPCC lahko primerjamo z že ustaljenimi pristopi za obdelavo oblakov točk v smislu natančnosti. V prihodnosti želimo algoritem uporabiti za avtomatsko gradnjo zemljevida, ki bo podlaga za lokalizacijo in navigacijo avtonomnega mobilnega sistema v notranjih prostorih.

Zahvala

Raziskovalni program št. P2-0219 je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

[1] X. Zhang, W. Wang, X. Qi, Z. Liao, and R. Wei, "Point-plane slam using supposed planes for indoor environments," *Sensors*, vol. 19, no. 17, p. 3795, 2019.

[2] G. Klančar, L. Teslić, and I. Škrjanc, "Mobile-robot pose estimation and environment mapping using an extended kalman filter," *International Journal of Systems Science*, vol. 45, no. 12, pp. 2603–2618, 2014.

[3] A. Zdešar, I. Škrjanc, and G. Klančar, "Visual trajectory-tracking model-based control for mobile robots," *International journal of advanced robotic systems*, vol. 10, no. 323, pp. 1–12, 2013.

[4] A. Zdešar, I. Škrjanc, and G. Klančar, "Homography estimation from circular motion for use in visual control," *Robotics and autonomous systems*, vol. 62, pp. 1486–1496, Oct. 2014.

[5] A. Mousavian, C. Eppner, , and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2901–2910, 2019.

[6] S. Woo, J. Shin, Y.-H. Lee, Y.-H. Lee, H. Kang, H. Choi, and H. Moon, "Stair-mapping with point-cloud data and stair-modeling for quadruped robot," *2019 16th International Conference on Ubiquitous Robots (UR)*, pp. 81–86, 2019.

[7] P. Tian, X. Hua, K. Yu, and W. Tao, "Robust segmentation of building planar features from unorganized point cloud," *IEEE Access*, vol. 8, pp. 30873–30884, 2020.

[8] D. Xu, F. Li, and H. Wei, "3d point cloud plane segmentation method based on ransac and support vector machine," *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 943–948, 2019.

[9] E. Grilli, F. Menna, and F. Remondino, "A review of point clouds segmentation and classification algorithms," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 339–344, 2017.

[10] I. Škrjanc, J. A. I. Martinez, A. Sanchis, D. Leite, E. Lughofer, and F. Gomide, "Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification : a survey," *Information sciences*, vol. 490, pp. 344–368, July 2019.

[11] I. Škrjanc, "Cluster-volume-based merging approach for incrementally evolving fuzzy gaussian clustering - egauss+," *IEEE transactions on fuzzy systems*, pp. 1–11, Apr. 2019.

[12] I. Škrjanc, B. Sašo, E. Lughofer, and D. Dovžan, "Inner matrix norms in evolving cauchy possibilistic clustering for classification and regression from data streams," *Information sciences*, vol. 478, pp. 540–563, Apr. 2019.

[13] I. Škrjanc and B. Sašo, "Incremental fuzzy c-regression clustering from streaming data for local-model-network identification," *IEEE transactions on fuzzy systems*, vol. 28, pp. 758–767, Apr. 2020.

[14] D. Dovžan and I. Škrjanc, "Fuzzy space partitioning based on hyperplanes defined by eigenvectors for takagi-sugeno fuzzy model identification," *IEEE transactions on industrial electronics*, vol. 67, pp. 5144–5153, June 2020.

[15] G. Klančar and I. Škrjanc, "Evolving principal component clustering with a low run-time complexity for lrf data mapping," *Applied Soft Computing*, vol. 35, pp. 349–358, 2015.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Readings in Computer Vision*, pp. 726–740, 1987.

[17] H. P. D. Poz and M. S. Y. Ywata, "Adaptive random sample consensus approach for segmentation of building roof in airborne laser scanning point cloud," *International Journal of Remote Sensing*, vol. 41, no. 6, pp. 2047–2061, 2020.

[18] H. L. Nguyen, D. Belton, and P. Helmholz, "Planar surface detection for sparse and heterogeneous mobile laser scanning point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 151, pp. 141–161, 2019.

[19] P. Hough, "Method and means for recognizing complex patterns," *in: US Patent*, 1962.

[20] E. Vera, D. Lucio, L. A. Fernandes, and L. Velho, "Hough transform for real-time plane detection in depth images," *Pattern Recognition Letters*, vol. 103, pp. 8–15, 2018.

[21] H. D. Khanh, V. S. Nguyen, A. Do, and N. T. Dzung, "An effective randomized hough transform method to extract ground plane from kinect point cloud," *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1053–1058, 2019.

[22] Z. Dong, B. Yang, P. Hu, and S. Scherer, "An efficient global energy optimization approach for robust 3d plane segmentation of point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 137, pp. 112–133, 2018.

[23] H. Wu, X. Zhang, W. Shi, S. Song, A. Cardenas-Tristan, and K. Li, "An accurate and robust region-growing algorithm for plane segmentation of tls point clouds using a multiscale tensor voting method," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 10, pp. 4160–4168, 2019.

[24] M. Huang, P. Wei, , and X. Liu, "An efficient encoding voxel-based segmentation (evbs) algorithm based on fast adjacent voxel search for point cloud plane segmentation," *Remote Sensing*, vol. 11, no. 23, 2019.

[25] M. Sung, H. Chae, D. Noh, N. H, and D. Hong, "Analysis and noise modeling of the intel realsense d435 for mobile robots," *International Conference on Ubiquitous Robots (UR)*, pp. 707–711, 2019.

[26] "Point cloud library." http://www.pointclouds.org/. Accessed: 2020-2-12.

Poligon z miniaturnimi avtomatsko vodenimi vozički za učenje in razvoj avtonomnih mobilnih sistemov

Andrej Zdešar, Matevž Bošnak, Gregor Klančar

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: {andrej.zdesar, matevz.bosnak, gregor.klancar}@fe.uni-lj.si

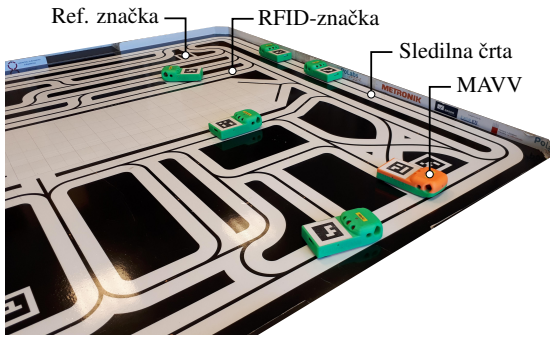
System of Small-Scale Automated Guided Vehicles for Education and Research of Autonomous Mobile Systems

The paper presents a cyber-physical model of a scaled-down industrial warehouse with automated guided vehicles (AGVs). The platform consists of a network of lines with RFID-tags that mark drivable roads and crossroads, along which the AGVs can travel. Above the platform it is a vision system that enables global tracking of all the AGVs in real time. AGVs are made from 3D printed, mechanical and electronic parts. The kinematic model of the AGVs drive train is a front-wheel driven tricycle. Since not all of the essential sensors are available in appropriate small-scale version, some virtual sensors are introduced. Various modular subsystems are integrated together using the Robot Operating System (ROS). The complete system enables study, development and evaluation of algorithms for autonomous driving. Although it is mainly used for educational purposes in the graduate course Autonomous Mobile Systems, it is also suitable for research activities.

1 Uvod

Na Univerzi v Ljubljani, Fakulteti za elektrotehniko se v drugem letniku magistrskega študija izvaja predmet Avtonomni mobilni sistemi (AMS). Predmet je sestavljen iz treh ur predavanj in dveh ur laboratorijskih vaj tedensko. Namen laboratorijskih vaj je, da študenti samostojno implementirajo metode za avtonomne mobilne sisteme, ki so jih spoznali tekom predavanj, na praktičnem primeru. Laboratorijske vaje pri predmetu AMS so v preteklosti potekale v skupinah dveh ali treh študentov na različne načine in na različni opremi. Eno študijsko leto je vsaka skupina delala na svojem projektu (na različnih mobilnih sistemih). Spet drugo leto so vsi študenti delali na istem projektu (na istem mobilnem sistemu), a vsaka skupina le na določenem podsistemu — na koncu pa so podsistemi, ki so jih razvile posamezne skupine morali delovati skupaj [1].

V zadnjih letih smo, zaradi naraščanja števila študentov pri predmetu (od študijskega leta 2018/19 poteka predmet vzporedno tudi v angleškem jeziku za študente na izmenjavi) in za dosego lažje primerljivosti študentskega dela, vaje preoblikovali tako, da imajo vse skupine študentov



Slika 1: Fizični model industrijske hale z MAVV-ji.

enako nalogo, ki jo vsaka skupina izvaja na svojem mobilnem sistemu. Naloga je zasnovana tako, da so postopoma obdelane vse glavne teme predavanj: modeliranje, vodenje, planiranje poti in lokalizacija. Po desetih terminih laboratorijskih vaj sledi ob zaključku semestra zagovor vaj, ki poteka kot demonstracija delovanja avtonomnega mobilnega sistema.

Za izvedbo laboratorijskih vaj smo morali zagotoviti več enakih kolesnih mobilnih sistemov, ki omogočajo izvedbo vsebine obravnavane na predavanjih. Znan sistem, ki se tudi uporablja za namen učenja in razvoja algoritmov za avtonomno vožnjo, je *Duckietown* [2], a je sistem veliko prevelik za vpeljavo v obstoječo učilnico. V Laboratoriju za avtomatiko in kibernetiko (LAK) smo zato razvili fizični model industrijske hale in miniaturne avtomatsko vodene vozičke (MAVV) ter pripadajoče pomožne sisteme (slika 1). Sistem predstavlja nadgradnjo našega predhodnega sistema, kjer smo imeli miniaturni poligon mesta in mobilne robote z diferencialnim pogonom [3]. V nadaljevanju je predstavljena zgradba celotnega sistema in možnosti, ki jih sistem ponuja — ne le za pedagoške namene, temveč tudi za raziskave in razvoj, npr. sledenje trajektoriji [4] ali vožnja več-agentov v formaciji [5] itd.

Pri izdelavi MAVV-jev nismo zahtevali natančne preslikave dejanske situacije iz industrijskega okolja, saj vseh sistemov (npr. laserskega merilnika razdalj, pogonskega mehanizma) ni moč primerno pomanjšati. S fizičnim modelom tako posnemamo le tiste lastnosti avtomatsko vodenih vozičkov (AVV), ki so potrebne za učenje in razvoj algoritmov za avtonomno vožnjo. Za namen izvedbe pedagoškega procesa smo želeli, da sistem omogoča razvoj

algoritmov za avtonomno vožnjo, tako da je možna vsaj izvedba odometrije, vodenja mobilnega sistema po poti, načrtovanje poti med poljubnimi cilji in lokalizacija mobilnega sistema v znanem zemljevidu okolja. Za namen raziskav in razvoja smo zaželi tudi, da sistem omogoča še razvoj algoritmov za več-agentno vodenje, načrtovanje in optimizacijo poti več agentov, iskanje alternativnih poti in obvozov v primeru zastojev in/ali ovir in podobno.

Zaradi boljše povezljivost med različnimi sistemi in modularnosti smo se odločili, da bomo uporabili okolje ROS (angl. *Robot Operating System*).

2 Poligon

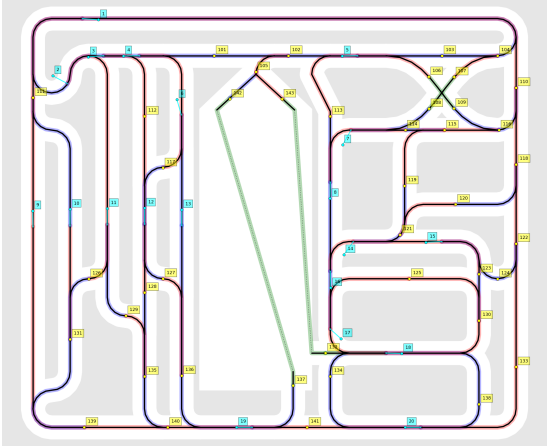
V industrijskih okoljih z AVV-ji se za označevanje poti mnogokrat uporabljajo magnetni trakovi, ki jih je mogoče zanesljivo zaznavati, in RFID-značke, ki delujejo na podlagi radio frekvenčne tehnologije, za označevanje imen poti in križišč. Naš poligon je sestavljen iz omrežja usmerjenih poti, ki se v križiščih razdružijo v več poti ali pa se združijo v eno pot (slika 2). Pri miniaturizaciji mobilnega sistema smo v našem primeru uporabili optični IR-senzor, ki omogoča detekcijo črte pod mobilnim sistemom, in RFID-bralnik, ki omogoča branje RFID-značk, ki se nahajajo na znanih položajih ob poteh. Glede na izbiro roba črte, ki mu sledimo, lahko v križišču izberemo želeno pot. RFID-značke se ne nahajajo ob vseh poteh in tudi ne pred vsemi križišči (na sliki 2 so položaji RFID-značk označeni s sinje modrimi oznakami).

V osrednjem delu poligona imamo tudi t. i. prosto področje, ki je brez sledilnih črt. V tem področju mora navigacija mobilnega sistema biti izvedena s pomočjo dodatnih senzorjev, in ne na podlagi optičnega senzorja za sledenje črti.

2.1 Predstavitev zemljevida poligona z grafom

Zemljevid poligona lahko zapišemo v obliki grafa. Vsaka pot ima določeno smer, ki jo v grafu predstavimo kot usmerjeno povezavo. Vozlišča grafa predstavljajo križišča, kjer se stikajo poti. Graf v programskem jeziku Python lahko zapišemo s pomočjo gnezdenega slovarja (podatkovni tip `dict`).

Za namen planiranja poti in vodenja smo graf predstavili v še nekoliko prirejeni obliki. Mesta na poteh, kjer se nahajajo RFID-značke, smo obravnavali kot postaje, kjer se mobilni sistem lahko ustavi oz. se odloči o svoji naslednji akciji — običajno izbiramo le med sledenjem levemu ali desnemu robu črte. Ker se na vseh poteh ne nahajajo RFID-značke, le te opremimo s t. i. virtualnimi značkami (njihov položaj lahko določimo le na podlagi odometrije/lokalizacije), ki jih tudi obravnavamo kot postaje. Vozlišča v grafu torej predstavljajo vse značke (postaje), poti med značkami pa so usmerjene povezave v grafu (na sliki 2 so povezave med značkami označene z modrimi, rdečimi in zelenimi črtami). V tej predstavitvi se določeni deli poti med postajami torej prekrivajo. Za to predstavitev grafa smo pripravili program, ki avtomatsko pretvori prej omenjeno predstavitev zemljevida z grafom v to obliko. Pri tem smo upoštevali tudi, da je



Slika 2: Zemljevid z označenimi potmi in položaji RFID-značk (sinje modre oznake) ter virtualnih značk (rumene oznake).

položaj RFID-senzorja izmaknjen glede na točko na mobilnem sistemu, ki sledi črti (glej poglavje 4).

3 Globalni sistem za merjenje lege s strojnim vidom

Nad poligonom se nahaja kamera, ki omogoča sledenje vseh objektov, ki se gibljejo po ravnini poligona. Na vsakem robotu se nahaja posebna značka z unikatnim vzorcem, ki ga sistem s kamero lahko robustno zazna in določi njegovo lego na sliki. Sledi opis postopka merjenja lege vseh značk oz. mobilnih sistemov glede na poligon, pri čemer je lega kamere poljubna.

Predpostavimo, da je koordinatni sistem kamere glede na svetovni koordinatni sistem podan z rotacijsko matriko $\boldsymbol{R} = [\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3]$ in translacijskim vektorjem $\boldsymbol{t}$. Glede na perspektivični model kamere je preslikava homogene točke $\boldsymbol{p}_W^T = [x_W\ y_W\ 1]$ na poljubni ravnini v prostoru in homogeno točko $\boldsymbol{p}_P^T = [x_P\ y_P\ 1]$ v koordinatnem sistemu slike:

$$\boldsymbol{p}_P \propto \boldsymbol{S} \begin{bmatrix} \boldsymbol{r}_1 & \boldsymbol{r}_2 & \boldsymbol{t} \end{bmatrix} \boldsymbol{p}_W = \boldsymbol{H} \boldsymbol{p}_W \,, \qquad (1)$$

kjer matrika $\boldsymbol{S}$ vsebuje notranje (intrinzične) parametre kamere. Matriko $\boldsymbol{S}$ lahko določimo s kalibracijo, npr. z uporabo postopka s šahovnico [6]. Preslikava $\boldsymbol{H}$ v (1) je znana kot homografija — predstavlja preslikavo med dvema ravninama.

Homografijo $\boldsymbol{H}$ lahko ocenimo na podlagi vsaj štirih korespondenčnih parov točk v ravnini tal in slikovni ravnini. Iz tega razloga se na poligonu nahajajo štiri referenčne značke (slika 1), katerih položaj glede na svetovni koordinatni sistem je znan. Za namen merjenja lege je vsak MAVV opremljen s posebno unikatno značko. Med vožnjo MAVV-ja se značka premika po ravnini, ki je vzporedna s tlemi, zato se položaja MAVV-ja ne da oceniti neposredno na podlagi homografije $\boldsymbol{H}$. Ker poznamo višino h, na katero je nameščena značka na MAVV-ju, lahko določimo homografijo $\boldsymbol{H}_h(h)$ za to vzporedno ravnino na podlagi homografije $\boldsymbol{H}$. Položaj značke $\boldsymbol{p}_P$ na sliki, ki se nahaja v vzporedni ravnini glede na ravnino

tal na višini h, lahko preslikamo na ravnino tal v točko $\boldsymbol{p}_W$, ki je pravokotna projekcija značke na ravnino tal:

$$\boldsymbol{p}_W \propto \boldsymbol{H}_h^{-1}(h)\boldsymbol{p}_P\,.$$

Da lahko ocenimo homografijo $\boldsymbol{H}_h(h)$ moramo poznati intrinzične parametre kamere $\boldsymbol{S}$:

$$\boldsymbol{H}_h(h) = \boldsymbol{S}\begin{bmatrix}\boldsymbol{q}_1 & \boldsymbol{q}_2 & (\boldsymbol{q}_1 \times \boldsymbol{q}_2)\frac{h}{n} + \boldsymbol{q}_3\end{bmatrix}\,, \qquad (2)$$

kjer parameter normiranja n določimo kot:

$$n = (\|\boldsymbol{q}_1\| + \|\boldsymbol{q}_2\|)/2\,. \qquad (3)$$

Vektorji $\boldsymbol{q}_1$ do $\boldsymbol{q}_3$ v (2) in (3) so:

$$\boldsymbol{S}^{-1}\boldsymbol{H} = \begin{bmatrix}\boldsymbol{q}_1 & \boldsymbol{q}_2 & \boldsymbol{q}_3\end{bmatrix}\,.$$

Vse dokler se mobilni robot giblje le po ravnih tleh, predstavljena metoda omogoča zelo točno sledenje, saj so omejitve sistema implicitno upoštevane v algoritmu za merjenje lege. Ker značka ni rotacijsko invariantna, lahko določimo tudi njeno orientacijo in tako celotno lego MAVV-ja v ravnini tal. Predstavljena metoda merjenja lege MAVV-jev deluje za poljubno postavitev kamere, dokler so vse značke v vidnem polju kamere, a natančnost sistema je lahko spremenljiva. Če ne pride do okluzij referenčnih značk na podlagi, deluje metoda tudi med gibanjem kamere, sicer pa se referenčna homografija $\boldsymbol{H}$ izračuna le kadar so vidne vse štiri referenčne značke.

Predstavljeni pristop je tudi računsko učinkovit, tako da ne zahteva izredno zmogljive strojne opreme. Algoritem za merjenje lege več MAVV-jev smo implementirali na vgradnem sistemu *Raspberry Pi 3 B+* s kamero *Raspberry Pi Camera Board v2*. Za sledenje značk smo uporabili knjižnico ArUco [7, 8], ki omogoča robustno in hitro sledenje značk z unikatnimi vzorci, ki je neodvisno od osvetljenosti prostora. S sistemom smo dosegli sledenje več značk in ocenjevanje lege MAVV-jev glede na svetovni koordinatni sistem s frekvenco 20 Hz. Podatki o legi MAVV-jev so na voljo v ROS-u na kanalu `tf`, ki vsebuje podatke o preslikavah med koordinatnimi sistemi. Sistem s strojnim vidom za merjenje lege omogoča hitro in enostavno postavitev, saj je potrebno nastaviti le položaje referenčnih značk in višine posameznih značk na MAVV-jih ter določiti notranje parametre kamere (npr. s standardnim postopkom za kalibracijo kamere).

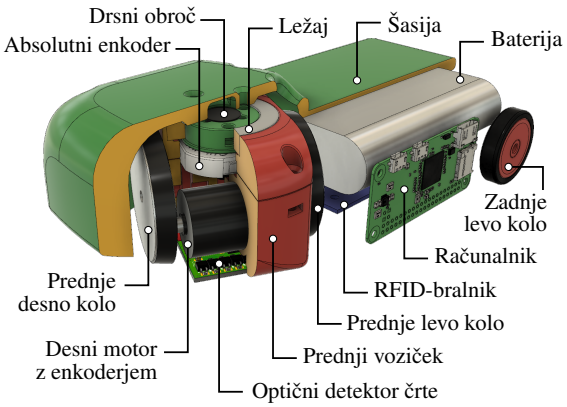
4 Miniaturni avtomatsko vodeni vozički

Želeli smo izdelati MAVV s kinematičnim modelom trikolesnika, ki ima gnano prednje kolo.

4.1 Zgradba

Prednji pogonski del smo izdelali z vozičkom z diferencialnim pogonom, ki je na šasijo pritrjen pasivno preko ležaja. Takšna oblika mehanskega mehanizma je bila najbolj enostavna za izvedbo, pri tem pa smo še vedno dosegli želeni kinematični model. Zgradba MAVV-ja z glavnimi sestavnimi deli je prikazana na sliki 3.

MAVV smo opremili z več senzorji, ki omogočajo implementacijo algoritmov za avtonomno vožnjo. Motorja na prednjem vozičku, ki ženeta obe kolesi, sta opremljena z inkrementalnim enkoderjem. Poleg tega imamo



Slika 3: Zgradba MAVV-ja v prerezu.

še absolutni enkoder, ki omogoča merjenje kota prednjega vozička glede na šasijo. Na dnu prednjega vozička se nahaja namensko tiskano vezje z mikrokrmilnikom, ki skrbi za nizkonivojsko obdelavo signalov v realnem času in regulacijo hitrosti vrtenja motorjev. Na tiskanem vezju se nahaja tudi 7 segmentni optični detektor črte.

Zadnji kolesi sta pasivni in v trenutni izvedbi nista opremljeni z enkoderji. Na spodnjem delu šasije se nahaja še RFID-bralnik.

Za obdelavo informacij s senzorjev, komunikacijo z zunanjimi sistemi in izvedbo regulacijskih algoritmov za avtonomno vožnjo se znotraj šasije nahaja računalnik *Raspberry Pi Zero W*. V šasiji se nahaja še baterija.

4.2 Dodatni virtualni senzorji

Zaradi majhnosti MAVV-ja, smo omejeni z naborom senzorjev, ki jih lahko uporabimo, saj določenih senzorjev, ki se običajno uporabljajo na AVV-jih, ni na voljo v tako majhni izvedbi. Takšen primer je laserski merilnik razdalj (LMR), ki je pri sodelujočih AVV-jih običajno obvezen kot varnostni element. AVV-ji so lahko opremljeni celo z več LMR-ji, ki omogočajo pokrivanje čim večjega vidnega kota. LMR-ji se pogosto uporabljajo za namen lokalizacije in gradnje zemljevida okolja [9].

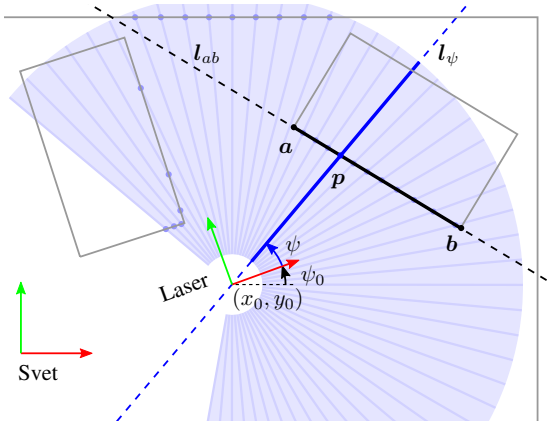
Čeprav LMR-ja ne moremo primerno pomanjšati, da bi ga vgradili na MAVV, pa lahko simuliramo meritve tega senzorja. Globalni sistem s strojnim vidom nam omogoča merjenje leg MAVV-jev, poznamo pa tudi njihove oblike in obliko poligona. Oblike vseh objektov lahko opišemo z daljicami, nato pa uporabimo algoritem za detekcijo presečišč laserskih žarkov z daljicami ovir.

Algoritem se sprehodi čez vse daljice vseh objektov, ki predstavljajo statične ali dinamične ovire, pri čemer za vsako daljico naredimo naslednje (slika 4):

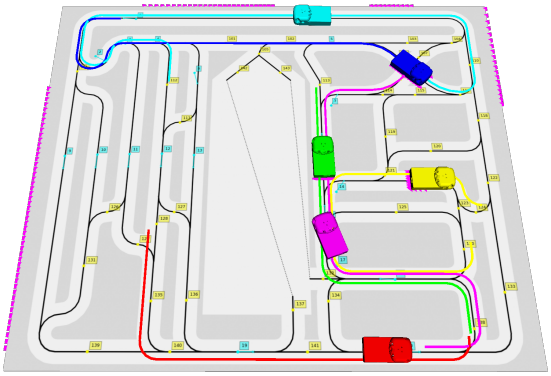
1. Če sta $\boldsymbol{a}$ in $\boldsymbol{b}$ robni točki daljice zapisani v homogenih koordinatah, potem je premica skozi ti dve točki $\boldsymbol{l}_{ab} = \boldsymbol{a} \times \boldsymbol{b}$ in enotski vektor daljice $\boldsymbol{e}_{ab} = \frac{\boldsymbol{b}-\boldsymbol{a}}{\|\boldsymbol{b}-\boldsymbol{a}\|}$.
2. Za vsak žarek LMR-ja

   $$\boldsymbol{l}_\psi = [-\sin(\psi)\ \cos(\psi)\ x_0 \sin(\psi) - y_0 \cos(\psi)]$$

   naredimo naslednje:



Slika 4: Modeliranje laserskega merilnika razdalj.



Slika 5: Vizualizacija meritev LMR-ja, leg MAVV-jev in načrtovanih poti v orodju RViz.

(a) Poiščemo presečišče

$$\boldsymbol{p} = [\kappa x \ \kappa y \ \kappa]^T = \boldsymbol{l}_\psi \times \boldsymbol{l}_{ab}$$

žarka s premico daljice in izračunamo faktor $q = \boldsymbol{e}_{ab}^T(\frac{\boldsymbol{p}}{\kappa} - \boldsymbol{a})$.

(b) Če velja $0 \leq q \leq 1$, potem presečišče $\boldsymbol{p}$ leži na daljici. V tem primeru izračunamo razdaljo r od objekta do izhodišča laserskega merilnika (v smeri vektorja žarka):

$$r = (\frac{\boldsymbol{p}^T}{\kappa} - \boldsymbol{p}_0^T) \begin{bmatrix} \cos\psi & \sin\psi & 0 \end{bmatrix}^T .$$

Če velja $0 \leq r \leq r_{max}$ in če je r tudi manjši od trenutne najkrajše razdalje r_ψ, posodobimo najkrajšo razdaljo: $r_\psi = r$.

Virtualni LMR lahko namestimo glede na katerikoli koordinatni sistem (npr. na prednji del MAVV-ja). Nastavimo lahko različne parametre senzorja, kot so doseg, natančnost, vidni kot, kotna ločljivost itd. Podatke meritev podamo v obliki standardnega ROS-sporočila tipa `sensor_msgs/LaserScan`. S stališča podatkov ne moremo ločiti med načinom uporabe realnega ali virtualnega senzorja. Virtualni senzor ne more zaznati objektov, katerih leg ne poznamo oz. jih ne merimo z globalnim sistemom za merjenje lege — poleg MAVV-jev lahko z značkami označimo tudi druge objekte na poligonu in tako omogočimo zaznavanje tudi teh.

Na sliki 5 je vizualizacija laserskih meritev (vijolične točke) v orodju RViz (del okolja ROS), kjer se laserski merilnik razdalj nahaja na vijoličnem MAVV-ju. Vidimo odboje laserskih žarkov na zelenem, modrem in rumenem MAVV-ju, ne pa tudi na sinje modrem MAVV-ju, saj je zasenčen z zelenim MAVV-jem; rdeči MAVV pa je izven vidnega kota senzorja. Zaradi omejenega dosega senzor ne zaznava objektov, ki so preveč oddaljeni.

5 Zaključek

V članku smo predstavili zgradbo miniaturnih avtomatsko vodenih vozičkov, poligona in podpornih sistemov, ki omogoča učenje in tudi razvoj algoritmov za avtonomno vožnjo. Čeprav fizikalne lastnosti dejanskih AVV-jev niso natančno posnemane, MAVV-ji omogočajo študijo vseh esencialnih problemov, ki se pojavljajo pri razvoju avtonomnih mobilnih sistemov. Sistem je modularen in ga je možno enostavno in hitro prilagoditi za želene namene. V prihodnosti imamo v načrtu opremiti MAVV-je z dodatnimi senzorji (kamera, senzorji bližine idr.). V načrtu pa je tudi nadgradnja sistema, ki bo omogočala oddaljen dostop do sistema in opravljanje laboratorijskih vaj oz. preizkusov na daljavo.

Zahvala

Raziskovalni program št. P2-0219 je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

[1] A. Zdešar, S. Blažič, and G. Klančar, "Engineering education in wheeled mobile robotics," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 12173 – 12178, 2017. 20th IFAC World Congress.

[2] L. Paull, J. Tani, H. Ahn, J. Alonso-Mora, L. Carlone, M. Cap, Y. F. Chen, C. Choi, J. Dusek, Y. Fang, *et al.*, "Duckietown: an open, inexpensive and flexible platform for autonomy education and research," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1497–1504, IEEE, 2017.

[3] M. Tavčar, U. Skalar, J. Perko, and A. Zdešar, "Implementation of model-based control systems for autonomous driving of miniature robotic vehicles," in *Proceedings of the 27th International Electrotechnical and Computer Science Conference ERK*, pp. 131–134, IEEE, 2018.

[4] G. Klančar and I. Škrjanc, "Tracking-error model-based predictive control for mobile robots in real time," *Robotics and autonomous systems*, vol. 55, no. 6, pp. 460–469, 2007.

[5] G. Klančar, D. Matko, and S. Blažič, "A control strategy for platoons of differential drive wheeled mobile robot," *Robotics and Autonomous Systems*, vol. 59, no. 2, pp. 57–64, 2011.

[6] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[7] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Generation of fiducial marker dictionaries using mixed integer linear programming," *Pattern Recognition*, vol. 51, pp. 481–491, 2016.

[8] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "Speeded up detection of squared fiducial markers," *Image and vision Computing*, vol. 76, pp. 38–47, 2018.

[9] L. Teslić, I. Škrjanc, and G. Klančar, "Using a lrf sensor in the kalman-filtering-based localization of a mobile robot," *ISA transactions*, vol. 49, no. 1, pp. 145–153, 2010.

Solution space of optimal heat pump schedules

David Kraljić[1], **Miha Troha**[1]

[1]*COMCOM d.o.o Idrija, Lapajnetova 29, Idrija, Slovenia*
E-pošta: david.kraljic@comcom.si

Abstract

We study the space of optimal schedules of a heat pump with thermal energy storage used in heating a residential building. We model the heating system as a Mixed Integer Linear Program with the objective to minimise the cost of heating. We generate a large number of realistic daily heat demands and calculate the optimal schedule for the heat pump. In addition to cost savings stemming from optimal running, we find that the space of optimal schedules is large in practice, even for the simplest model of the heating system we use, and that the optimal schedules are difficult to reproduce with statistical models. These findings strengthen the case for the use of mathematical optimisation in real-life applications.

1 Introduction

Mathematical optimisation is a generic framework for formulating and solving problems where some goal (or objective) needs to be optimised (maximised or minimised) given some constraints. It has been applied successfully across many domains such transport and logistics[1], power plant dispatch and electric grid balancing [2], economics [3], machine learning [4], biology [5] and many others.

With the advent of smart buildings and smart devices[1] which enable users to actively control them, the problem arises of how to do that optimally. Mathematical optimisation framework is well suited to tackle this problem. The goal of a smart device owner is to minimise utility costs (or equivalently maximise profits) subject to the physical characteristics[2] of a smart device.

Usually, in practice, smart devices are not actively controlled by mathematical optimisation algorithms, but instead they follow simple rules or some predefined daily schedules. The question appears whether modelling the devices and surrounding processes to produce their optimal schedules adds any value in comparison to schedules or logic fixed in advance. The aim of this contribution is to study this question in a simplified but realistic heating system for an actual residential building. We focus on the properties of the solution space for optimal running of the heating system given the realistic demand for heat for the building.

[1]Examples include air conditioning units, heat pumps, charging electric cars.

[2]Examples include size, losses, speed of charging or discharging etc.

In Section 2 we describe the devices and other data, in Section 3 we discuss the mathematical and statistical modelling used, in Section 4 we show the results and discuss them.

2 Set-up, Data and Benchmark

The optimisation problem that we address in this contribution aims to minimise costs of operating a heat pump by considering physical constraints of the heat pump and the requirement that enough heat must be generated to satisfy heat demand of the building. Data for costs, physical constraints of the heap pump and demand of the building were determined based on realistic historical data of a residential building[3].

2.1 Costs

The main cost of heating comes from the cost of electric energy which is split into two tariffs: lower cost tariff (MT) applicable between 10pm and 6am and a higher cost tariff (VT) applicable between 6am and 10pm. We assume that VT stands at 1.5 of MT, which is a realistic split given the current electricity prices.

2.2 Physical constraints

The heat pump is capable of providing roughly 160 kW of heat at a coefficient of performance[4] ~ 1.6. The heat storage is capable of storing up to 200 kWh of energy.

2.3 Demand

We use real heat demand data which is measured by a calorimeter at the exit of the heating system. A typical daily demand profile based on 480 days during three years of heating season (mid October to April) is depicted in Fig. 1.

2.4 Benchmark

The heat pump operates in the 'on/off' fashion, that is, either it runs at maximum power or is idle at low power. Schedules produced by the mathematical optimisation algorithms are to be benchmarked against actual historical

[3]Retirement home 'Dom Sv. Lenart', Lenart, Slovenia

[4]The coefficient of performance determines the heat generated by one unit of electrical power.

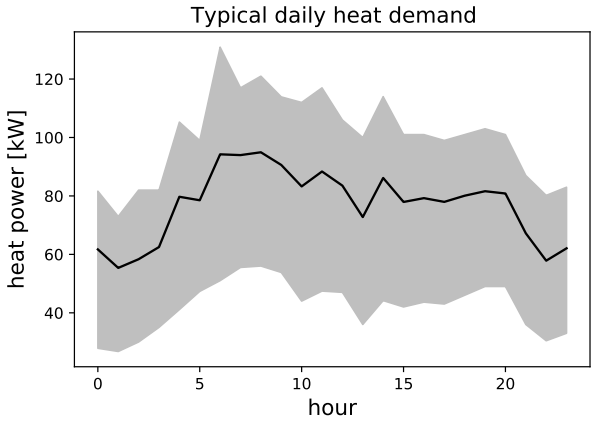


Figure 1: Typical daily heat demand. The shaded region around the mean line corresponds to area between the first and third quantile.

schedules of the heat pump. For an example of a historical schedule, as captured by a power meter located at the heat pump, see Fig. 2.

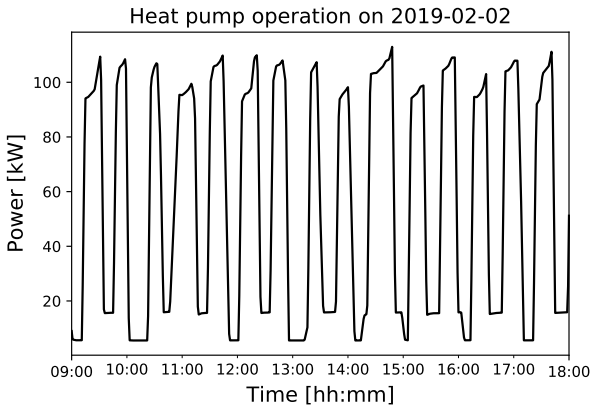


Figure 2: Typical operation of the heat pump.

3 Methods

3.1 Optimal schedule of the heat pump

We obtain the optimal daily schedule for the heat pump by building an optimisation model for the heating set-up with the objective to minimise the cost of operation. We follow the modelling of [6], [7].

Our aim is to evaluate the variability of optimal solutions given the heat demand. Therefore, we simplify a realistic description as far possible without losing the crucial elements of the set-up - that is the variation in schedules is only determined by the variations in heat demand and not in the variation of the modelling parameters. Thus, we neglect the variable, fixed, and cycling costs of the heat pump as well as heat losses in the storage. We also assume the coefficient of performance is constant throughout the day (even though it varies with outside temperature). We also model the heat pump as only being on at maximum power or off. Finally, we chose hourly granulation for all the variables and parameters in our model. This means that the solution space for the heat pump can be encoded as a sequence of zeros and ones of length 24, corresponding to heat pump either running or not in a specific hour. The size of the solution space is 2^{24}.

The optimisation problem we solve reads as follows:

$$\min_{P_t, \overline{P}_t, S_t} \quad \sum_{t \in T} P_t^{\text{HP}} \cdot \pi_t^{\text{elec}} \tag{1}$$

$$\text{s. t.} \quad S_{t+1} = S_t + \text{cop}_t \cdot P_t^{\text{HP}} - \text{hd}_t \; t \in \overline{T} \tag{2}$$

$$P_t^{\text{HP}} = p^{\max} \cdot \overline{P}_t^{\text{HP}} \; t \in T \tag{3}$$

$$s^{\min} \leq S_t \leq s^{\max} \; t \in T \tag{4}$$

$$P_t, S_t \in \mathbb{R} \; t \in T \tag{5}$$

$$\overline{P}_t \in \mathbb{B} \; t \in T \tag{6}$$

where P_t^{HP} is the heat pump power, π_t^{elec} is the electricity price, S_t is the heat energy stored, cop_t is the coefficient of performance, hd_t is the building heat demand, $p^{\max}$, $s^{\max}$, $s^{\min}$ are the power and energy limits, and $\overline{P}_t^{\text{HP}}$ is the binary decision variable denoting whether the heat pump is on or off. Time sets T and $\overline{T}$ are defined as $\{0, 1, ..., 23\}$ and $\{0, 1, ..., 22\}$.

The objective in Eq. 1 is the cash paid for the electricity consumed to heat the building, which we minimise[5]. Eq. 2 describes the energy balance for the heat storage. The heat pump is constrained to be either off or on at max power whcih is enforced by Eq. 3. The storage limits are set by Eq. 4.

We use the following parameters: $p^{\max} = 100\,\text{kW}$, $s^{\max} = 200\,\text{kWh}$, $s^{\min} = 0\,\text{kWh}$, $\text{cop}_t = 1.6$. For all runs we set the initially stored heat, $S_0 = 100\,\text{kWh}$.

3.2 Daily heat demand profiles

We are interested in running the optimisation problem from Sec. 3.1 a large number of times to explore as much of solution space as computationally possible. For a 24 hour granulation of the daily optimal schedule of an 'on/off' heat pump the total number of potential solutions is $2^{24} \approx 1.7 \cdot 10^7$. Since we only have the actual data for 480 days during the heating season, we need to generate synthetic but realistic heat demand profiles.

Heat demand for each hour of the day is a random variable with a mean and a variance in the simplest terms. Because the demand is bounded from below with 0, a simple normal distribution cannot be used to model it. We chose a lognormal distribution[6] which is bounded by 0 and is also only described by a mean and variance. We illustrate our modelling for a specific hour in Fig. 3.

Because the heat demand from hour to hour within a single day is not independent, we need to also model these hourly cross-correlations. The size of these is illustrated by the covariance matrix for the hourly demand in Fig. 4.

[5] We add to electricity prices a very small random increase with each hour of the day, thus ensuring that the optimal objective has a unique solution that prefers taking decisions earlier in the day.

[6] logarithm of a random variable is distributed normally

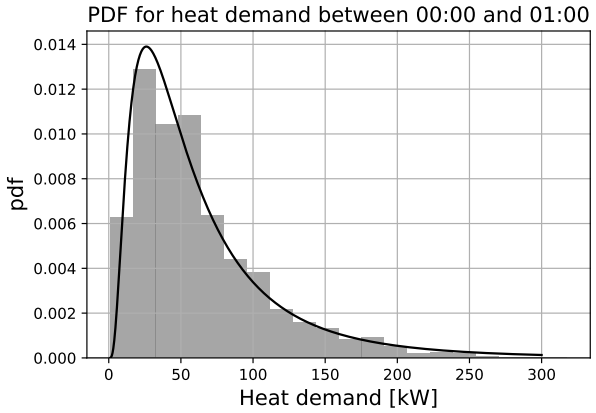


Figure 3: Shaded is the empirical pdf for the heat demand between hour 00:00 and 01:00. The line is our best fit lognormal model.

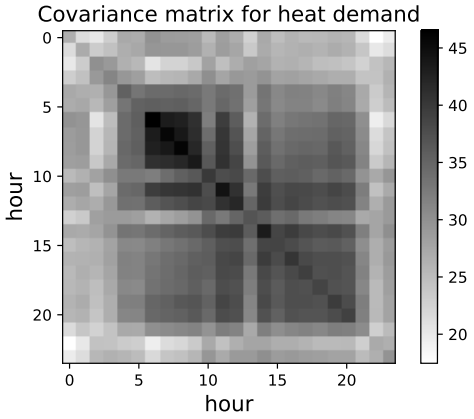


Figure 4: Empirical covariance matrix for the hourly heat demand.

Our final model for the daily demand is the lognormal distribution derived from empirical mean (Fig. 1) and the covariance matrix (Fig. 4) of the actual data.

4 Results & Discussion

We have generated 100 000 daily heat demand profile samples and solved the optimisation problem (Sec. 3.1) the same number of times, obtaining 100 000 optimal solutions.

4.1 Size of solution set

To determine if our solution set is localised in the space of all possible solutions, we compare it to the baseline case – the set where all 2^{24} solutions for the heat pump schedule are possible and equally probable. For such a uniformly distributed set a single solution appears on average $\lambda = n/2^{24}$ times, where $n = 100\,000$ is the number of generated solutions. Then it follows that the *number* of times a solution appears is Poisson distributed with mean λ.

In Fig. 5 we plot the fraction of solutions we found against the multiplicity of the solutions. That is, we plot the fraction of solutions that appear at most once, twice, three times etc., in our solution set. We also plot the expectations stemming from the uniform distribution of solutions, as explained above.

We see (Fig. 5) that the set of optimal schedules we obtain has solutions appearing more often and at higher multiplicities than expected if the baseline case would be true. Note, that the fraction of solutions that appear only once in the baseline case is practically all of them (0.9999823). This suggests that the solution space to our optimisation problem is localised and not completely random. This is expected, as the variability in the heat demand is not completely random, but constrained.

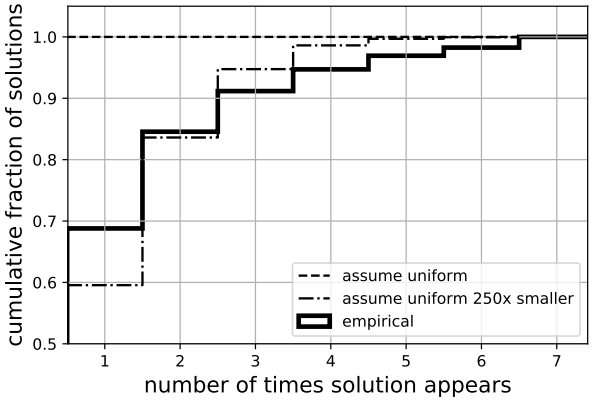


Figure 5: Empirical multiplicities of our solution set compared to expectations if solutions were uniformly distributed in the space of all possible solutions.

However, the fraction of solutions that appear once, twice etc., is still small. For example, there are no solutions that would appear more than 7 times in our set of 100 000, suggesting that different demand profiles lead to very different optimal solutions.

If we assume the solution set to our optimisation problem is uniformly distributed (but smaller) then by fitting the Poisson distribution to our empirical data in Fig. 5 we deduce that the solution space is roughly 250 times smaller than 2^{24} ($\approx 7 \cdot 10^4$). This means a single solution to our optimisation problem is expected to repeat roughly every $7 \cdot 10^4$ daily runs or 180 years.

4.2 Complexity of solution set

We measure how easy it is to describe the solutions to our heat pump optimisation problem by training machine learning models to predict optimal heat pump schedules given the heat demand and evaluating their performance.

The generated dataset of 100 000 heat demand profiles forms the input set with the corresponding optimal solutions being the output set. We split the dataset into train, validation, and test subsets. We selected tree-based (Decision tree, Random Forest) models and linear models (Logistic regression)[7]. All model parameters were fine-tuned on the validation set, with the final performance

[7]These models are selected as they can be expressed easily in code and thus can be deployed to control the heat pump

evaluated on the test set. We also evaluate a trivial model which produces random schedules.

We measure how well machine learning models predict the optimal solution by counting the average number of hours that differ between true and predicted schedules. In other words, we count the minimum number of errors that transform the true optimal solution to the one predicted by a machine learning model (known as the Hamming distance[8] in coding theory). We summarise the results in Table 1.

Table 1: Performance of machine learning models predicting optimal solutions.

Model	Best param.	Hamming error [hour]
Trivial model	-	10.1
Decision Tree	max_depth = 10	5.94
Random Forest	max_depth = 10 num_trees = 400	5.72
Log. Regression	-	6.36

Solving the optimisation problem in Sec. 3.1 performs a mapping from the daily heat demand profile to the optimal schedule. Mimicking this mapping with either linear or tree-based models does not work well since on average the predicted optimal solutions are still wrong on about 6 hours of the day.

4.3 Cost savings

Currently (i.e. in our benchmark data), the heat pump operates in short bursts of 15-30minutes (e.g. Fig. 2), and follows the real-time demand for heat almost exactly without utilising the large heat storage available. We evaluate the cost of this current strategy on the 480 days of heating seasons using the actual data. We run the simplified optimisation problem from Sec. 3.1 for the same days and plot the differences in Fig. 6

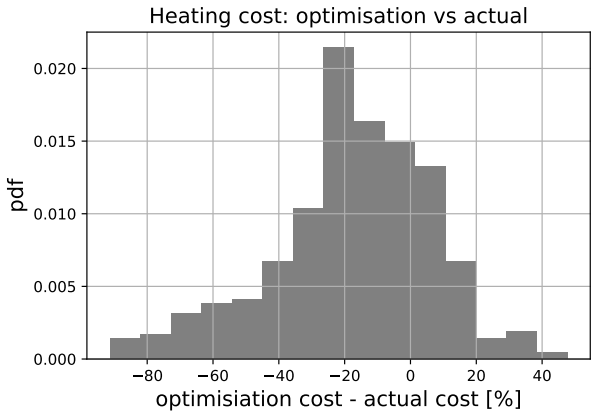


Figure 6: Percentage difference between the costs of schedules produced by our simple optimisation model and actual costs.

We note that even the simple modelling of the heating system we use manages to save on average 15% in heating costs. We expect that formulating the optimisation problem with higher granularity (sub-hourly), as well as using time-varying COP, would make the gains even bigger, as the problem would be less constrained.

5 Conclusion

We studied the set of optimal solutions for a simplified heating system consisting of a heat pump&storage satisfying realistic heat demand profiles. The set-up and data are based on an actual residential building. We find that the solution space for optimal heat pump schedules is smaller than the maximum solution space, but still large enough to justify using the optimisation framework in practice. We also show that the optimal schedules, once they are known, are not easy to predict using statistical models. Finally, we show that even simplistic modelling of the optimisation problem can lead to significant cost savings.

Acknowledgement

This contribution was done as part of the 'BETAi' project supported by public tender 'DEMO PILOTI II 2018', that is partly financially supported by European Union from European Regional Development Fund and Ministry of Economic Development and Technology.

We thank Martin Turk for data preparation and processing and Gasper Zadnik for implementing the optimisation problem in code.

References

[1] Pursula, Matti & Niittymäki, Jarkko. (2001). Mathematical Methods on Optimization in Transportation Systems. 10.1007/978-1-4757-3357-0.

[2] Lin, J., & Magnago, F. H. (2017). Electricity markets: Theories and applications. Somerset: John Wiley & Sons, Incorporated.

[3] Mikulás Luptácik, 2010. Mathematical Optimization and Economic Analysis, Springer Optimization and Its Applications, Springer, number 978-0-387-89552-9.

[4] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. 2011. Optimization for Machine Learning. The MIT Press.

[5] Reali Federico, Priami Corrado, Marchetti Luca. 2017. Optimization Algorithms for Computational Systems Biology, Frontiers in Applied Mathematics and Statistics. 10.3389/fams.2017.00006

[6] Renaldi, Renaldi & Kiprakis, Aristides Friedrich, Daniel. (2016). An optimisation framework for thermal energy storage integration in a residential heat pump heating system. Applied Energy. 10.1016/j.apenergy.2016.02.067.

[7] Schütz, Thomas & Harb, Hassan & Streblow, Rita Mueller, Dirk. (2015). Comparison of models for thermal energy storage units and heat pumps in mixed integer linear programming.

[8] R. W. Hamming, Error detecting and error correcting codes, The Bell System Technical Journal, 10.1002/j.1538-7305.1950.tb00463.x.

Močnostna elektrotehnika

Power Engineering

Electric field strength inside a residential house due to a high voltage overhead line proximity

Amar Zejnilović[1], Breda Cestnik[1], Boštjan Blažič[2]

[1] Milan Vidmar Electric Power Research Institute (EIMV), Hajdrihova 2, Ljubljana
[2] University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, Ljubljana
E-pošta: amar.zejnilovic@eimv.si

Abstract *- One of the primary sources of the electric field in the environment are high voltage transmission lines. The calculation of the electric field strength is important for the proper design of transmission lines, and for verification that the calculated values of this field are within the prescribed values.*

As it is well known, every building has a shielding effect on electric field distribution, i.e. the electric field strength inside the building is reduced. The aim of this paper is to determine the electric field strength inside a house that is caused by a high voltage transmission line proximity. First of all, the computation of the unperturbed electric field was done, using the EFC-400 software. Then the measurements of electric field strength inside the house were performed. Comparing these two values, a percentage value of the electric field strength inside the house according to the calculated unperturbed electric field is determined. Also, the obtained electric field strength inside the house is compared with the limit values.

Keywords—electric field measurement, electric field calculation, shielding factor, house

1 Introduction

Electrical energy has become one of the basic goods and probably one of the greatest achievements of the civilization. Direct and indirect impact of electric energy on the quality of our lives is huge. Modern lighting, electrical transportation, automatization, robotization, telecommunication etc. are all benefits of the usage of electrical energy. On the other hand, the usage of electrical energy has also some side products. One of them is generation of electromagnetic fields.

One of the primary sources of the electric field in the environment are high voltage transmission lines. The calculation of electric field is important for the proper design of transmission lines, and for verification that the calculated values of this field are within the prescribed values. This is the reason for establishing the organizations and commissions that are dealing exclusively with negative impacts of electric and magnetic fields. One such organization is the International Commission on Non-Ionizing Radiation Protection – ICNIRP, which published in 2010 new guidelines regarding the human exposure to the limit values of the field.

Also, some countries have legislations regarding electromagnetic field. In Slovenian Decree, electromagnetic radiation in natural and living environment has been adopted in 1996, which gives limit values for electric and magnetic fields in natural and living environment in a specific manner.

Before the construction of new high voltage (HV) transmission line, the simulation and calculation of electric field around that line has to be done and only if the calculated values of electric field strength are lower than recommended ones, the line could be constructed. Also, if a new object (for example house) is planned near an existing or planned HV transmission line, the electric field inside that object needs to be determined and compared with the limit values before the design of object starts. Using the appropriate software, for example EFC-400, the calculation of the unperturbed electric field is relatively easy. On the other hand, the calculation of electric field inside the object would be very complex and would require quite different type of software.

The problem of the determination of the electric field strength inside a house that is caused by high voltage transmission line proximity is analyzed in this paper. First of all, the computation of the unperturbed electric field was done, using the EFC-400 software. After calculating the unperturbed electric field around the transmission line, the measurements of electric field strength inside the house were performed. Comparing these two values, a percentage value of the electric field strength inside the house according to the calculated unperturbed electric field is determined.

2 Electric field generated by transmission line

Transmission lines are one of the most important parts of every Electro Power System (EPS). They transmit electrical energy on various voltage levels (high, very high and extremely high voltages). Transmission lines are part of the transmission system that connects power generation system with the distribution system and without them, the functioning of the EPS, which we are familiar with, would be impossible.

Transmission lines are also one of the primary sources of the electric field in the environment. The voltage of transmission line is practically the same, which means that electric field generated by this source is constant with regard to time. [1]

The most important reasons for calculating electric fields from transmission lines are:

- Determination and evaluation of compliance with prescribed or recommended limit values with regard of protection of humans.
- Impact on objects in the surrounding area (pipelines, telecommunication lines etc.)

2.1 Limit values for electric field strength by ICNIRP

The International Commission on Non-Ionizing Radiation Protection (ICNIRP) is an independent organization that provides scientific advice and guidance on the health and environmental effects of non-ionizing radiation (NIR) to protect people and the environment from detrimental NIR exposure. In 2010, this organization published new guidelines regarding the limit values of the field.[2]

Two different types of reference levels are mentioned in this guideline: reference level for professionally exposed persons, and the reference level for the general population.[2]

- Professionally exposed persons are adults, who are exposed to the known value of field no more than 8 hours per day. They are also introduced to the potential risks of working in this environment.
- General population are all other individuals who can be exposed to the field even 24 hours per day. In this population, there are also young and old people, and this is the reason why for this type of population, the reference level for the electric field is lower.

The limits values which are set up by ICNIRP are given in Table 1 as the effective values of electric field strength (*E*) for frequency 50Hz:

Table 1: Limit values of electric field strength for frequency 50 Hz, according to ICNIRP.

Category of population	***E*** **(kV/m)**
Professionally exposed persons	10
General population	5

2.2 Limit values for electric field strength by Slovenian standard

In Slovenia, the limit values are defined with regulatory act "Uredba o elektromagnetnem sevanju v naravnem in življenskem okolju" (Hereinafter "Uredba") and these limit values are stricter for new sources in the areas of higher degree of protection than the ones from ICNIRP. Limit values for electric field strength for frequency 50Hz, defined by Uredba are given in the table below: [3]

Table 2: Limit values for electric field strength for frequency 50Hz, according to *Uredba.*

	area I, for new and reconstructed transmission lines	area II, for new and reconstructed transmission lines, area I and II for existing transmission lines
E	0,5kV/m	10kV/m

2.3 Electric field strength factors

The electric field strength generated from transmission lines depends on the following factors: [4]

- Transmission line voltage: There is a linear relation between the voltage and electrical field strength, which means the electric field strength is changing linearly with the voltage change.
- Tower structures: Height of the conductors above the ground, radius of the conductors, space between the phase conductors, and between the sideway conductors from each other, are all have influence on the electric field strength.
- Distance from the conductors: Electric field strength decreases from the conductors with the inverse square of distance.
- Number of systems

The influence of each factor on the electric field strength can be seen from Figure 1. Three types of towers are presented on the figure: Donau 220kV, Donau 400kV and Plane 400kV. Comparing electric field produced from Donau 220kV and Donau 400kV, the voltage influence on the amplitude of the electric field strength (*E*) can be seen. For example, comparing the values of *E* just below the tower, the electric field strength is almost 2,5 times higher for the 400kV system than for the 220kV system for the same type of tower.

The distribution and amplitude of electric field strength depend also on the tower construction. The best way to confirm this is to compare two different types of towers of the same voltage level. That analysis could be done by taking a look at the Figure 1 and comparing towers Donau 400kV and Plane 400kV. Electric field strength just below the towers is almost the same, but the distribution of field is totally different and the main reason for this is in the position of conductors.

The dependence of electric field strength on distance is obvious from all three towers, and it can be concluded that *E* decreases rapidly with the distance.

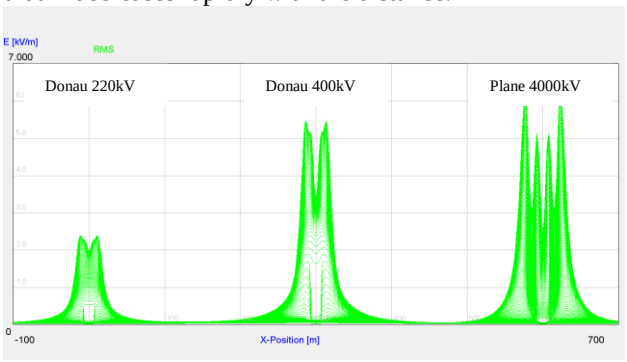


Figure 1: Electric field strength in dependence on distance for different types of towers and different voltages. The figure is obtained using the software EFC-400.

On the figure above, the distribution of unperturbed electric field is presented. In practice, the electric field is usually shield/reduced by trees, houses, fences etc.

All values for *E* that are shown and that will be shown below are values 1m above the ground [5] (unless otherwise stated).

3 Results

Although most of the electric field inside the house is reduced, it still has to be determined due to the fact that electric field strength could have relatively high values, especially on the upper floors of the houses located in the vicinity of the power transmission lines. As it was already said, the calculation of the electric field inside the house with the usage of software is very complex. The best way of determining the exact value of electric field inside the house is by measuring it. But in certain situations, the measurement of the electric field inside the house is not possible. The example is when the construction of the new overhead transmission line (OHL) is planned near the house; the determination of the electric field inside the house has to be done, but it cannot be measured since the OHL hasn't been constructed yet. In this situation, approximate value of electric field strength inside the house is determined as the combination of the unperturbed electric field (calculated with the software) and the shielding factor of the house. It is said approximate value since the value of the shielding factor depends on the material of the walls, material of the roof, objects inside the house etc. The shielding factor of the house is determined as the combination of the calculated unperturbed electric field and measured values of electric field strength as it will be shown further in the paper.

The house where the measurements were performed is located between the towers SM 34 and SM 35 (Figure 2). House is in the vicinity of the 400kV Beričevo-Divača transmission line. Lowest distance between the house and transmission line axis is 12,7m. Environmental conditions and transmission line parameters are given in the table below:

Table 3: Environmental conditions and transmission line parameters

Ambient temperature T	9,2 °C
Humidity	56,5°
Voltage at the time of measurements	404,1kV - 409,8kV
Height of the lowest conductor	17,1m
Types of towers	SM 34: Y2 and SM 35:Y7

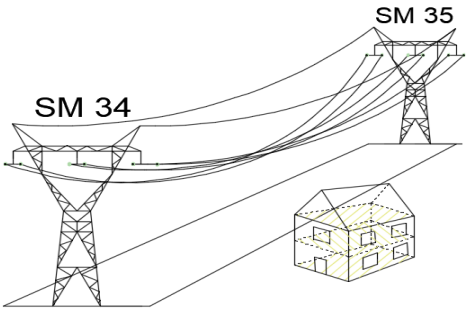


Figure 2: Sketch of the situation

Two types of measurements were done:

- Inside the house: the aim of this measurements is to determine the shielding factor of the house for the electric field.
- Measurements of unperturbed electric field: these measurements provide the information whether the model for calculation of unperturbed electric field is a precise representation of the practical case.

All the measurements were performed with the EFA-300 Field Analyser, manufactured by Narda STS.

3.1 Measurements of unperturbed electric field

Measurements of unperturbed electric field were done in order to compare these values with the ones that are obtained with software. If a model that is used in the software is done right, the calculated and measured values should have similar values. The measurements were done up to 20m on both sides of the transmission line axis, with the measurement step of 2m as it is shown on the figure below.

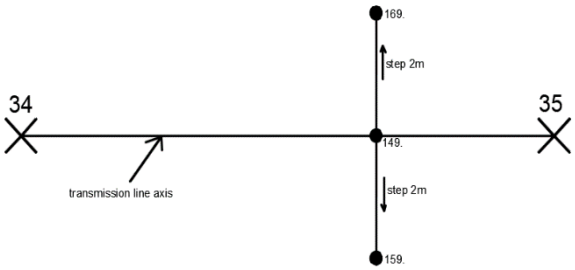


Figure 3: Measurement points of unperturbed electric field - top view

There are 21 measurement points at total (10 on one and another side of the axis and 1 point on the axis of the transmission line). The distribution of the electric field that is shown on Figure 4 is calculated with the software EFC-400, while the distribution of electric field on Figure 5 is measured with the EFA-300 field analyzer. Comparing these two figures it can be seen that values of electric field are same (or very similar) in each point of the space, in both mentioned cases. This means that the model that is used for the calculation is a precise representation of the practical case and it can be used in further calculations for determination of the electric field shielding factor of the house.

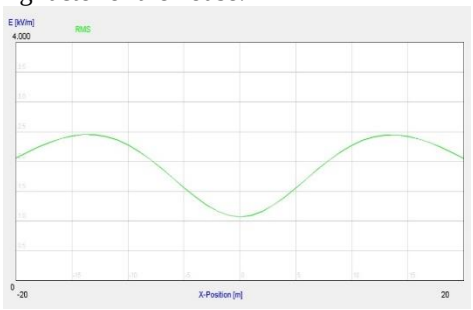


Figure 4: Electric field calculated with the software EFC-400

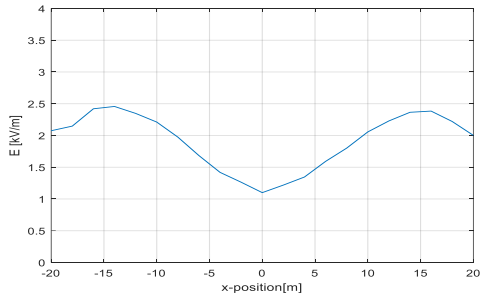


Figure 5: Electric field measured with the EFA-300 Field Analyzer

3.2 Electric field shielding factor of the house

The chosen house has 3 floors: ground floor, 1st floor and 2nd floor (attic). Generally, the measurements were done in three directions: 0,2m, 1m and 2m from the wall and multiple measurements were done on each direction as it will be shown further in the paper. Each floor will be analyzed separately.

On the Figure 6 top view of ground floor is shown. The measurements of electric field are numbered from 1. to 24. and all the measurements were done on 1m height, unless otherwise is stated in *Notes*. Despite the fact that all 24 points were measured, just some of them will be compared to the calculated values, and these points will be called *representative points* from now on. Some of the measurement points can't be compared with the calculated values due to the fact that there are near power sockets so the main source of electric field strength in that points is power socket and not transmission line or there are just behind fence or flowers and electric field strength is too small that EFA-300 Field Analyzer couldn't measure it with satisfactory accuracy.

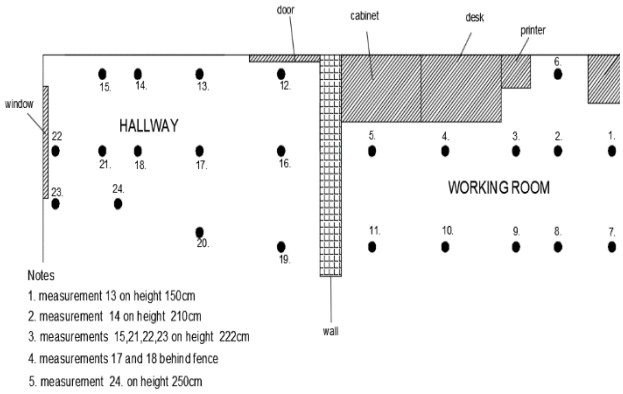


Figure 6: Top view of ground floor

As it was already said few times, the software EFC-400 could calculate only unperturbed electric field and doesn't take into account the shielding factor of the house. The distribution of unperturbed electric field on the house location and on 1m height is shown on Figure 7. The house is located between -12m and –25m on x-axis and between 160m and 170m on y-axis on the graph.

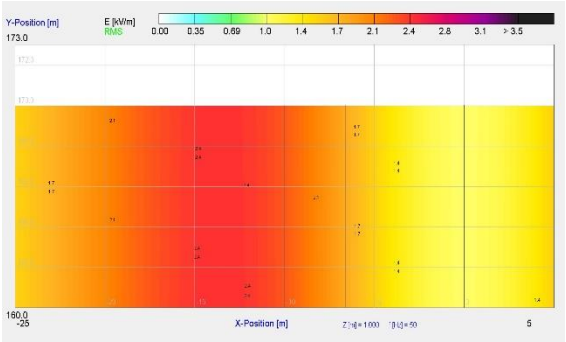


Figure 7: Distribution of unperturbed electric field on the house location on 1m height

In the Table 4 the measured and calculated unperturbed values[1] for representative points are shown, as well as the shielding factor in the same points. The shielding factor is calculated according to the following formula:

$$shielding\ factor\ (\%) = \left(1 - \frac{measured\ value}{calculated\ unpertubed\ value}\right) \cdot 100\% \quad (1)$$

Table 4: Shielding factor in representative points on the ground floor

Measurement point	Measured value (V/m)	Calculated unperturbed value (V/m)	Shielding factor (%)
1.	14,6 ± 0,6	2439	99,4
7.	10,7 ± 0,4	2406	99,5
12.	5,6 ± 0,2	2395	99,7
13.	45,2 ± 1,8	2385	98,1
			average: 99,175

The average shielding factor of the house on the ground floor is 99,175% which means that electric field inside the house, on ground floor, is just 0,825% of the unperturbed electric field.

Before analyzing the first floor, it is interesting the take a look at measured values from table above and to compare these values with the limit ones from Table 1 and Table 2. As it can be seen, the measured values are much lower compared to the limit ones.

The top view of the first floor is shown on Figure 8. The measurement points are numbered from 25. to 39. and the measurements were done on 3,42m height, unless otherwise is stated in the Notes. The measurements 35-39 were done on the mezzanine (space between first and second floor) and that's the reason why are these measurements were done on 4,9m height, while the measurements 33 and 34 were done on the stairs.

The distribution of unperturbed electric field on the house location and for the first floor is shown on the Figure 9.

[1] unperturbed value is calculated without the house

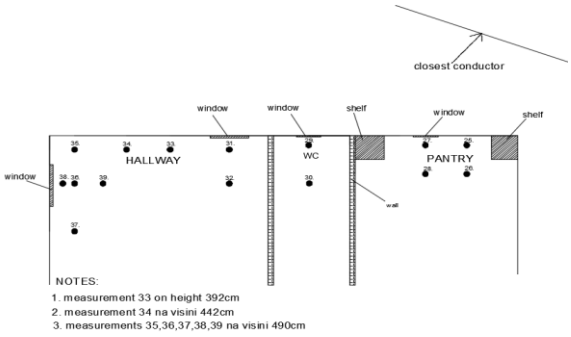


Figure 8: Top view of the first floor

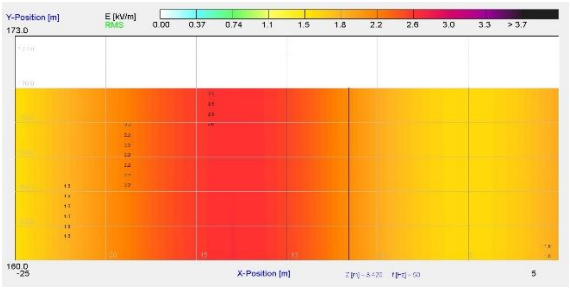


Figure 9: Distribution of the unperturbed electric field on the house location on 3.42m height

In the Table 5 the measured and calculated unperturbed values for representative points on the first and mezzanine floor are shown, as well as the shielding factor in the same points.

Table 5: Shielding factor in the representative points on the first

Measurement point	Measured value (V/m)	Calculated unperturbed value (V/m)	Shielding factor (%)
25.	20,5 ± 0,8	2611	99,2
27.	41,2 ± 1,6	2596	98,4
31.	16,2 ± 0,6	2519	99,3
32.	8,3 ± 0,3	2468	99,6
33.	18,4 ± 0,7	2524	99,3
34.	10,0 ± 0,4	2521	99,6
			average: 99,23

The average shielding factor of the house on the first + mezzanine floor is 99,23% which means that electric field inside the house, on the first floor with mezzanine floor, is 0,77% of the unperturbed electric field. The value of shielding factor on first floor is similar as the shielding factor on the ground floor.

Before analyzing the measurement points on attic (second floor) it is interesting to compare shielding factors in points from the previous table, especially between the points 25. and 27. As it can be seen from the Figure 8, the measurement point 25. is closer to the transmission line compared to the measurement point 27. and therefore it is logical to assume that the electric field strength in this point is higher than in the point 27. But from Table 5, it can be seen that the measured electric field in point 27. is two times higher than electric field in point 25. This is mostly due to the fact that measurement point 27. is just behind the window and point 25. is behind the wall and comparing the shielding factor in these two points it can be presumed that a window has lower shielding factor compared to a wall.

The top view of the second floor is shown on the Figure 10. All the measurements were done on the height of 6,2m.

The distribution of unperturbed electric field on the house location and for the second floor is shown on the Figure 11. Comparing the values of electric field strength from this figure with the Figure 9 and Figure 7, it is obvious that the electric field has higher values at attic compared to the ground and first floor and that was expected since the distance between the attic and transmission line is lower in comparison with the distance between transmission line and first (ground) floor.

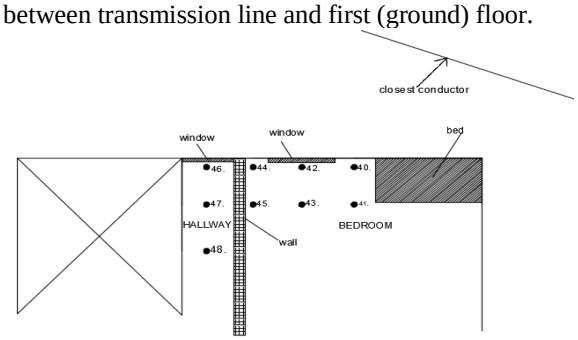


Figure 10: Top view of the attic

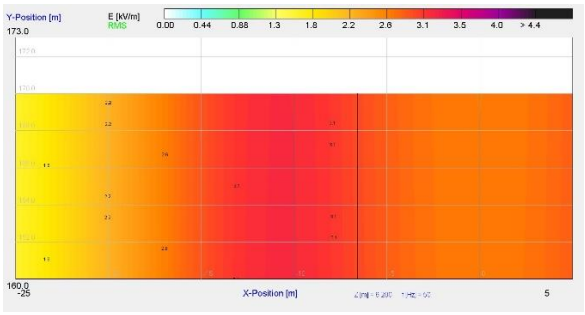


Figure 11: Distribution of unperturbed electric field on the house location on 6.2m height

In the Table 6 the measured and calculated unperturbed values for representative points on the second floor are shown, as well as the shielding factor in the same points.

Table 6: Shielding factor in representative points on the second floor (attic)

Measurement point	Measured value (V/m)	Calculated unperturbed value (V/m)	Shielding factor (%)
40.	184,7 ± 7,4	3055	93,95
41.	215,7 ± 8,6	2970	92,74
42.	58,4 ± 2,3	3001	98,05
43.	69,7 ± 2,8	2910	97,61
44.	45,4 ± 1,8	2944	98,46
45.	34,4 ± 1,4	2946	98,80
46.	42,9 ± 1,7	2885	98,52
47.	34,9 ± 1,4	2780	98,71
48.	42,1 ± 1,7	2637	97,28
			average: 97,12

The average shielding factor on the attic is 97,12% and it is lower compared to the shielding factors on ground and first floor. This shielding factor means that the real electric field on the attic is 2,88% of the calculated unperturbed electric field.

The highest electric field strength is in 41. measurement point and there are two reasons for that: first reason is that this point is closest to the transmission line as it can be seen from Figure 10. Second reason is that this point is just below roof window and thus the shielding factor is lower compared to the points that are behind/under the wall. Nevertheless, the electric field strength in this point is still significantly lower compared to the limit values from Table 1 and Table 2.

In the Table 7, the minimum and average shielding factors of each floor are presented. As it can be seen, on ground and first floor, both average and minimum shielding factor have similar values and it can be concluded that the shielding factor of these two floors is same. On the other hand, the shielding factor of second floor (attic) is considerably smaller.

Taking a look at the average shielding factors at each of the floors, it can be concluded that the average shielding factor of this house is 97% or more. This means that the real electric field inside the house is just 3% (or less) of the calculated unperturbed electric field. But observing the minimum shielding factors the different conclusion can be made. In certain occasions, the electric field inside the house is up to 7,5% of the calculated unperturbed electric field as it is situation with the electric field in measurement point 41. This measurement point is specific for two reasons. The first one is that the electric field is measured on the attic, and it is already said that the shielding factor on the attic is lower compared to the lower floors. The other reason is that this measurement point is just below the roof window and it is presumed that the window has lower shielding factor compared to a roof.

Table 7: Average and minimum shielding factor of each floor

Shielding factor					
Ground floor		**First floor**		**Attic**	
Average	**Minimum**	**Average**	**Minimum**	**Average**	**Minimum**
99,2%	**98,1%**	**99,2%**	**98,4%**	**97,1%**	**92,7%**

4 Conclusion

Shielding factor of the house was determined as the combination of approximate measured value of electric field strength inside the house and the unperturbed electric field calculated with the EFC-400 software. For the analyzed house (side wall made of EFE block +insulation and with clay roofing), electric field inside the house is up to 3% of the calculated unperturbed electric field (shielding factor of the house is 97% or more), except in some specific situations. Specific situations refer to measurement locations that are just under the roof where the electric field is up to 5% of calculated unperturbed electric field, or even up to 7,5% if the measurement point is just below roof window.

Literature

[1] N. Y. Jayalakshmi in S. N. Deepa, „Modelling Of Electric And Magnetic Fields Under High Voltage Ac Transmission Line“, str. 9.

[2] „ICNIRP: Guidelines for Limiting Exposure to Time-Varying Electric and Magnetic Fields (1 Hz to 100 kHz), Health Physics, Vol.99, No.6, pp. 828-836, 2010.“ .

[3] B. Cestnik, „“Electric and Magnetic Field in the Vicinity of transmission line”. Original title: “Električno in magnetno polje v okolju elektroenergetskih vodov”, 1999.

[4] H. Javadi, A. Mobarhani, in A. Ameli, „Electric field evaluation of 400kV overhead transmission lines on the earth surface“, v *5th IET International Conference on System Safety 2010*, Manchester, UK, 2010, str. PO3–PO3, doi: 10.1049/cp.2010.0841.

[5] E. Lunca, S. Ursache, in A. Salceanu, „Computation and analysis of the extremely low frequency electric and magnetic fields generated by two designs of 400 kV overhead transmission lines“, *Measurement*, let. 124, str. 197–204, avg. 2018, doi: 10.1016/j.measurement.2018.04.012.

Dinamična *u-t* karakteristika varistorjev

Anton Zhezhov, Mislav Trbušić in Anton Hamler

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, 2000 Maribor, Slovenija

E-pošta: anton.zhezhov@student.um.si

Dynamic *u-t* characteristic of varistors

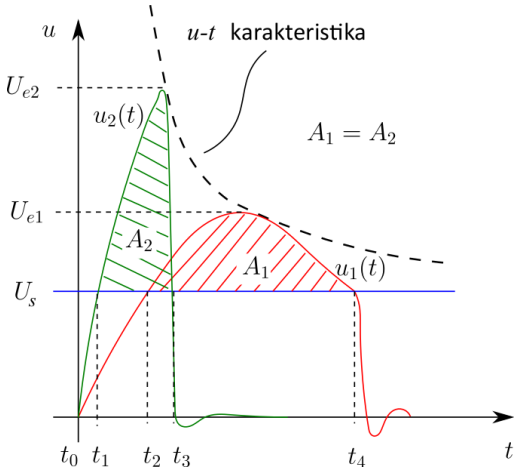
Abstract. *This short paper presents a brief introduction to surge voltage wave impact on a metal-oxide material based varistors, which are commonly used to suppress overvoltages in electrical power systems and devices. The article is focused on the measurements of dynamic u-t characteristic of low voltage application varistor. The results show a good agreement with theoretical predictions which are based on the breakdown mechanism in gasses.*

1 Uvod

Ščitenje elektronskih naprav proti hitrim tranzientnim prenapetostim predstavlja s stališča zanesljivosti pomemben vidik pri načrtovanju vezij, saj so polprevodniški elementi še posebej občutljiva na tovrstne motnje, ki imajo nemalokrat za posledico uničenje elementa. Iz omenjenega razloga je zaželeno omejiti vpliv prenapetosti na delovanje vezja, za kar se uporabljajo metal oksidni nelinearni upori, ki jih zaradi namena imenujemo tudi prenapetostni odvodniki oziroma varistorji. Glede na širok razpon in značaj prenapetosti, ki se pojavljajo v nizkonapetostnem omrežju je smotrno poznati odziv tovrstnih zaščitnih elementov pri počasnih in hitrih prenapetostnih pojavih. Z namenom ovrednotenja odziva varistorja na udarno napetost, so v članku predstavljene meritve napetostnih razmer na varistorju pri različnih strminah čela udarnega vala [1][2].

2 Merjenje udarne *u-t* karakteristike varistorja

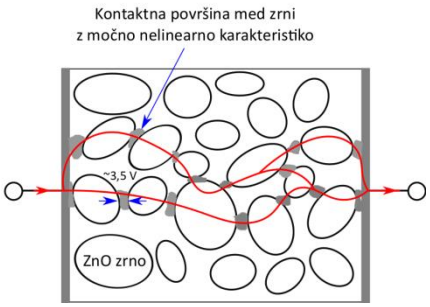
Namen preskušanja varistorjev z udarno napetostjo je posneti *u-t* karakteristike, na podlagi katerih bi se dalo sklepati ali pri kovinsko oksidnih materialih velja zakon enakih površin, podobno kot je to pri plinastih izolantih, kjer je prebojna trdnost odvisna od površine, ki jo napetostni sunek oklepa z vrednostjo statične prebojne napetosti, ki je najnižja napetost pri kateri pride do preboja, kot je shematsko prikazano na sliki 1. Za določeno geometrijo in izolacijski plin je ta površina konstantna, kar pomeni, da bo pri hitrejših napetostnih spremembah, pri bolj strmem čelu udarnega vala, napetost pri kateri pride do preboja višja, kot pa pri počasnih spremembah napetosti, oz. pri bolj položnem čelu udarnega vala [3]. Pri varistorjih je zaradi drugačnega mehanizma prevajanja, smiselno udarno napetost primerjati s statično prevodno napetostjo, oz. napetostjo pri kateri skozi varistor steče nek v naprej predpisani tok npr. 10 mA.



Slika 1: Zakon enakih površin in *u-t* karakteristika, ki velja pri plinskih izolantih

2.1 Varistor

Varistor je nelinearen upor, sestavljen na osnovi ZnO zrn (~90%) in drugih komponent Bi_2O_3, MnO_2, Cr_2O_3, Sb_2O_3, Co_3O_4 (~10%), ki kot celota predstavljajo nehomogen material [2][4]. ZnO zrna, oziroma kontaktna površina med njimi, se obnašajo podobno kot polprevodniški elementi s kopico zanimivih fizikalnih lastnosti, kot so: visoka mobilnost elektronov, visoka toplotna prevodnost, široka energijska reža (energijski pas) in relativno visoka vezalna energija, kar omogoča uporabo ZnO spojin ne samo v varistorjih ampak širše na področju polprevodniške tehnike, npr. v tankoslojnih tranzistorjih, foto-detektorjih, LED diodah, ipd. [5].
Pri varistorjih oz. pri prenapetostnih zaščitnih elementih na osnovi ZnO je ključno, da kontaktna površina med zrni izkazuje usmerniški efekt z relativno visoko pragovno napetostjo, ki znaša približno ~ 3,5 V, tako lahko napetostno in tokovno vzdržnost varistorja pogojujemo s številom zaporednih in vzporednih kontaktnih površin, kjer zaporedna vezava zvišuje napetostni, vzporedna pa tokovni nivo (Slika 2) [6][7]. Povedano drugače, z velikostjo ZnO zrn, ki jih dobimo med procesom sintranja in geometrijo, so v veliki meri določeni parametri varistorja.



Slika 2: Sestava varistorja; prevajalno napetost določajo mejne ploskve med ZnO zrni

Z ozirom na statično oz. kvazistatično *U-I* karakteristiko varistorja, ki jo kaže slika 3, je mogoče delovanje varistorja opredeliti v tri področja, in sicer:

- področje 1, v katerem varistor deluje pri normalnih obratovalnih napetostih. V tem področju ima varistor zelo visoko upornost, pa vendarle je značilno, da kljub temu skozenj teče nek električni tok, ki mu pravimo uhajavi tok.

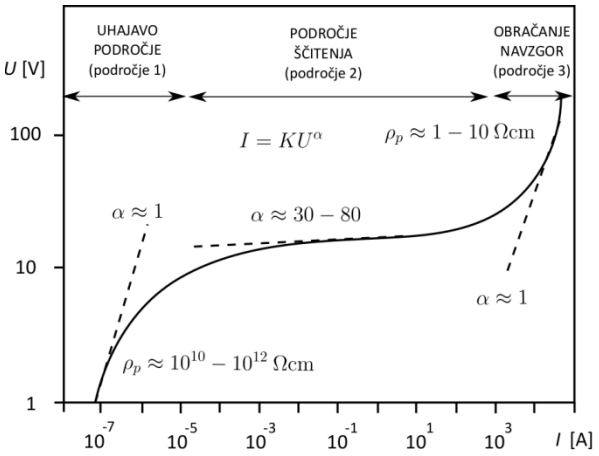
- Področje 2 na *U-I* karakteristiki je izrazito nelinearno in značilno za zaščitno delovanje varistorja pred prenapetostmi. V tem področju želimo da varistor omejuje napetostne sunke.

- Področje 3, ki je ponovno linearno, kjer se *U-I* karakteristika obrača navzgor. Varistor se v tem področju delovanja obnaša podobno kot nizkoohmski upor.

Za posamezno področje delovanja je tokovno napetostna odvisnost podana z izrazom (1)

$$I = KU^{\alpha}\,, \qquad (1)$$

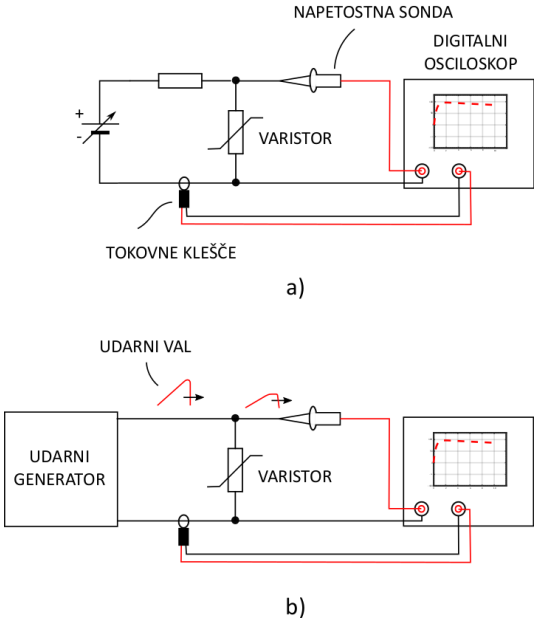
pri tem so: I tok skozi varistor, U napetost na varistorju, α koeficient nelinearnosti in K konstanta varistorja.



Slika 3: Shematski prikaz statične oz, kvazistatične *U-I* karakteristike varistorja za nizkonapetostne aplikacije.

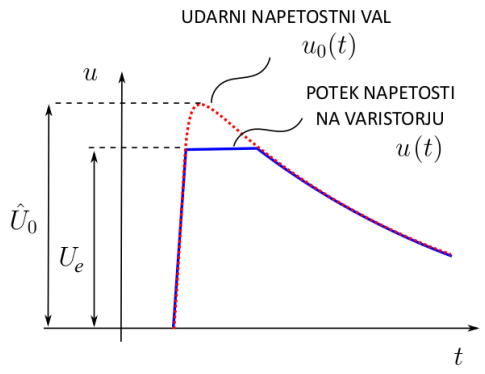
2.2 Merilno vezje in potek meritev

Shematski prikaz merilnega vezja za določanje statične in dinamične prevodne napetosti podaja slika 4. Kot izvor enosmerne napetosti je bil uporabljen enosmerni laboratorijski vir. Udarna napetost za ovrednotenje dinamične prevajalne napetosti je bila ustvarjena z laboratorijskim udarnim napetostnim generatorjem. Merjenje napetosti in toka je bilo izvedeno preko napetostne sonde in tokovnih klešč priključenih na digitalni osciloskop. Uporabljena merilna oprema in tip varistorja so podani v tabeli 1 v dodatku.



Slika 4: Shema merilne vezave za preizkušanje varistorjev z udarno napetostjo; a) za merjenje statične prevajalne napetosti in b) za merjenje dinamičnega odziva.

Ker je bil namen meritev ovrednotiti dinamično *u-t* karakteristiko varistorja, oziroma vrednost udarne napetosti U_e, pri kateri preide varistor v prevodno stanje, je kot referenčna vrednost bila izbrana statično prevodna napetost U_s, oz. tista vrednost enosmerne napetosti, pri kateri skozi varistor teče tok 10 mA. Shema merilnega vezja za določitev U_s je podana na sliki 4-a.
V naslednjem koraku je bil varistor preizkušan še z udarnim napetostnim valom $u_0(t)$ (Slika 4-b), kjer je bilo opravljenih več meritev potekov napetosti $u(t)$ na varistorju pri različnih temenskih vrednostih udarnega vala $\hat{U}_0$, oziroma pri različnih strminah naraščanja napetosti. Pri tem nas je zanimala vrednost napetosti pri kateri varistor omeji špico udarnega vala U_e. Shematsko je prikaz časovnega poteka napetosti na varistorju podan na sliki 5, kjer je prav tako pikčasto vrisan udarni napetostni val.



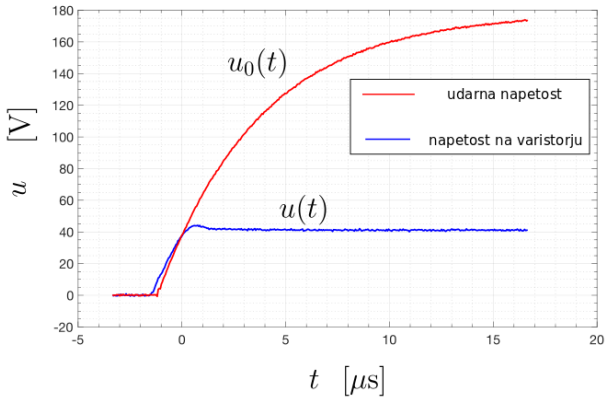
Slika 5: Shematski prikaz napetostnih razmer na varistorju pri preizkušanju z udarno napetostjo

3 Rezultati meritev in komentarji

Na podlagi opisanih merilnih postopkov, smo v prvem koraku izvedli meritev statične prevodne napetosti, ki je za izbrani ZnO varistor tipa ZOV-10D330K znašala $U_s = 34{,}8$ V (pri toku 10 mA).

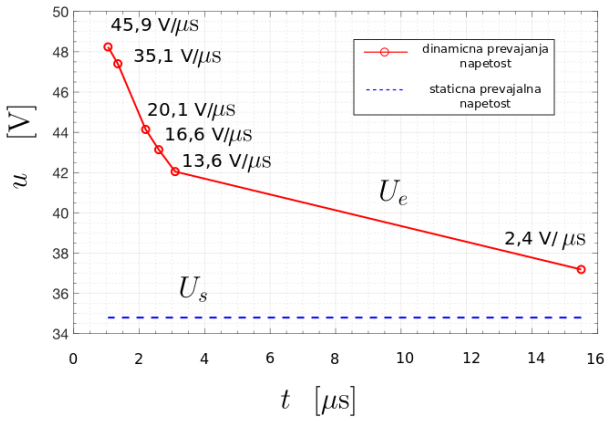
Z namenom določitve dinamične *u-t* karakteristike varistorja smo nato po merilni shemi, ki jo kaže slika 4-b, opravili šest meritev z enojnim udarnim valom, kjer smo vršno vrednost udarnega vala $\hat{U}_0$ spreminjali v mejah od 54-289 V. Za primer $\hat{U}_0 = 179{,}8$ V je na sliki 6 prikazan z osciloskopom izmerjeni časovni potek napetosti na varistorju, kot tudi referenčni udarni val, ki je bil izmerjen brez priključenega varistorja. V konkretnem primeru vidimo, da pri udarnem valu z dvižnim časom $t_d = 10{,}6\ \mu$s varistor omeji napetost v času $\Delta t = 2{,}2\ \mu$s na vrednost $U_e = 44{,}1$ V, kar je občutno višja vrednost kot statična prevodna napetost $U_s = 34{,}8$ V. Razmerje med napetostji U_e in časom Δt v katerem začne varistor prevajati podaja strmino naraščanja napetosti k_e, ki je odločilen podatek pri določanju dinamične *u-t* karakeristike varistorja (2).

$$k_e = \frac{U_e}{\Delta t}, \qquad (2)$$



Slika 6: Časovni potek meritev napetosti na varistorju pri preizkušanju z udarnim valom.

Rezultati predstavljeni v obliki dinamične *u-t* karakteristike varistorja so podani na sliki 7, iz katere je razvodno, da ima strmina naraščanja udarnega napetostnega vala oz. t.i. strmina čela udarnega vala občuten vpliv na dinamično prevajalno napetost U_e, ki se s strmejšim naraščanjem (krajši dvižni čas do začetka prevajanja) pomika k višjim vrednostim.



Slika 7: Izmerjena dinamična *u-t* karakteristika ZnO varistorja, pri preizkušanju z udarnim napetostnim valom (rdeče) in statična prevajalna napetost izmerjena pri enosmerni napetosti (modro-črtkano).

4 Zaključek

Na podlagi rezultatov meritev z udarnim valom, ki smo jih opravili na varistorju ZOV-10D330V lahko sklepamo, da je dinamična prevajalna napetost varistorja močno odvisna od strmine udarnega vala in vsekakor višja, kot pa statična prevajalna napetost, ki jo izmerimo pri enosmernih oz. počasi spreminjajočih napetostih. Zanimivo je povleči vzporednico s prebojno napetostjo v plinih in tekočinah, ki je na podoben način odvisna od strmine čela udarnega vala, kar je tudi motiviralo avtorje, da pristopijo k raziskovanju na tem področju.

Načrtovanju prenapetostne zaščite predstavlja v praksi velik inženirski izziv. Ne le, da so oblike vršne vrednosti in pojavi prenapetosti v omrežju stohastične narave, marveč vidimo, da se tudi materiali, ki se uporabljajo v zaščitnih elementih odzivajo drugače na hipne in počasne napetostne spremembe. Iz tega razloga je obravnavana tematika aktualna tudi v širšem kontekstu uporabe elektrotehniških materialov v izolacijski in polprevodniški tehniki.

Dodatek

Tabela 1: Uporabljena merilna oprema.

Oprema	Tip
Osciloskop	Keysight MSO-X 4054A
Napetostna sonda	Keysight N2790A (max 1000 V)
Tokovne klešče	CAG E3N
Napetostni vir	Laboratorijski napajalnik 24V
Udarni generator	Dvostopenjski Marx - 600 V / 0,6 J
Preizkušai varistor	ZOV-10D330K

Zahvala

Zahvaljujem se vodstvu Laboratorija za aplikativno elektromagnetiko na UM-FERI, ki je podprlo nastajanje članka in mi omogočilo sodelovanje na študentkem tekmovanju v sklopu ERK 2020.
Anton Zhezhov, študent zaključnega letnika na univerzitetnem študijskem progrmu elektrotehnika UM-FERI.

Literatura

[1] M. Lipovec, »Zaščita proti prenapetosti v nizkonapetostnem omrežju«, diplomsko delo, Fakulteta za elektrotehniko, računalništvo in informatiko, UM, Maribor, 2019.
[2] Vishay BC components, Resistive Products - technical note, document number: 29079, 2019.
[3] M. Bizjak, M. Bekovic and A. Hamler, *Spark Breakdown in Gas-Discharge-Tube Surge Arrester at Voltage Pulse*, IEEE Transactions on Power Delivery, vol. 30, No. 3, 2015.
[4] J. Hrastnik, »Razvoj metode za ugotavljanje funkcionalnosti kovinsko oksidnih odvodnikov prenapetosti«, doktorska disertacija, Fakulteta za elektrotehniko, računalništvo in informatiko, UM, Maribor, 2011.
[5] A. Janotti and C. G. Van de Walle, Fundamentals of zinc oxide as a semiconductor, Rep. Prog. Phys. 72, 126501, 2009.
[6] Z. Topcagic, T. Tsovilis and D. Krizaj, *Modeling of current distribution in zinc oxide varistors using Voronoi network and finite element method*, Electirc Power System research, Vol. 164, pp. 253-262, 2018.
[7] A. Rečnik, S. Bernik and N. Daneu, *Microstructural engineering of ZnO-based varistors ceramics*, J. Mater. Sci., Vol. 47, pp. 1655-1668, 2012.

Načrtovanje laboratorijskega udarnega napetostnega generatorja

Mislav Trbušić, Marko Jesenik, Mladen Trlep in Anton Hamler

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, 2000 Maribor, Slovenija

E-pošta: mislav.trbusic@um.si

Design of a surge voltage generator

Abstract. *To asset a withstand voltage level of semiconductor elements and circuits on fast transient voltages such as surge voltage, it is desirable to have a device that is capable to produce a surge wave with the prescribed wave shape and voltage level. Such devices are called surge wave generators. For low voltage applications, up to 10 kV, it is possible to build up such generators based on MOSFET technology, which turns to be a proper solution for research experiments. The paper presents a brief introduction to the design principles of surge wave generators, where the parameters of the generator are calculated by the usage of Angelini's method. The verification of the method is confirmed by simulations.*

1 Uvod

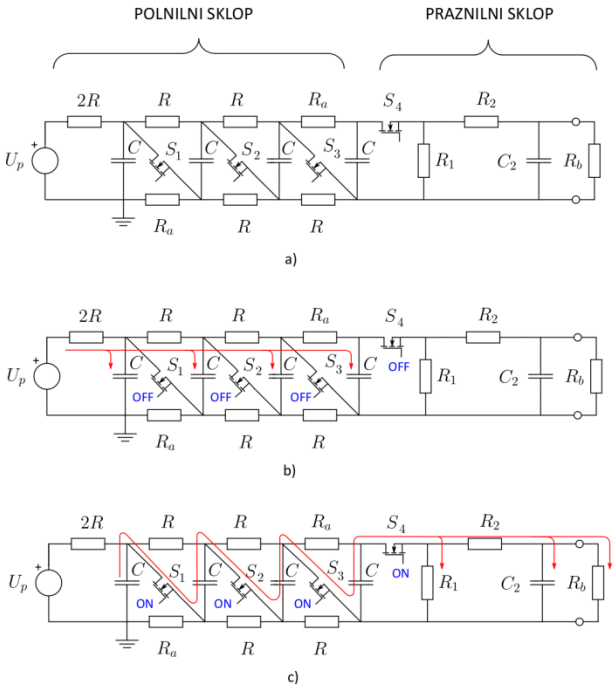
Naprave priključene na električno omrežje, so med obratovanjem pogosto izpostavljene prenapetostim, ki so posledica atmosferskih razelektritev, stikalnih manevrov, prehodnih in resonančnih pojavov v omrežju. Takšne napetostne anomalije lahko poškodujejo ali uničijo električno napravo, oziroma elemente v njej. Še posebej občutljive na prenapetosti so elektronske komponente (tranzistorji, diode, integrirana vezaja, ipd.) [1]. V splošnem, je za ovrednotenje vzdržnosti izolacije elementov, te potrebno testirati na udarno napetost, ki je hipnega in tranzientnega značaja. Za tak namen se uporabljajo udarni napetostni (in tokovni) generatorji, kateri so zmožni proizvesti aperiodično obliko udarnega vala s karakterističnim strmim čelom in položnim hrbtom [2][3]. Za namen preizkušanja polprevodniških elementov, ki se nahajajo v napravah z napetostnim nivojem do 300 V, je v Laboratoruju za aplikativno elektromagetiko (UM FERI) bil izdelan 900 V štiristopenjski udarni napetostni generator, kjer je napetostni množilnik izveden z elektronskim proženjem preko MOS tranzistorjev (MOS je okrajšava za Metal-Oxide-Semiconductor). V članku so predstavljeni osnovni principi načrtovanja udarnega generatorja.

2 Udarni napetostni generator

V praksi se srečuje več tipov udarnih napetostnih generatorjev, med katerimi je najbolj pogost t.i. Marxov kaskadni množilnik, oziroma kar Marxov generator [4][5]. V referatu je predstavljeno načrtovanje laboratorijskega udarnega napetostnega generatorja z uporabo MOSFET tehnologije.

2.1 Nadomestno vezje

Splošno so udarni generatorji sestavljeni iz dveh sklopov, in sicer iz polnilnega in praznilnega sklopa. V prvem sklopu se generatorju iz enosmernega vira napetosti dovaja energija, ki se skladišči v kondenzatorjih in se nato v praznilnem procesu sprosti preko praznilnega sklopa, kot je prikazano na sliki 1. Med procesom polnjenja MOS tranzistorji ne prevajajo, medtem, ko se udarni val na bremenu doseže s prehodom MOS tranzistorjev v prevodno stanje.



Slika 1: Električna shema štiristopenjskega Marksovega generator v izvedbi z MOS tranzistorji namesto iskrišč: a) opredelitev polnilnega in praznilnega sklopa; b) polnilni proces; c) praznilni proces

2.2 Polnilni proces

Med procesom polnjenja udarnega generatorja, MOS tranzistorji ne prevajajo, zato je aktiven le polnilni sklop, ki ga sestavljajo polnilni upori R in R_a ter kondenzatorji C. Z ozirom na sliko 2, so kondenzatorji priključeni v t.i. lestvično vezje, ki s stališča polnjenja predstavlja vzporedno vezavo kondenzatorjev, saj se pri dovolj dolgem času polnjenja T_p vsi napolnijo na

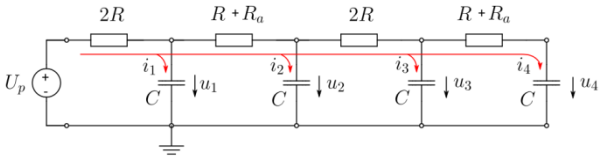
napetost vira U_p. Praviloma so vrednosti kapacitivnosti kondenzatorjev C izbrane tako, da so med seboj enake in da med preizkušanjem zagotovijo neko želeno vrednost električne energije (1):

$$W_e = \frac{n\,C\,U_p^2}{2}, \tag{1}$$

kjer C pomeni vrednost polnilnega kondenzatorja, n število kaskadnih stopenj udarnega generatorja in U_p vrednost enosmerne, polnilne napetosti.
Ker se kondenzatorji ne napolnijo hkrati, je najbolj merodajen podatek čas, v katerem zadnja stopnja doseže vrednost napetosti polnilnega vira U_p. Za natančno določitev časovnih potekov napetosti med polnjenjem je potrebno simultano rešiti sistem linearnih diferencialnih enačb, vendar je za praktično uporabo bolj priročno uporabiti obrazec (2) [6]:

$$T_p \approx 5\,R'\,C\,n^2\ . \tag{2}$$

Pri tem R' predstavlja povprečno vrednost polnilnega upora, ostale oznake pomenijo enako kot prej.



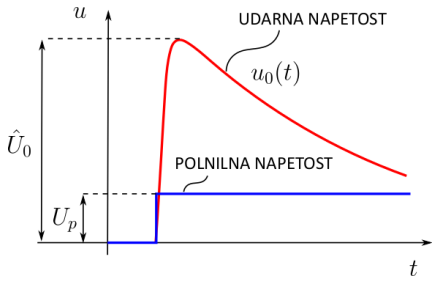
Slika 2: Nadomestno električno vezje med polnilnim procesom udarnega generatorja

2.3 Praznilni proces

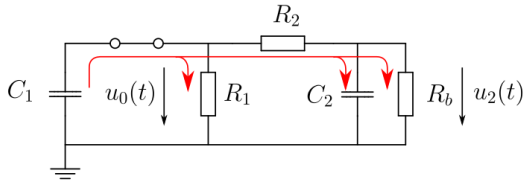
Kot je že bilo omenjeno, je namen napetostnega udarnega generatorja proizvesti določeno aperiodično obliko napetosti t.i. udarni val, katerega vršna vrednost $\hat{U}_0$ nekaj krat presega vrednost polnilne napetosti U_p (Slika 3). To je doseženo tako, da se preko MOS tranzistorjev kondenzatorji povežejo zaporedno, kar pomeni, da se napetosti na kondenzatorjih seštejejo. Razmerje med vršno vrednostjo napetosti $\hat{U}_0$ udarnega vala in polnilno napetostjo U_p je odvisna od karakterja preizkušanca (kapacitiven, ohmski, induktivni oz. kombinacija teh). Vendar za večino primerov, ki so ohmskega ali kapacitivnega značaja velja, da je to razmerje enako številu kaskadnih stopenj generatorja n (3).

$$n = \frac{\hat{U}_0}{U_p} \tag{3}$$

Proces praznjenja Marxovega generatorja lahko opišemo glede na sliko 1-c. Predpstavimo, da so kondenzatorji naelektreni tako, da je na njih napetost enosmernega vira. Delovanje MOS tranzistorjev poveže kondenzatorje zaporedno, kar da na izhodu štirikratno vrednost napetosti. Na sliki 4 je podano nadomestno vezje udarnega generatorja, ki je merodajno v procesu praznjenja. Pri tem je impulzna kapacitivnost enaka $C_1 = \frac{C}{n}$.



Slika 3: Shematski prikaz razmerja med enosmerno polnilno napetostjo U_p in aperiodično obliko udarne napetosti $u_0(t)$



Slika 4: Nadomestno električno vezje med procesom praznjenja udarnega generatorja

Napetost na preizkušancu $u_2(t)$ je razlika dveh eksponencialnih funkcij, katere časovne konstante določajo elementi udarnega generatorja in preizkušanec s svojo upornostjo in kapacitivnostjo (4).

$$u_2(t) = \hat{U}_0\left(e^{-\frac{t}{\tau_2}} - e^{-\frac{t}{\tau_1}}\right) \tag{4}$$

Pri tem sta časovni konstanti v eksponencialnih funkcijah enaki: $\tau_1 = R_2C_2$ in $\tau_2 = R_1(C_1 + C_2)$. Povezava med časom čela T_s in časom polovične vrednosti T_r ter časovnima konstantama τ_1 in τ_2 je za standardni udarni val 1,2/50 z upoštevanjem pogoja $R_b \gg R_1, R_2$ sledeč: (5) in (6) [3].

$$T_s = 3\,\tau_1 \tag{5}$$

$$T_r = 0.72\,\tau_2 \tag{6}$$

2.4 Določanje vrednosti elementov

Pri načrtovanju napetostnega udarnega generatorja je potrebno posebno pozornost nameniti izbiri vrednosti elementov v udarnem oz. praznilnem sklopu, saj le ti v veliki meri določajo obliko udarnega vala in časovni konstanti čela ter hrbta. Praviloma so polnilni upori R_a in R izbrani tako, da je njihova ekvivalentna upornost v udarnem režimu bistveno večja kot sta upornosti R_1 in R_2, zato polnilne upornosti praktično nimajo vpliva na obliko udarnega vala in je njihova vrednost pogojena z želenim časom polnjenja kondenzatorjev v polnilnem procesu.
Za štiristopenjski udarni generator ($n = 4$) in z izbrano vrednostjo polnilnega kondenzatorja $C = 1\ \mu\mathrm{F}$ ter

časom polnjenja $T_p = 350$ ms, je po enačbi (2) vrednost polnilnega upora približno enaka $R \approx 4\text{k}7\ \Omega$.

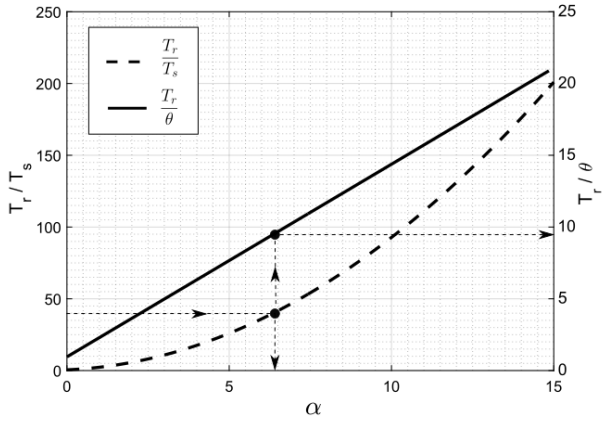
Pristopi k določitvi elementov udarnega generatorja so raznoliki, bolj napredni in računalniško podprti se temeljijo na optimizacijskih postopkih, modalni analizi ali pa karakterističnih vrednostih nadomestnega vezja [7]-[9]. Kakorkoli, v članku bo za ovrednotenje elementov uporabljena Angelinijeva grafična metoda [5].

Za nadomestno vezje udarnega sklopa, kot ga kaže slika 4 se vrednosti uporov R_1 in R_2 določita z ozirom na sliko 5 in enačbe (7)-(9), pri tem je potrebno poznati razmerje med časom polovične vrednosti in časom čela T_r/T_s ter kapacitivnosti C_1 in C_2. Izbira vrednosti kondenzatorja C_2 je poljubna vendar je pri izbiri potrebno upoštevati $C_2 \ll C_1$, v konkretnem primeru smo izbrali $C_2 = 10$ nF.

$$R_1 = \frac{\alpha\,\theta}{C_1 + C_2}\left[1 + \sqrt{1 - X}\right] \tag{7}$$

$$R_2 = \frac{\alpha\,\theta}{C_2}\left[1 - \sqrt{1 - X}\right] \tag{8}$$

$$X = \frac{1}{\alpha^2}\left(1 + \frac{C_2}{C_1}\right) \tag{9}$$



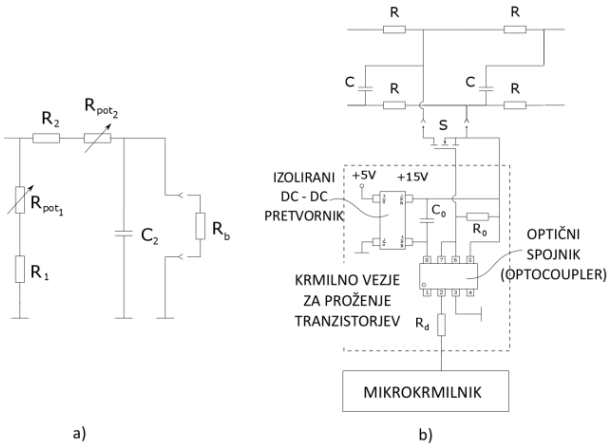
Slika 5: Graf za določaje parametrov α in θ po Angelinijevi metodi [5]

Vrednosti α in θ odčitamo iz grafa (Slika 5) na sledeč način:

- iz krivulje $\left(\frac{T_r}{T_s}\right)$ odčitamo vrednost α pri $T_r/T_s = 41{,}7$ ta znaša $\alpha = 6{,}5$;
- iz krivulje $\left(\frac{T_r}{\theta}\right)$ odčitamo vrednost $T_r/\theta = 9{,}5$ iz katere sledi $\theta = 5{,}26\ \mu s$.

Z upoštevanjem vrednosti $\alpha = 6{,}5$, $\theta = 5{,}26\ \mu$s, $C_1 = 250$ nF in $C_2 = 10$ nF, sta vrednosti uporov določljivi preko (7) in (8), ter enaki $R_1 = 261\ \Omega$ in $R_2 = 42\ \Omega$.

Pri dejanski sestavi generatorja je potrebno predvideti odstopanja fizičnih elementov od nazivnih vrednosti, zato se za prilagoditev oz. umerjanje časovnih konstant T_s in T_r, v praznilni sklop namestijo potenciometri, kot je prikazano na sliki 6-a. Ker je proženje oz. množenje napetosti izvedeno preko MOS tranzistorjev, ki so na različnih električnih potencialnih nivojih, je potrebno zagotoviti galvansko ločitev med krmilnim vezjem in tranzistorji, za kar so bili uporabljeni optični spojniki (optocoupler-ji). Izvedba proženja ene stopnje je podana na sliki 6-b, kjer je povezava prožilnega vezja izvedena med vrati (G) in izvorom (S) tranzistorja. Pri izbiri MOS tranzistorjev je potrebno izbrati takšne, ki imajo zadosti visoko vzdržno napetost med ponorom (D) in izvorom (S), ki naj bo vsaj 20% višja od maksimalne pričakovane napetosti. Smotrno je omeniti, da naj imajo izbrani tranzistorji hiter odziv in nizko prevajalno upornost v nasičenju.



Slika 6: a) Shematski prikaz izvedbe praznilnega ali udarnega oz. impulznega sklopa, kjer sta za umerjnje časovnih konstant T_s in T_r predvidena potenciometra R_{pot_1} in R_{pot_2}. b) Krmilno vezje za proženje tranzistorjev, ki mora zagotavljati galvansko ločitev od mikrokrmilnika.

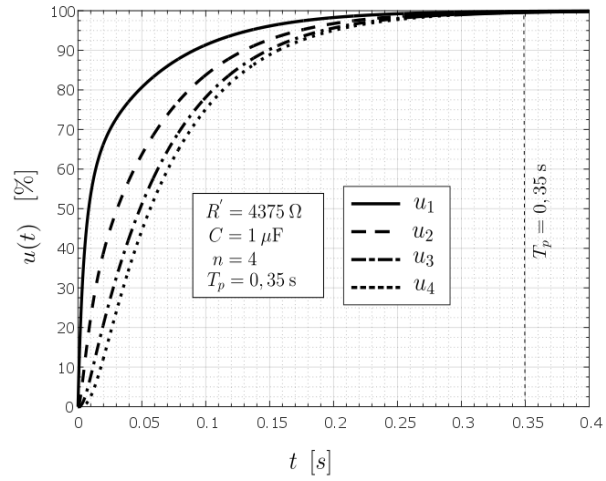
3 Simulacije in komentarji k rezultatom

Na podlagi izračunov so podani rezultati simulacij za štiristopenjski udarni napetostni generator s standardno obliko udarnega vala 1,2/50. Simulacije so bile opravljene v programskem okolju Matlab, kjer je pri polnilnem procesu bila upoštevana vezava, ki jo prikazuje slika 2, medtem ko je bilo v praznilnem procesu upoštevano poenostavljeno nadomestno vezje s slike 4. Vrednosti elementov udarnega generatorja so podane v tabeli 1 na koncu poglavja.

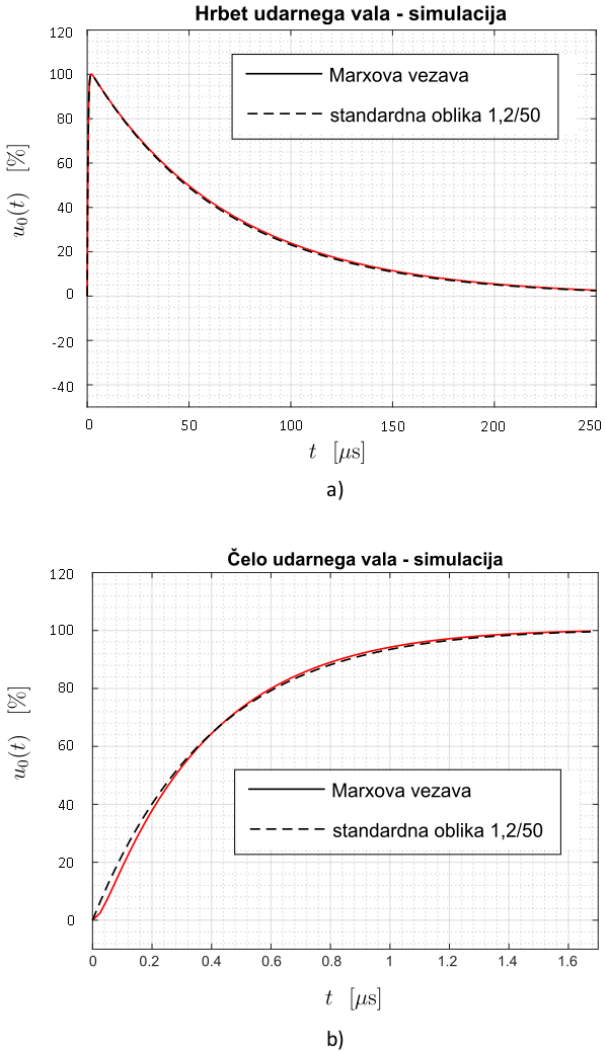
Slika 7 prikazuje izračunane časovne poteke napetosti na polnilnih kondenzatorjih C tekom polnjenja stopenj, pri tem se oznake napetosti navezujejo na sliko 2. Na sliki je prav tako označena vrednost časa polnjenja, izračunana na podlagi (2), iz česar lahko vidimo praktičen vidik izraza (2) pri načrtovanju polnilne stopnje udarnega generatorja.

Izračunane vrednosti potekov udarne napetosti, in sicer za hrbet ter čelo udarnega vala so podane na sliki 8. Pri izračunanih potekih je za primerjavo podana tudi s standardom predpisana oblika udarnega vala 1,2/50. Iz primerjave je razvidno, da ovrednotenje elementov udarnega generatorja po Angeliniju da zadovoljive rezultate, ter se predstavljena metoda lahko priporoči za

uporabo pri načrtovanju laboratorijskega udarnega napetostnega generatorja v Marxovi vezavi.



Slika 7: Časovni poteki napetosti na kondenzatorjih pri polnjenju Marxovega generatorja z ozirom na sliko 2



Slika 8: Časovni poteki napetosti pri praznjenju Marxovega generatorja: a) potek na hrbtu udarnega vala in b) potek na čelu udarnega vala. Za primerjavo je vrisana standardna oblika udarne napetosti 1,2/50 (črtkano), ki je podana z razliko dveh eksponencialnih funkcij $u_{std}(t) = (e^{-15000\,t} - e^{-2470000\,t})$.

Tabela 1: Parametri udarenga generatorja upoštevani v simulacijah

Parameter	vrednost	Opomba
n	4	število stopenj
R	4700 Ω	polnilni upor
R_a	2350 Ω	polnilni upor
C	1000 nF	polnilni kondenzator
C_1	250 nF	impulzna kapacitivnost
C_2	10 nF	praznilna kapacitivnost
R_1	261 Ω	upor za določanje hrbta vala
R_2	42 Ω	upor za določanje čela vala
R_b	100 kΩ	upornost preizkušanca
T_p	350 ms	polnilni čas stopenj
T_s	1,26 μs	časovna konstanta čela vala
T_r	48,7 μs	časovna konstanta hrbta vala

4 Sklepna beseda

V članku je predstavljena metoda izračuna udarnega napetostnega generatorja, ki se v veliki meri opira na Angelinijevo grafično metodo. Na podlagi simulacij lahko podamo sklep, da je predstavljena metoda primerna za praktično uporabo pri načrtovanju tovrstnih generatorjev. Čeprav je v konkretnem primeru bilo obravnavano načrtovanje udarnega generatorja z vršno napetostjo 900 V, pa enaki principi izračuna veljajo tudi za generatorje z večjim številom stopenj in višjo preizkusno napetostjo.

Literatura

[1] M. Majcen, »Rešitve v proizvodnji varistorjev v ohišji SMD«, diplomsko delo, Fakulteta za elektrotehniko, računalništvo in informatiko, UM, Maribor, 2016.

[2] IEC 60076-1, Power Transformers, Part 1: General, IEC Standard, 2000.

[3] M. Babuder, »Visokonapetostna tehnika – skripta«, UL-Fakulteta za elektrotehniko,UL, Ljubljana, 2006.

[4] K. Schon, »High Impulse Voltage and Current Measurement Techniques«, Sprnger, 2013

[5] B. Heller, A. Veverka, »Surge Phenomena in Electrical Machines«, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1968.

[6] M. Trbušić, J. Pihler, A. Hamler, *Kompenzacija časa polovične vrednosti pri preskušanju z Marksovim generatorjem*, 25. Posvetovanje Komunalna energetika, Maribor, 2015.

[7] W.M. Al-Hasawi, K.M. El-Naggar, *A Genetic Based algorithm for Digitally Recorded Impulse Parameter Estimation*, IEEE Bolognia Power Tech conference Proceedings, vol. 2, 23-26, 2003

[8] E.A. Feilat, A. Al-Hinai, *HV Impulse Parameter Estimation using ESPRIT-based method*, 2nd International Conference on Electric Power and Energy Conversion Systems (EPECS), 2011.

[9] P. Yutthagowith, P. Tuethong, N. Pattanadech, *Efective Circuit Parameter Determination in lighting Impulse Voltage Tests of air Core Inductors*, 12th IET International Conference on AC and DC Power Transmission, Beijing, China, 2016

Več-vejno delovanje paralelnih močnostnih filtrov

Mitja Nemec

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: mitja.nemec@fe.uni-lj.si

Multi-leg operation of parallel active power filters

Abstract. *The presented research focuses on multi-leg operation of parallel active power filters (PAPF). After the introduction and a brief explanation of basic operation of PAPF the article discusses the impact of inductance design on active power filter operation. Then a multi-leg operation of PAPF is explained. The simulation results show that while multi-leg operation does not have significant impact on reducing switching frequency ripple on the output current, the increased bandwidth improves performance in lower frequency region and reduces the overall size of magnetic components.*

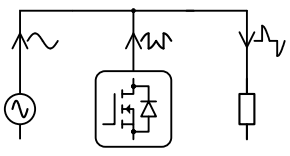
1 Uvod

Eden najbolj izrazitih trendov v močnostni elektroniki v zadnjem času je paralelna vezava močnostnih topologij [1]. Najbolj očitna prednost paralelne vezave je višja dosegljiva moč, ob izbranih stikalnih elementih [2] Dodatno se delno oz. celoti izniči valovitost toka in/ali napetosti, ki je posledica stikalnega delovanja pretvornika. Posledično se zmanjšajo motnje, ki jih naprava oddaja v sistem. Prav tako lahko tudi zmanjšamo velikost pasivnih komponent (dušilke, kondenzatorji) in s tem zmanjšamo in pocenimo napravo.

Vplivi paralelne vezave v DC/DC pretvornikih so že dobro znani in opisani. Na področju AC/DC pretvornikov pa je raziskav nekoliko manj. V članku si bomo pogledali specifike paralelne vezave paralelnih močnostnih filtrov.

2 Osnove delovanja paralelnega aktivnega močnostnega filtra

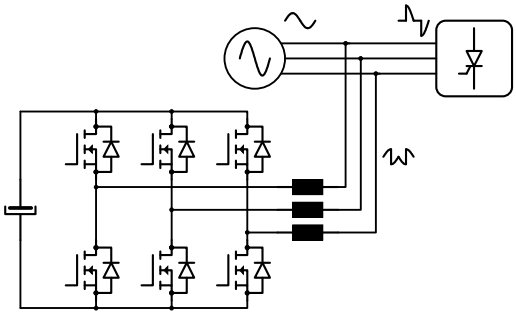
Paralelni aktivni močnostni filter je naprava ki se priklopi paralelno k bremenu, ki je priklopljeno na omrežje [3]. Preko meritve bremenskega toka aktivni filter generira tak tok, da je vsota tokov, ki teče v/iz omrežja sinusna in pri trifaznih izvedbah tudi fazno enakomerno porazdeljena (slika 1)



Slika 1: Paralelni aktivni močnostni filter

Najbolj pogosta izvedba je prikazana na sliki 2. Glavni del filtra predstavlja napetostno napajan trifazni tranzistorski mostič, ki je preko dušilk priklopljen na omrežje.

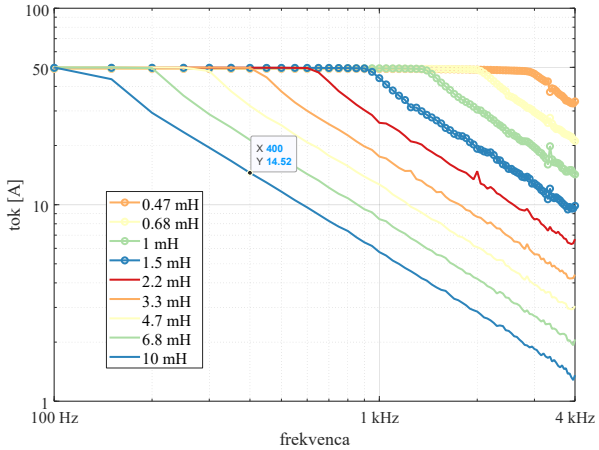
Ob normalnem obratovanju v stacionarnem stanju paralelni aktivni filter potrebuje delovno moč samo za pokrivanje izgub. Vsa preostala moč ki teče preko njega je jalovega značaja. Ob primerno izbrani kapacitivnosti enosmernega tokokroga ter nekoliko bolj zahtevni regulacijski shemi [4], aktivni filter ne potrebuje ločenega napajanja in se napaja kar iz omrežja.



Slika 2: Tipična topologija aktivnega močnostnega filtra

Glavni namen dušilk je omejitev strmine toka, ki je posledica stikalnega delovanja pretvornika. Z večjo induktivnostjo zmanjšamo vpliv stikalnega delovanja vendar s tem omejimo dinamiko filtra (slika 3). Ker pa je filter naprava s katero želimo iz omrežja izločati predvsem tokove višjih frekvenc si poslabšanja dinamike ne želimo. Tako je pri načrtovanju filtra treba iskati kompromis

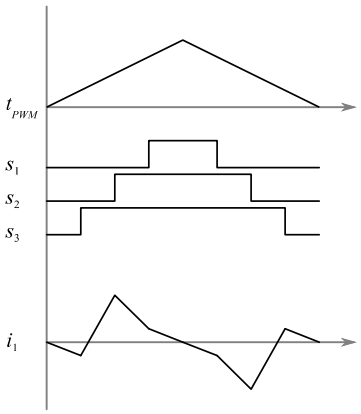
.



Slika 3: Pasovna širina toka v odvisnosti od induktivnosti

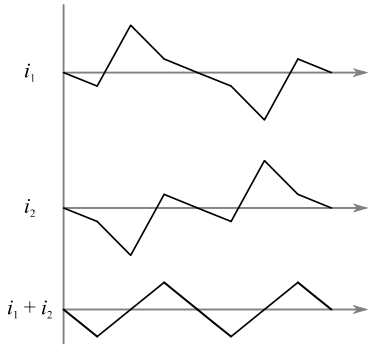
3 Več-vejno obratovanje

Najbolj pogosto uporabljena strategija proženja močnostnih stikal v trifaznem mostiču je modulacija prostorskega vektorja napetosti (SVM) [5]. Pri SVM se na podlagi simetričnega nosilnega signali tvorijo prožilni signali za posamezno vejo (slika 4). Zaradi simetričnega nosilnega signala je sredinsko simetrična tudi valovitost faznega toka.



Slika 4: Nosilni signal (t_{PWM}, prožilni signali (s_1, s_1 in s_3) in valovitost toka pri SVM modulaciji

Pri paralelnem obratovanju več trifaznih pretvornikov je smiselno, da se nosilni signali med seboj fazno premaknejo. S tem lahko nekoliko zmanjšamo valovitost vsote tokov posamezne faze (slika 5) kot tudi valovitost toka v enosmernem tokokrogu [6]. Učinek je v tem primeru bistveno manjši kot pri DC/DC pretvornikih zaradi dveh razlogov. Prvič, posamezna veja nikoli ne obratuje v stacionarnem stanju saj se ji vklopno razmerje stalno spreminja. Drugič, valovitost toka v posamezni fazi je posledica preklopnega stanja te faze in tudi ostalih dveh faz.

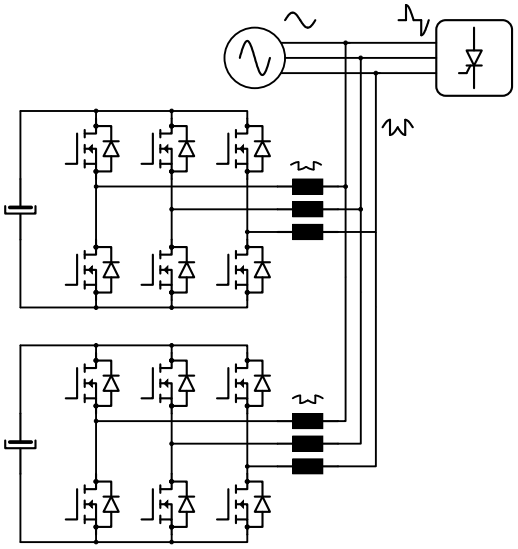


Slika 5: Vsota dveh tokov (i_1 in i_2) pri paralelni vezavi dveh trifaznih pretvornikov

Pri več-vejnem obratovanju (slika Error: Reference source not found) se pri enaki moči tudi spremeni tokovna obremenitev dušilk. Če privzamemo, da je volumen dušilke (V_1) proporcionalen produktu njene induktivnosti in kvadrata toka, ki teče čez njo [7] je volumen dušilke eno-vejnega pretvornika

$$V_1 \propto L_1 \cdot i_1^2 \tag{1}$$

Skupni volumen dušilk pri n-vejnem pretvorniku pa znaša



Slika 6: Dve-vejni paralelni aktivni močnostni filter

$$V_n \propto n \cdot L_n \cdot i_n^2 \tag{2}$$

In ker je tok čez posamezno dušilko n-vejnega pretvornika n-krat manjši manjši

$$i_n = \frac{i_1}{n} \tag{3}$$

Je lahko skupni volumen dušilk n-vejnega pretvornika pri enaki vrednosti ($L_1 = L_n$) induktivnosti n-krat manjši.

$$V_n \propto n \cdot L_n \cdot \frac{i_1^2}{n^2} = L_n \cdot \frac{i_1^2}{n} = \frac{V_1}{n} \tag{4}$$

Ker pa je valovitost toka pri n-vejnem pretvorniku še nekoliko manjša, so lahko za doseganje enake valovitosti skupnega toka tudi vrednosti induktivnosti nekoliko manjše in tako je volumen induktivnosti lahko še manjši.

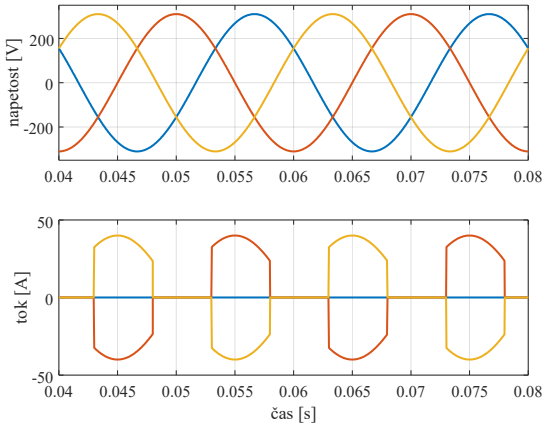
Pri več-vejnem obratovanju je treba omeniti tudi področje zanesljivosti delovanja. Zaradi večjega števila komponent je verjetnost odpovedi seveda večja kot pri enovejnih pretvornikih, vendar pa je v primeru okvare še vedno možno obratovanje z zmanjšano močjo.

4 Rezultati simulacij

Vpliv več-vejnega obratovanja paralelnih močnostnih filtrov smo analizirali simulacijskim paketom MATLAB/Simulink. Simulirano je bilo tudi stikalno delovanje pretvornika vendar so imeli stikalni elementi idealne parametre (upornost 0 Ω v prevodnem stanju, hipni preklop, …) Parametri simulacijskega modela so zbrani v tabeli 1.

Tabela 1. Parametri simulacije

Parameter	Vrednost
Napetost omrežja	400 V
Frekvenca omrežja	50 Hz
Napetost enosmernega tokokroga	720 V
Preklopna frekvenca	20 kHz
Simulacijski korak	500 ns



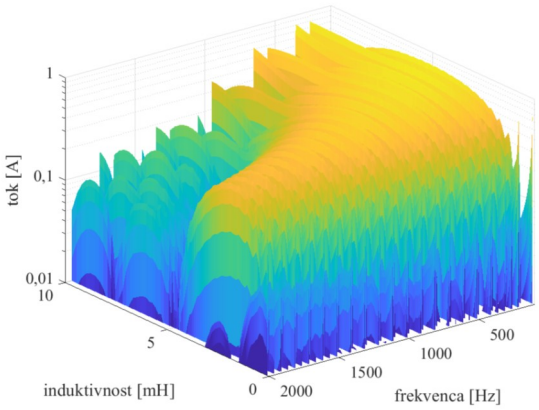
Slika 7: Omrežna napetost in bremenski tok

Slika 7 prikazuje napetost in bremenski tok uporabljen pri simulacijah. Bremenski tok je popolnoma sintetičen, saj v praksi nikoli ne dosežemo neskončne strmine toka in tako predstavlja največji možen izziv za aktivni filter.

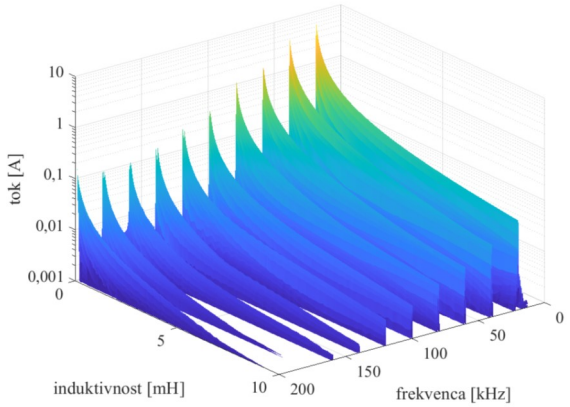
V aktivnem močnostnem filtru je bila za regulacijo toka izbrana metoda neposredne tokovne regulacije [8], s katero dosežemo najboljšo možno dinamiko in se tako izognemo vplivov regulacija na rezultate simulacij. Simulirali smo eno-vejni, dve-vejni, tri-vejni in štiri-vejni paralelni aktivni močnostni filter.

Prvi rezultati (slika 8 in 9) prikazujejo učinkovitost delovanja eno-vejnega aktivnega močnostnega filtra. V nizko frekvenčnem delu spektra (50 Hz - 2000 Hz) prikazanem na sliki 8 vidimo, da pri zelo nizkih induktivnosti (pod 2 mH) filter dobro opravlja svojo vlogo odstranjevanja harmonikov iz omrežnega toka. Delovanje se poslabša pri večjih induktivnosti kar je posledica zmanjšane pasovne širine.

V območju stikalnih frekvenc (20 kHz – 200 kHz) prikazanem na sliki 9 pa amplitude stikalnih harmonikov pričakovano padajo ob povečevanju induktivnosti.



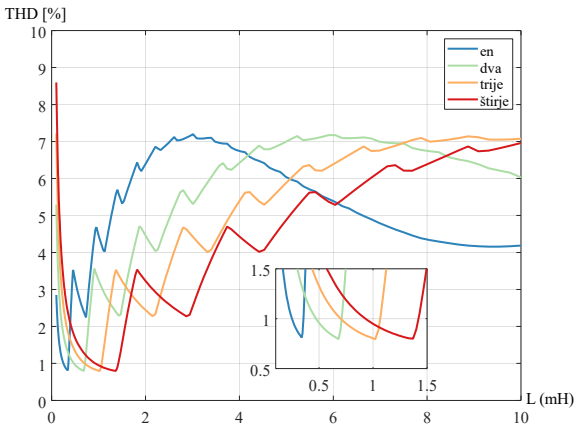
Slika 8: Nizkofrekvenčni del spektra omrežnega toka v odvisnosti od induktivnosti



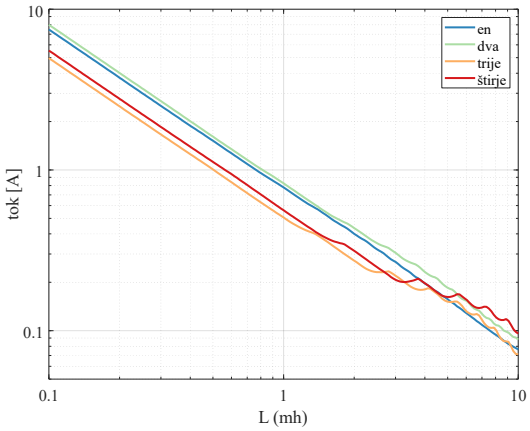
Slika 9: Visokofrekvenčni del spektra v odvisnosti od induktivnosti

Primerjava eno, dve, tri in štiri-vejnega aktivnega močnostnega filtra pa je prikazana na slikah 10 in 11. V nizko frekvenčnem delu spektra je na sliki 10 prikazano harmonsko popačenje v območju med 50 Hz in 2 kHz. Tudi tu je opazno povečanje harmonskega popačenja, ko se poveča induktivnost, vendar je področje nizkega harmonskega popačenja večje, več kot ima aktivni filter paralelnih vej. Glavni razlog za to povečanje je večja pasovna širina katero dosežemo pri enaki induktivnosti z večjim številom vej.

Slika 11 prikazuje efektivno vrednost visokofrekvenčnih komponent toka v odvisnosti od induktivnosti. Pričakovano se le-ta zmanjšuje z naraščajočo induktivnostjo. Nekoliko presenetljiva pa je ugotovitev, da imata razliki med eno in dve-vejnim aktivnim filtrom ter tri in štiri-vejnim aktivnim filtrom obraten smisel in sta zelo majhni (dve-vejni filter ima več visokofrekvenčnih komponent kot eno-vejni).



Slika 10: T.H.D. omrežnega toka v odvisnosti od induktivnosti za eno, dve, tri in štiri-vejni aktivni filter



Slika 11: Efektivna vrednost visokofrekvenčnih komponent toka v odvisnosti od induktivnosti za eno, dve, tri in štiri-vejni aktivni močnostni filter

5 Zaključek

V članku predstavljeni rezultati simulacij potrjujejo, da je več-vejno obratovanje paralelnih močnostnih filtrov možno, vendar pa učinek na zmanjšanje visokofrekvenčnih (stikalnih harmonikov) v izhodnem toku ni izrazit. Vendar pa z več vejnim delovanjem efektivno povečamo pasovno širino regulacije toka. Tako lahko pri več-vejnih pretvornik povečamo vrednost induktivnosti s čimer zmanjšamo visokofrekvenčno valovitost toka pri tem pa ne zmanjšamo učinkovitosti delovanja v nizkofrekvenčnem področju.

V prihodnje velja raziskati vzrok zakaj je pri visokofrekvenčnih komponentah razlika med eno in dve-vejnim paralelnim filtrom v prid eno-vejnega ter potrditi rezultate z meritvami.

Zahvala

Delo je bilo sofinancirano iz programa ARRS »Pretvorniki električne energije in regulirani pogoni« P2-0258 (B).

Literatura

[1] A. Rihar, M. Nemec, and H. Lavrič, “Trendi v razvoju močnostne elektronike za vodenje električnih strojev,” *Elektrotehniški Vestn.*, vol. 86, no. 5, pp. 237–247, 2019.

[2] A. Rihar, P. Zajec, M. Nemec, M. Petkovšek, and D. Vončina, “Načrtovanje, zasnova in izdelava večfaznega razsmernika z MOSFET stikali,” Portorož, Slovenija, 2018, pp. 217–220.

[3] H. Akagi, “Active Harmonic Filters,” *Proc. IEEE*, vol. 93, no. 12, pp. 2128–2141, Dec. 2005, doi: 10.1109/JPROC.2005.859603.

[4] M. Nemec, K. Drobnič, D. Nedeljković, and V. Ambrožič, “Regulacija napetosti pri aktivnem močnostnem filtru,” 2008, p. str. 307-310.

[5] X. Wen and X. Yin, “The SVPWM fast algorithm for three-phase inverters,” in *2007 International Power Engineering Conference (IPEC 2007)*, Dec. 2007, pp. 1043–1047.

[6] M. Nemec, “Tok enosmernega tokokroga pri večfaznih pogonih,” in *ERK 2018*, 2018, p. str. 205-208.

[7] “Ferrites and accessories - Application notes.” Epcos AG, 2017.

[8] K. Drobnič, M. Nemec, D. Nedeljković, and V. Ambrožič, “Predictive direct control applied to AC drives and active power filter,” vol. 56, no. 6, pp. 1884–1893, Jun. 2009.

Iskanje in odpravljanje elektromagnetnih motenj pri štirikvadrantnem pogonu

Mitja Nemec, Andraž Rihar

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: mitja.nemec@fe.uni-lj.si

EMC faults diagnostics and mitigation in four quadrant drive

Abstract. *The article presents the EMC issues during the development of four quadrant frequency inverter. The device consists of two converters. A four leg transistor bridge acts as an active rectifier while three leg bridge operates as an inverter. The source of interference caused by common mode currents was firstly incorrectly attributed to du/dt transients. Following thorough analysis, common mode currents through the control electronics were correctly identified and successfully suppressed using common mode choke.*

1 Uvod

Naprave močnostne elektronike že v sami naravi svojega delovanja zahtevajo tudi obravnavo elektromagnetne združljivosti (EMC). V kolikor gre za naprave ki gred na trg, so standardi, ki jih morajo naprave na področju EMC izpolnjevati, dobro znani. Pri razvoju prototipov in eksperimentalnih modelov pa je bistveno, da se najprej implementira osnovna funkcionalnost. Tako se področja EMC pogosto sploh ne obravnava, oziroma se mu posvetimo le v kolikor se to izkaže kot potrebno.

V laboratoriju za regulacijsko tehniko in močnostno elektroniko (LRTME) smo se pri nadgradnji eksperimentalnega modela štirikvadrantnega frekvenčnega pretvornika, zaradi pojavljajočih motenj, bili primorani ukvarjati s tem področjem.

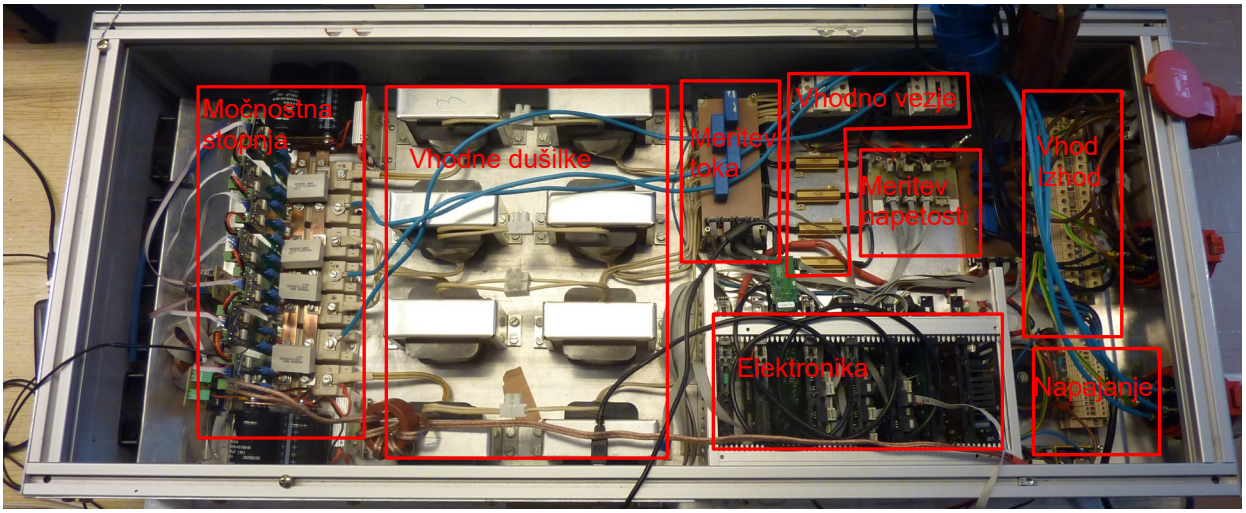
2 Opis sistema

Zgodovina omenjenega štirikvadrantnega frekvenčnega pretvornika se je začela kot aktivni močnostni filter [1]. Pred nadgradnjo je bilo jedro naprave štiri-vejni dvo-nivojski tranzistorski pretvornik. Ob nadgradnji so se dodale še tri tranzistorske veje, ki bi služile kot frekvenčni pretvornik za napajanje električnih strojev, obstoječi štiri-vejni pretvornik pa bi služil kot aktivni usmernik, s katerim bi lahko zagotavljali dvosmerni pretok energije. Naprava je prikazana na sliki 1, kjer lahko vidimo položaje najbolj ključnih komponent.

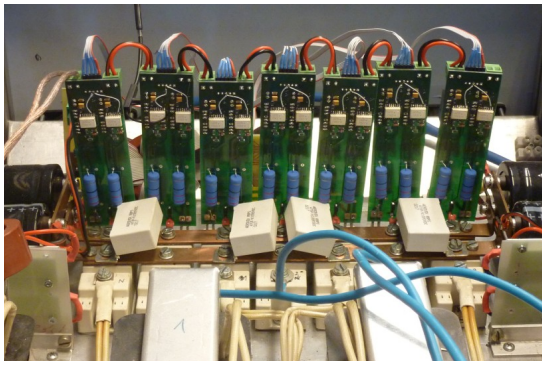
Naprava je grajena za direktni priklop na 400 V trifazno omrežje z možnostjo tri-žilnega (brez nevtralnega vodnika) ali štirižilnega obratovanja. Za meritev toka so uporabljeni zaprtozančni Hallovi senzorji LA 55P, meritev napetosti pa je izvedena preko uporovnih napetostnih delilnikov in instrumentacijskih ojačevalnikov. Digitalizacijo opravljajo trije šest-kanalni, 16 bitni AD pretvorniki ADS8556. Glavna procesna enota pa je mikrokrmilnik TMS320F28346 s taktom 300 MHz. Preostali parametri so navedeni v tabeli 1.

Tabela 1. Nazivni parametri naprave

Parameter	Vrednost
Omrežna napetost	400 V
Omrežni tok	40 A
Napetost enosmernega tokokroga	650 - 850 V
Izhodna napetost	400 V
Izhodni tok	40 A
Preklopna frekvenca	20 kHz



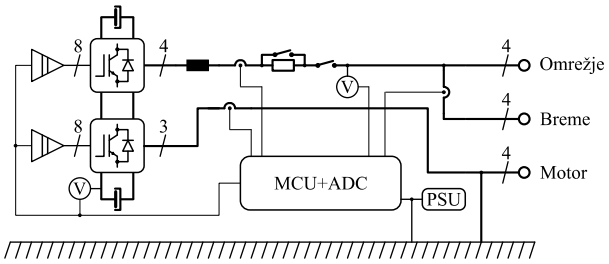
Slika 1: Fotografija pretvornika



Slika 2: Močnostna stopnja s prožilnimi vezji

Močnostna stopnja temelji na IGBT tranzistorskih modulih SKM75GB12, ki imajo nazivno napetost 1200 V in nazivni tok 75 A. V enem modulu se nahajata dva tranzistorja, ki skupaj tvorita eno tranzistorsko vejo. Na tiskanem vezju, ki je pritrjeno neposredno na posamezen tranzistorski modul, se nahajata dve galvansko ločeni prožilni vezji (slika 2). Prožilna vezja so grajena okoli integriranega optično ločenega prožilnega vezja ACPL331J in galvansko ločenega DC/DC pretvornika.

Blokovna shema pretvornika je prikazana na sliki 3.



Slika 3: Blokovna shema pretvornika

3 Kronološki potek

Ko je bila naprava fizično nadgrajena smo pričeli s pripravo programa in preizkušanjem posameznih funkcionalnosti. V začetku je bilo implementirano in preizkušeno delovanje aktivnega usmernika pri znižani omrežni napetosti. Ko je to delovalo smo omrežno napetost povečevali, dokler nismo uspešno potrdili delovanja pri nazivni omrežni napetosti. Ker nismo imeli bremena smo delovanje pri povečani tokovni obremenitvi preizkusili tako, da je aktivni usmernik obratoval z jalovim tok in tudi to obratovanje je potekalo brez večjih težav.

Ko je bilo delovanje aktivnega usmernika preverjeno, smo pričeli s pripravo pisanja programa in preizkušanjem izhodnega pretvornika. Le to je v začetku potekalo pri znižani napetosti omrežja in enosmernega tokokroga.

3.1 Prvi simptomi

Prvi simptomi motenj so se pokazali, ko smo pričeli zviševati omrežno napetost in napetost enosmernega tokokroga. Ko je bila napetost enosmernega tokokroga višja od 150 V, se je sprožila zaščita prožilnih vezij. Proženje zaščite je bilo sporadično, vendar višja kot je bila napetost enosmernega tokokroga, hitreje se je sprožila zaščita. Ker smo ista prožilna vezja kot tudi iste močnostne module uporabljali pri aktivnem usmerniku in tam dodobra preverili njihovo delovanje (pri bistveno višji napetosti enosmernega tokokroga) smo hitro ovrgli sum, da bi bilo kaj narobe s samo močnostno stopnjo.

3.2 Prva razlaga in ukrepi

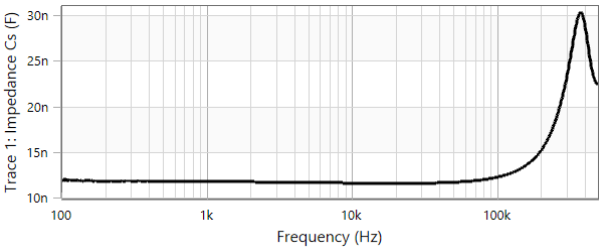
Prva razlaga za to je bila, da velika strmina du/dt, ki je prisotna na izhodih posameznih tranzistorskih vej inducira motnje na signalnih povezavah. Ker je signal s katerim prožilno vezje proži napako tipa *odprti kolektor* (open collector), je signal v normalnem obratovalnem stanju bolj dovzeten za inducirane motnje.

Razlago zakaj se ta motnja ni izražala, ko smo preizkušali vhodni pretvornik smo poiskali v tem, da so izhodi posamezne tranzistorske veje vhodnega pretvornika z zelo kratkimi povezavami priključeni na omrežne dušilke. Tako so visoke strmine du/dt omejene na majhen prostor. Pri izhodnem pretvorniku pa te povezave potekajo čez celotno dolžino naprave in tako širijo motnje po vsem sistemu.

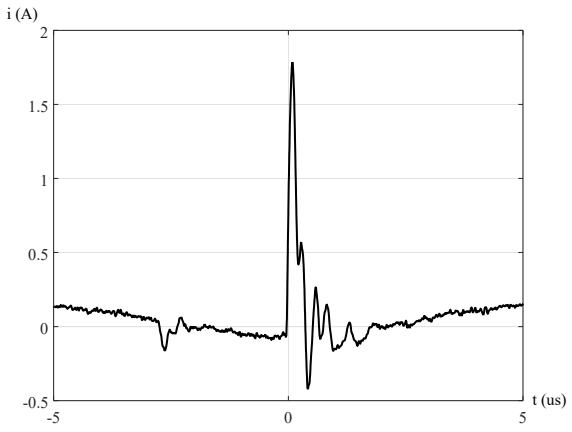
Da bi to omejili smo povezave izhodnega pretvornika proti motorju v celoti izvedli z oklopljenim kablom. Ker oklop ni imel niti najmanjšega vpliva na zmanjšanje motenj se je ta razlaga izkazala za napačno, mi pa smo morali iskati razloge za napačno delovanje naprave drugje.

3.3 Druga razlaga

Naslednji vzrok za omenjeno delovanje smo poiskali v sofaznih tokovih. Zaradi kapacitivnega sklopa navitij električnega stroja proti ohišju (slika 4) ob preklopu tranzistorjev stečejo veliki sofazni tokovi (slika 5).

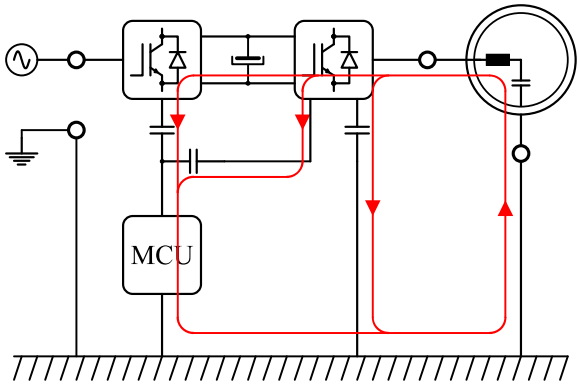


Slika 4: Kapacitivnost navitij stroja proti ohišju



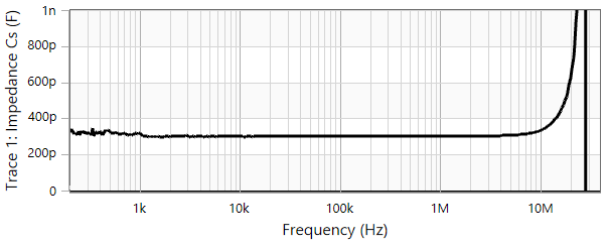
Slika 5: Sofazni tok v električni stroj pri preklopu

Z dodatnimi meritvami celotnega sistema smo ugotovili, da sofazni ne tečejo le med strojem in močnostno stopnjo temveč po celotnem sistemu (slika 6).



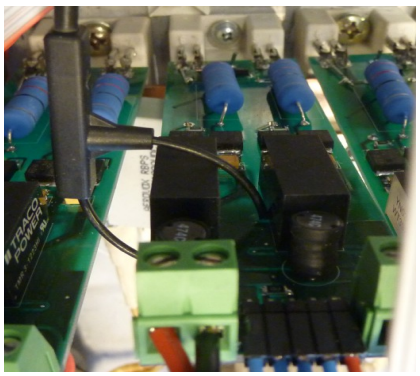
Slika 6: Potek sofaznih tokov

Sofazni tokovi, ki tečejo med električnim strojem in pretvornikom so bili pričakovani. Le ti tečejo od pretvornika proti stroju, nato preko ozemljitvenega vodnika nazaj v pretvornik in po ohišju oz. hladilnem rebru preko kapacitivnega sklopa med hladilnikom in močnostnimi moduli nazaj v pretvornik. Kapacitivnost posameznega modula proti ohišju je približno 300 pF (slika 7).

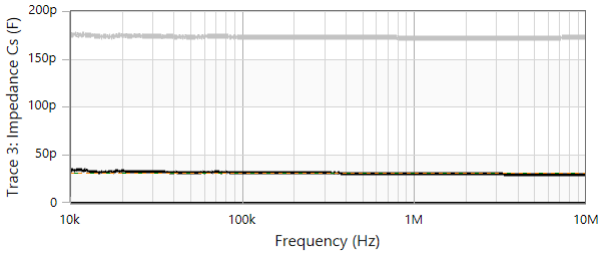


Slika 7: Kapacitivnost močnostnega modula proti ohišju

Presenetili pa so nas predvsem tokovi, ki tečejo med močnostno stopnjo preko prožilnih vezij proti mikrokrmilniku in na ozemljitev. Z meritvami preko posameznega prožilnega vezja (slika 8) in z meritvijo kapacitivnosti med primarjem in sekundarjem DC/DC pretvornika (slika. 9). smo ugotovili, da le-ti tečejo predvsem preko DC/DC pretvornika, saj je kapacitivni sklop optično ločenega prožilnega vezja za dva velikostna razreda manjši. Zaradi tega smo se odločili za zamenjavo DC/DC pretvornikov z modelom, ki ima bistveno nižjo kapacitivno sklopitev. Ker je sofazni tok od prožilnih vezji proti elektroniki, tekel po kablu skupaj s signalom s katerim je prožilno vezje prenašalo status napake, je tako lahko prišlo do proženja napake.

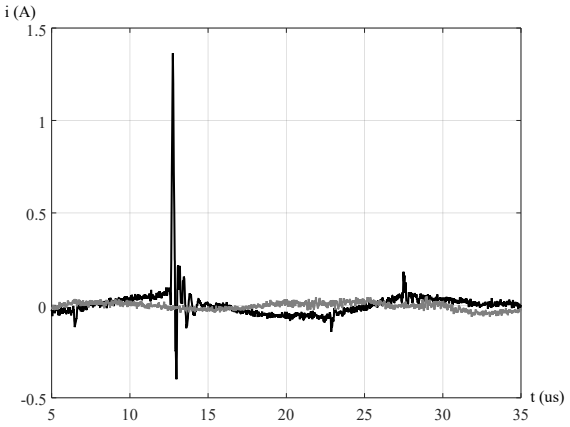


Slika 8: Meritev sofaznega toka preko prožilnega vezja



Slika 9: Kapacitivnost med primarjem in sekundarjem DC/DC pretvornika (siva TMR3-1223, črna TMR3-1223HI)

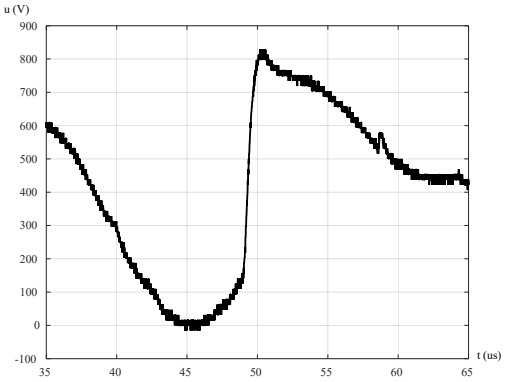
Z zamenjavo DC/DC pretvornikov in z dodanimi sofaznimi dušilkami med elektroniko in prožilnimi vezji smo sofazni tok med prožilnimi vezji in elektroniko zadovoljivo zmanjšali (slika 10) in tudi odpravili razlog za pojav napake.



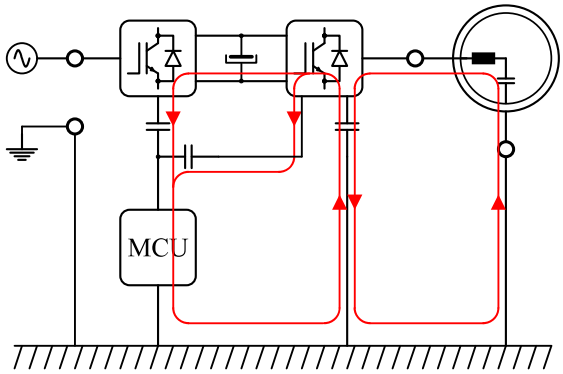
Slika 10: Sofazni tok med prožilnimi vezji in elektroniko pred (črna) in po uvedbi zaščitnih ukrepov (siva)

3.4 Tretja razlaga

Povečanje sofazne impedance med močnostno stopnjo in elektroniko ni imelo nikakršnega vpliva na zmanjšanje sofaznih tokov med močnostno stopnjo in električnim strojem. V nadaljevanju smo se zato nekoliko bolj poglobili v razloge za nastanek sofaznih tokov in z dodatnimi meritvami smo prišli do naslednjega sklepa. Ob preklopu močnostne stopnje med močnostno stopnjo in električnim strojem steče sofazni tok. Ker je kapacitivnost med navitji stroja in ohišjem bistveno večja kot kapacitivnost med močnostno stopnjo in ohišjem, sofazni tokovi bistveno bolj spremenijo sofazno napetost med močnostno stopnjo in ozemljitvijo, kot pa napetost med navitji in ohišjem stroja. Tako se napetost med enosmernim tokokrogom in ozemljitvijo ob vsakem preklopu bistveno spremeni (slika 11). Ta sprememba potenciala povzroči, da stečejo tudi sofazni tokovi med močnostno stopnjo in elektroniko (slika 12).

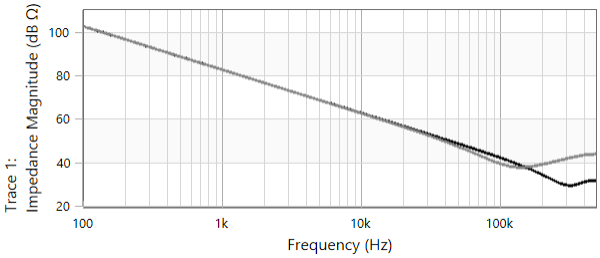


Slika 11: Sofazna napetost enosmernega tokokroga

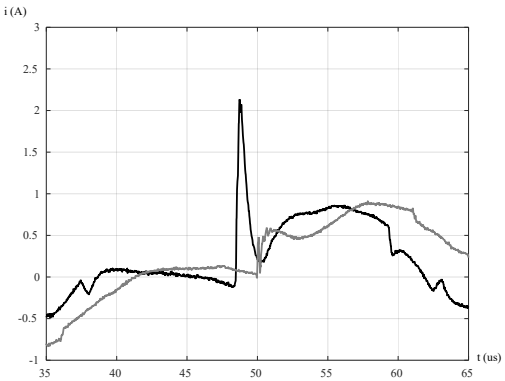


Slika 12: Alternativni pogled na potek sofaznih tokov

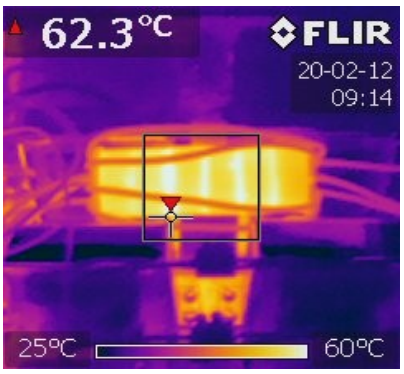
Z vstavitvijo sofazne dušilke med močnostno stopnjo in električni stroj povečamo impedanco predvsem v področju višjih frekvenc (slika 13) in s tem drastično zmanjšamo sofazni tok (slika 14). Vendar pa je pri dimenzioniranju sofazne dušilke treba biti pozoren tudi na izgube v njej, saj je v našem primeru prišlo do pregrevanja (slika 15)



Slika 13: Impedanca stroja proti ohišju z (siva) in brez (črno) sofaznih induktivnosti



Slika 14: Sofazni tok električnega stroja brez (črna) in z vstavljeno sofazno dušilko (siva)



Slika 15: Pregrevanje toroidnih jeder sofazne dušilke

4 Postranske ugotovitve

Tekom raziskav, ki so se primarno ukvarjale z EMC na omenjeni napravi, smo prišli tudi do nekaterih s tem nepovezanih ugotovitev.

Razvoj naprave je potekal po majhnih korakih, posamezni razvojni cikli pa so bili kronološko med seboj precej oddaljeni/razmaknjeni. zato bi morala biti dokumentacija bistveno boljša, podrobnejša in obširnejša.

Za poglobljene študije sofaznih tokov bi bilo zelo praktično, v kolikor bi hitrost preklopov močnostnih tranzistorjev nastavljali s programsko opremo in ne preko strojne opreme (vrednosti uporov v vratih tranzistorja).

Na spreminjanje sofazne napetosti enosmernega tokokroga bi lahko posumili že bistveno bolj zgodaj. Zaradi nezanesljive meritve sofaznih tokov v okolici enosmernega tokokroga smo namreč napak sklepali, da so izenačevalni upori na kondenzatorjih enosmernega tokokroga prav tako kapacitivno povezani z ohišjem, kar pa kasnejša meritev ni potrdila.

Induktivnost sofazne dušilke ni edini parameter, ki določa njeno učinkovitost. Pri frekvencah, s katerimi smo imeli opravka, je bistven vpliv imel tudi material jedra dušilke (ferit oz. vitroperm).

5 Zaključek

Čeprav naprava sedaj zadovoljivo služi svojemu namenu, delo na njej na področju EMC še ni končano. V prvi fazi bo treba pravilno načrtovati in vstaviti sofazni filter v smeri proti električnemu stroju.

V nadaljevanju bi veljalo tudi raziskati, kako na sofazne tokove vplivajo hitrosti preklopov, kar je zanimivo predvsem v luči vpeljave sodobnih polprevodniških tehnologij (SiC, GaN), kjer so hitrosti preklopov še bistveno višje.

Zahvala

Delo je bilo sofinancirano iz programa ARRS »Pretvorniki električne energije in regulirani pogoni« P2-0258 (B).

Literatura

[1] M. Nemec, “Prediktivne metode pri reguliranih trifaznih pretvorniških sistemih : doktorska disertacija / Mitja Nemec,” Univerza v Ljubljani, Ljubljana, 2008.

Dvojni DCT regulator za regulacijo harmonskih komponent

Denis Sušin

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana
E-pošta: denis.susin@fe.uni-lj.si

Dual DCT controller for harmonics control

Abstract. *In this paper the dual DCT controller for harmonics control of RLC circuit is introduced. Dual DCT controller is a novelty that has not yet been mentioned in the relevant literature before. It poses an alternative to the multiple resonant controllers. The total number of harmonics controlled with multiple resonant controllers is limited by the microprocessor computing power. In contrary, dual DCT controller can control arbitrarily selected harmonics with constant computing time regardless of the total number of controlled harmonics.*

1 Uvod

Na področju elektrotehnike ima regulacijska tehnika pomembno vlogo. Električni prehodni pojavi hitro izzvenijo, sistem pa preide v stacionarno stanje, ki ga moramo zaradi varnosti in funkcionalnosti imeti pod nadzorom. Ta nadzor zagotavlja regulacijski algoritem. V stacionarnem stanju električni signali običajno zavzemajo konstantne vrednosti ali pa so periodični. Tovrstni signali so v praksi zelo pogosto obremenjeni s periodičnimi motnjami, ki se odražajo kot višjeharmonske komponente v napetostih in tokovih ter morebiti tudi v drugih veličinah.

Če tovrstne motnje niso zaželene, jih je treba odpraviti. S klasičnimi prijemi (P, PI ali PID regulator) periodičnih motenj ne moremo dovolj dobro zadušiti, zato moramo uporabiti naprednejše regulacijske algoritme [1]. V literaturi za namensko regulacijo osnovne harmonske komponente in višjeharmonskih komponent napetosti in/ali tokov v stacionarnem stanju zasledimo uporabo naslednjih regulatorjev: več resonančnih (mRES) regulatorjev, repetitivni regulator in regulator s filtrom, ki temelji na diskretni kosinusni transformaciji (DCT – angl. Discrete Cosine Transform) [2]. Našteti regulatorji največkrat delujejo vzporedno z drugim regulatorjem (npr. PI), ki skrbi za ustrezno obratovanje med prehodnimi pojavi. Pravilno parametrirani namenski regulatorji harmonskih komponent ne povzročajo dodatnih prenihajev med prehodnimi pojavi, pač pa občutno izboljšajo razmere v stacionarnem stanju.

Regulacijo harmonskih komponent najbolj pogosto srečamo na področju močnostne elektronike, predvsem pri omrežno priključenih napravah [3]–[5]. Za uspešnejše odpravljanje negativnih vplivov stikalne frekvence na frekvenčni spekter izhodnih napetosti in/ali tokov se namesto RL pasivnih filtrov vedno bolj uporablja filtre višjih redov (LC, LCL in CLC filtre) [6]. S stališča zagotavljanja stabilnosti regulacijskega sistema je uvajanje filtrov višjih redov neugodno. V tovrstnih situacijah parametriranje repetitivnega in DCT regulatorja postane zelo zahtevno. Repetitivni regulator in DCT regulator sicer lahko regulirata večje število harmonskih komponent H kot mRES, a imata eno prostorsko stopnjo manj, kar lahko povzroča resne težave s stabilnostjo. Zaradi tega je treba zmanjšati vrednosti ojačanj, kar posledično poslabša dinamični odziv repetitivnega in DCT regulatorja.

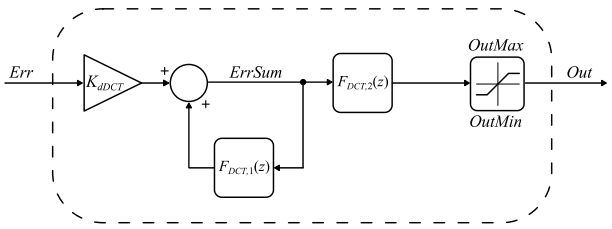
Zaradi enostavnega parametriranja, manjših težav pri zagotavljanju stabilnosti, dobre dinamike in zaradi neodvisnosti med harmonskimi komponentami pri parametriranju bi si želeli vedno uporabiti mRES regulator. Žal pa to ni vedno možno, saj je tovrstni regulacijski algoritem realiziran v mikroprocesorju (μC). Posledično je čas, ki je na voljo za izračun regulacijskega algoritma omejen. Z vsakim dodatnim resonančnim regulatorjem pa skupni čas izračuna narašča. Če želimo skupno regulirati H harmonskih komponent (npr. 20), potem moramo uporabiti tudi H resonančnih regulatorjev (tudi 20, po en za vsako harmonsko komponento). Skupno število resonančnih regulatorjev je navzgor omejeno, saj je omejen tudi čas, ki je na voljo za izračun regulacijskega algoritma znotraj vzorčne periode.

Obe omenjeni slabosti (omejeno skupno število harmonskih komponent H in slabše dinamične lastnosti) lahko zaobidemo z dvojnim DCT regulatorjem (dDCT), ki združuje dobre lastnosti mRES in DCT regulatorja. Ta regulator sem zasnoval sam in je noviteta na tem področju. Do sedaj v literaturi ni bilo moč najti nobenega zapisa o tovrstnem regulatorju. Ker gre za nadgradnjo (enojnega) DCT regulatorja sem ga poimenoval dvojni DCT regulator, saj sta uporabljena dva DCT filtra. Predstavlja direktno alternativo mRES regulatorju, saj lahko nastavljamo parametre za vsako posamezno harmonsko komponento neodvisno od drugih harmonskih komponent. Regulirati je možno poljubno število skupnih harmonskih komponent H, podobno kot pri DCT regulatorju, pri čemer je čas izračuna konstanten in neodvisen od H. Dinamične lastnosti dvojnega DCT regulatorja so primerljive s tistimi pri mRES. Seveda je ta regulator kompleksnejši, da je njegova realizacija sploh možna pa mora biti na razpolago še ustrezen digitalni signalni procesor (DSP).

V nadaljevanju je predstavljen dvojni DCT regulator, opisana je njegova uporaba pri regulaciji harmonskih komponent napetosti RLC vezja ter opravljena primerjava z mRES regulatorjem.

2 Dvojni DCT regulator

Shema dvojnega DCT regulatorja je prikazana na sliki 1. DCT filtra, ki prepuščata le izbrane harmonske komponente sta predstavljena z blokoma $F_{DCT,1}$ in $F_{DCT,2}$. Vhodni signal pogreška *Err* je akumuliran v interni pozitivni povratni (integracijski) zanki. Dodana je še omejitev izhoda regulatorja *Out*.



Slika 1: Shematski prikaz delovanja dvojnega DCT regulatorja.

Parametri, s katerimi nastavljamo delovanje dvojnega DCT regulatorja, so:

1. število vzorcev v eni periodi osnovne harmonske komponente N v obsegu $[0,\infty]$,
2. N koeficientov prvega DCT filtra $\mathbf{coeff_1}$, ki je del $F_{DCT,1}$, v obsegu $[-\infty,\infty]$,
3. N koeficientov drugega DCT filtra $\mathbf{coeff_2}$, ki je del $F_{DCT,2}$, v obsegu $[-\infty,\infty]$,
4. ojačenje regulatorja K_{dDCT} v obsegu $[0,\infty]$,
5. limiti izhoda regulatorja *OutMin* in *OutMax*.

Število vzorčnih intervalov v eni periodi reguliranega signala N izračunamo kot razmerje med vzorčno frekvenco (tudi frekvenca izvajanja regulacijskega algoritma) f_S in frekvenco osnovne harmonske komponente f_1 (v podanem primeru N = 400). V eni vzorčni periodi je treba opraviti konvolucijo med N koeficienti filtra $\mathbf{coeff_1}$ oziroma $\mathbf{coeff_2}$ ter signalom **ErrSum** za oba DCT filtra z enačbama

$$Out(n) = \sum_{i=0}^{N-1} coeff_2(i) \cdot ErrSum(n-i)\,, \qquad (1)$$

$$ErrSum(n) = K_{DCT} \cdot Err(n) + \\ + \sum_{i=0}^{N-1} coeff_1(i) \cdot ErrSum(n-1-i)\,. \qquad (2)$$

Posameznih N koeficientov obeh DCT filtrov izračunamo z enačbama

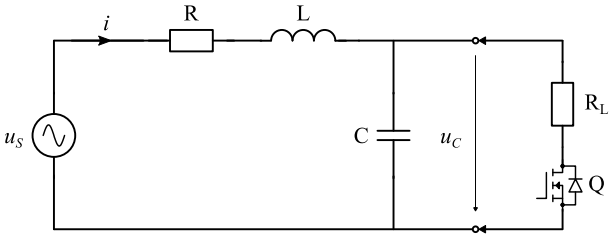
$$coeff_1(i) = \frac{2}{N} \sum_{h=1}^{h_{\max}} A_{h,1} \cdot \cos(2\pi \cdot h \cdot \frac{i}{N})\,, \qquad (3)$$

$$coeff_2(i) = \frac{2}{N} \sum_{h=1}^{h_{\max}} A_{h,2} \cdot \cos(2\pi \cdot h \cdot \frac{i}{N} + \varphi_{comp,h})\,, \qquad (4)$$

kjer s faktorjema korekcije amplitude $A_{h,1}$ in $A_{h,2}$ nastavimo ojačanje h-te harmonske komponente, s kompenzacijo zakasnitve $\varphi_{comp,h}$ pa kompenziramo fazni premik pasivnega filtra (v podanem primeru LC) in dodatne zakasnitve zaradi diskretizacije in modulacije, s čimer zagotovimo stabilno delovanje dvojnega DCT regulatorja. Za doseganje najboljših rezultatov je treba nastaviti pravilne vrednosti kompenzacije zakasnitve $\varphi_{comp,h}$, ojačanja K_{dDCT} in faktorja korekcije amplitude $A_{h,1}$ in $A_{h,2}$. To parametriranje predstavlja pri regulaciji harmonskih komponent največji izziv.

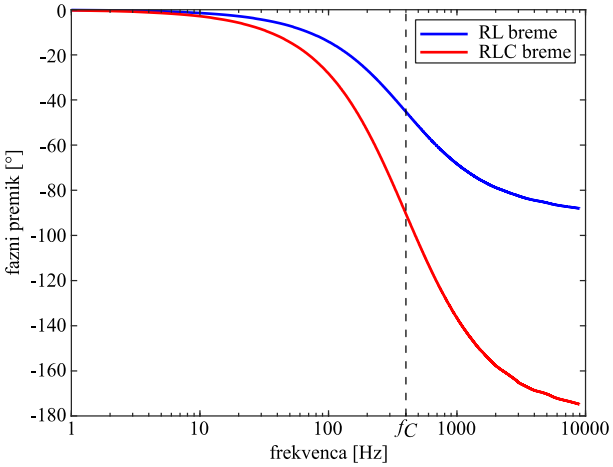
3 Testiranje dvojnega DCT regulatorja – eksperimentalni rezultati

Dobre lastnosti dvojnega DCT regulatorja pridejo do izraza, ko je treba regulirati večje število harmonskih komponent H, prav tako pa se zahteva še visoka dinamika. Delovanje tega regulatorja je preizkušeno na RLC vezju (slika 2), ki ponazarja pasivni filter v močnostnih pretvornikih. Modeliramo ga lahko kot člen drugega reda s faktorjem ojačanja K, časovno konstanto T (oziroma mejno frekvenco f_C) in faktorjem dušenja z. V močnostnih pretvornikih mora biti zaradi izgub nadomestna serijska upornost R (najpogosteje gre za parazitno upornost dušilke) čim manjša. Z naraščajočo močjo se R še dodatno zmanjšuje, zmanjšuje pa se tudi faktor dušenja z, kar povzroča težave z zagotavljanjem stabilnosti. Za preizkušanje dvojnega DCT regulatorja je izbrano vezje s slike 2, s parametri: $R = 257\ \Omega$; $L = 47$ mH; $C = 3{,}3\ \mu$F; $K = 1$; $T = 400\ \mu$s; $f_C = 1/T = 398$ Hz; $z = 1{,}08$; $R_L = 1\ k\Omega$.



Slika 2: Shema RLC vezja z možnostjo obremenitve napetosti kondenzatorja u_C.

Fazna frekvenčna karakteristika RLC vezja je prikazana na sliki 3, skupaj s fazno karakteristiko primerljivega RL vezja. Opazimo lahko bolj zakrivljeno karakteristiko faznega premika RLC vezja, ki se asimptotično približuje -180°. Fazni premik RL vezja dosega le -90°. Zaradi tega nastopijo težave pri zagotavljanju stabilnosti regulacijskega sistema z RLC vezjem oziroma regulacijskih sistemov višjih redov. Brez ustrezno nastavljene kompenzacije zakasnitve $\varphi_{comp,h}$ je v tovrstnih sistemih regulacija zagotovo nestabilna. Pri mRES in dDCT je možno nastaviti kompenzacijo zakasnitve za vsako harmonsko komponento posebej, kar je prav tako nujno, saj fazne karakteristike sistema niso linearne.



Slika 3: Primerjava fazne karakteristike sistema RLC vezja in RL vezja.

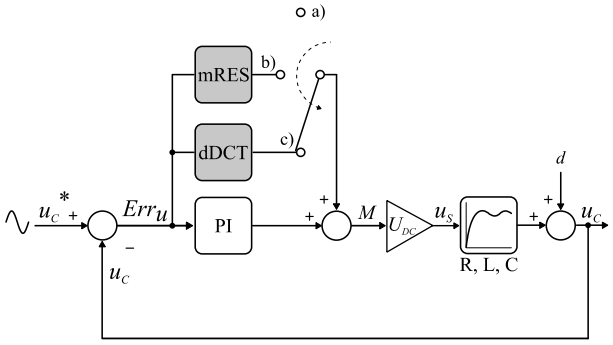
S pomočjo pulzno-širinske modulacije (PWM) je možno nastavljati srednjo vrednost napetosti u_S na izhodnem digitalnem pinu mikroprocesorja TMS32F28069, na katerega je priključeno RLC vezje. Z ustreznim nastavljanjem vklopnega razmerja in posledično napetosti u_S je regulirana napetost kondenzatorja u_C. Želena vrednost izhodne napetosti je sestavljena iz enosmerne komponente in izmenične komponente

$$u_C^* = U_{C,avg}^* + U_{C,amp}^* \cdot \sin(2\pi \cdot f_1 \cdot t)\,, \qquad (5)$$

kjer je frekvenca $f_1 = 50$ Hz, srednja vrednost $U^*_{C,avg}$ je 1,5 V, amplituda izmenične komponente $U^*_{C,amp}$ pa 1 V. Napetost u_S lahko zavzema vrednosti med 0 V in $U_{DC} = 3{,}3$ V (napajanje mikroprocesorja).

Vsako periodo $T_1 = 1/f_1 = 20$ ms je napetost u_C obremenjena z uporom R_L, za čas trajanja 1 ms (tranzistor Q prevaja). To povzroči periodično motnjo d, ki se odraža kot prisotnost višjeharmonskih komponent v napetosti u_C. Višjeharmonske komponente so seveda neželene in jih želimo odpraviti, v skladu z enačbo (5).

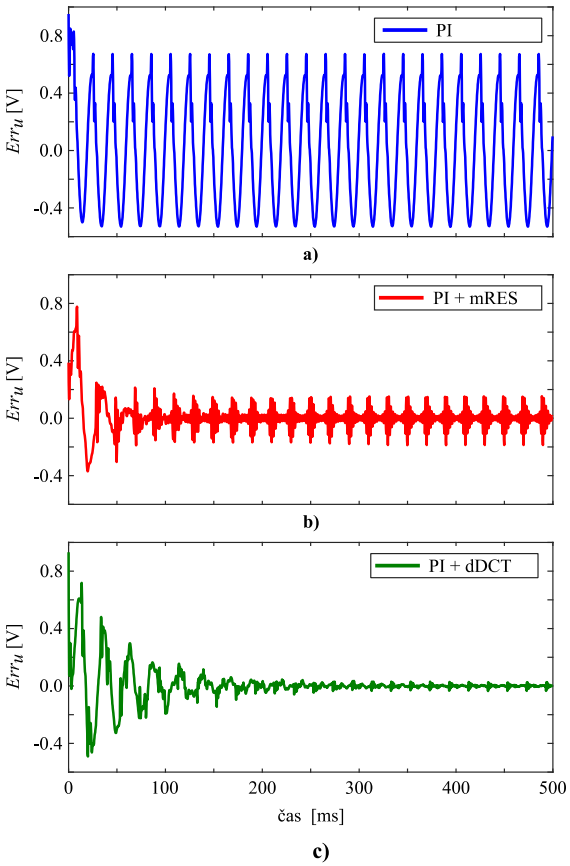
Slika 4 prikazuje regulacijski algoritem za sledenje želeni vrednosti napetosti u_C. Osnovni regulator napetosti kondenzatorja u_C je PI regulator, ki pa ne zmore odpraviti periodičnih pogreškov. Zato z namensko regulacijo harmonskih komponent (bodisi z mRES ali z dDCT regulatorjem) pogrešek v stacionarnem stanju znatno zadušimo.



Slika 4: Shema regulacijskega algoritma za regulacijo napetosti kondenzatorja u_C.

Zaradi časovnih omejitev lahko znotraj uporabljenega mikroprocesorja TMS32F28069 realiziramo največ osem resonančnih regulatorjev in reguliramo lahko prav toliko harmonskih komponent. Z mRES regulatorjem torej reguliramo prvih 8 harmonskih komponent (do vključno 400 Hz). Z dvojnim DCT regulatorjem reguliramo prvih 20 harmonskih komponent (do vključno 1000 Hz). Poudariti velja, da je naenkrat lahko aktiven le en regulator (bodisi PI, bodisi PI + mRES, bodisi PI + dDCT).

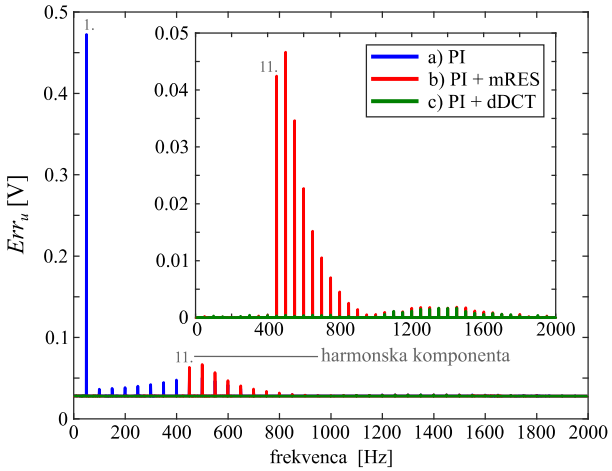
Na sliki 5 je prikazan časovni potek pogreška napetosti Err_u ob vklopu sistema. Napetost kondenzatorja pred vklopnim manevrom je enaka 0. Sledi prehodni pojav z največjim pogreškom napetosti tik po preklopnem manevru. Ta postopoma izzveni, sistem pa v nekaj periodah preide v stacionarno stanje. Ko je uporabljen le PI regulator, maksimalni pogrešek napetosti Err_u v stacionarnem stanju dosega približno 0,7 V. V primeru, ko je PI regulatorju vzporedno priključen še mRES regulator, se maksimalni pogrešek zmanjša na manj kot 0,3 V. Uporaba dvojnega DCT regulatorja valovitost pogreška zaduši do zanemarljive vrednosti 0,02 V.



Slika 5: Prehodni pojav pogreška napetosti Err_u ob vklopu za tri različne implementacije regulatorjev.

Slika 6 prikazuje frekvenčni spekter pogreška napetosti Err_u v stacionarnem stanju. PI regulator ne zmore regulirati osnovne harmonske komponente napetosti brez znatnega pogreška, prav tako ne zmore popolnoma zadušiti višjeharmonskih komponent. Vzporedna vezava resonančnih regulatorjev odpravi pogrešek pri osnovni

harmonski komponenti in pri višjeharmonskih komponentah do 400 Hz. Z uporabo dDCT regulatorja zadušimo višjeharmonske komponente do 1000 Hz. Amplitude nereguliranih harmonskih komponent med 1 kHz in 2 kHz so manjše od 5 mV in so zato zanemarljive.



Slika 6: Frekvenčni spekter pogreška napetosti Err_u v stacionarnem stanju za tri različne implementacije regulatorjev.

V tabeli 1 so prikazani faktorji harmonskega popačenja. Najboljši THD dosega dDCT regulator, najslabšega pa seveda samostojni PI regulator. Ker mRES regulator regulira le 8 harmonskih komponent, očitnega izboljšanja faktorja THD ni.

Tabela 1. Izmerjeno popačenje napetosti u_C (THD) za različne implementacije regulatorjev.

Regulator	THD [%]
PI	10,04
PI + mRES	7,85
PI + dDCT	0,48

Tabela 2. Primerjava časa izračuna za različne implementacije regulatorjev.

Regulator	Čas izračuna [µs]	Obremenitev µC [%]
PI	0,8	1,5
PI + RES	5,7	11,4
PI + mRES (8x)	39,0	78,0
PI + dDCT	35,2	64,0

Ker je frekvenca izvajanja algoritma $f_S = 20$ kHz, ima mikroprocesor za izračun regulacijskega algoritma na voljo $T_S = 50$ µs. Implementacija PI regulatorja ni problematična, saj ta zahteva le 1,5 % razpoložljivega časa (tabela 2). Tudi implementacija enega resonančnega regulatorja ne bi bila problematična, saj zahteva 11,4 % razpoložljivega časa. Težava nastopi, ko je resonančnih regulatorjev več. Zaradi prevelike časovne zahtevnosti 9 resonančnih regulatorjev ni možno realizirati, kar je bilo tudi preizkušeno. V tem primeru se zgodi, da v nekaterih trenutkih mikroprocesor celo preseže omejitev 50 µs, kar seveda ni dovoljeno, zato mikroprocesor zaustavi izvajanje regulacijskega algoritma. Tako se izkaže, da se s časovnega vidika splača uporabiti dDCT regulator za primere, ko želimo regulirati več kot 7 harmonskih komponent. Ne splača pa se dDCT regulatorja uporabiti za regulacijo le nekaj harmonskih komponent.

4 Zaključek

Novo zasnovani dvojni DCT regulator predstavlja alternativo vzporedni vezavi resonančnih regulatorjev, saj lahko regulira večje število harmonskih komponent, pri čemer je čas izračuna vedno enak. Za enostavne sisteme je kompleksnost tega regulatorja prevelika, zato je njegova uporaba smiselna v sistemih višjih redov, kjer postane prva izbira, saj začnejo prevladovati njegove dobre lastnosti. Zaradi vedno pogostejše uporabe LC, LCL in CLC filtrov v močnostni elektroniki, lahko za regulacijo harmonskih komponent pričakujemo porast uporabe dvojnega DCT regulatorja. Nadaljnje delo predvideva preučevanje lastnosti tega regulatorja in njegovo parametriranje ter izvedba dodatnega testiranja.

Literatura

[1] D. Sušin, M. Nemec, V. Ambrožič, in D. Nedeljković, „Single Phase Active Rectifier with Low Harmonic Content in Input Current“, v *2018 26th Mediterranean Conference on Control and Automation (MED)*, Zadar, 2018, str. 302–307.

[2] S. Buso in P. Mattavelli, „Digital Control in Power Electronics“, *Synth. Lect. Power Electron.*, let. 1, št. 1, str. 1–158, jan. 2006, doi: 10.2200/S00047ED1V01Y200609PEL002.

[3] A. Lidozzi, L. Solero, S. Bifaretti, in F. Crescimbini, „Sinusoidal Voltage Shaping of Inverter Equipped Stand-Alone Generating Units“, *IEEE Trans. Ind. Electron.*, str. 1–1, 2014.

[4] P. Mattavelli in F. P. Marafao, „Repetitive-Based Control for Selective Harmonic Compensation in Active Power Filters“, *IEEE Trans. Ind. Electron.*, let. 51, št. 5, str. 1018–1024, okt. 2004.

[5] C. Lorenzini, J. V. Flores, L. F. A. Pereira, in L. A. Pereira, „Resonant–repetitive controller with phase correction applied to uninterruptible power supplies“, *Control Eng. Pract.*, let. 77, str. 118–126, avg. 2018.

[6] R. N. Beres, X. Wang, M. Liserre, F. Blaabjerg, in C. L. Bak, „A Review of Passive Power Filters for Three-Phase Grid-Connected Voltage-Source Converters“, *IEEE J. Emerg. Sel. Top. Power Electron.*, let. 4, št. 1, str. 54–69, mar. 2016.

Optimizacija izgub električnega pogona z asinhronskim motorjem

Žiga Fabijan, Rastko Fišer, Henrik Lavrič, Klemen Drobnič

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: ziga.fabijan@gmail.com

Optimization of losses in induction motor drive

Abstract. *Modern induction motor drive contains a frequency converter, which enables an adjustment of the rotational speed and the amplitude of the input voltage. An open-loop control method (U/f) is typically employed in low-cost drives, but there remains a problem of an excessive magnetization under light loads. In this paper, we have developed a procedure to calculate losses in a frequency converter and induction motor. An optimal magnetization, which guarantees the best possible efficiency for a given operational point of the drive, is found by adjusting the magnetizing flux linkage at different rotational speeds and torques. A clear benefit of using optimal magnetization instead of constant one is shown by analyzing annual electricity consumption for a 2.2 kW water pump drive.*

Keywords: induction motor, frequency converter, *U/f* method, field weakening, pump, Matlab

1 Uvod

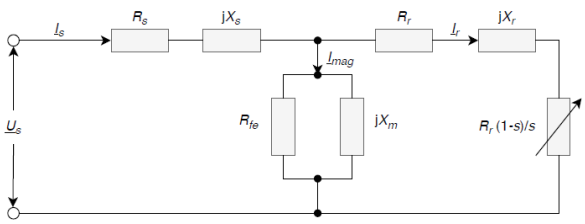
Asinhronski motor s frekvenčnim pretvornikom navadno krmilimo v stacionarnem načinu s skalarno *U/f* metodo, kjer je razmerje med statorsko napetostjo in frekvenco motorja običajno konstantno. Takšne izvedbe električnih pogonov se uporablja v manj zahtevnih napravah nižjih moči, kjer izkoristek ni tako pomemben. Vendar tovrsten koncept vodenja ni vezan le na naprave nižjih moči, temveč je pogost tudi pri napravah večjih moči, kot so razni ventilatorji, kompresorji in črpalke, kjer se izkorišča večja robustnost asinhronskih motorjev v primerjavi z ostalimi izvedbami pogonov.

Na temo optimizacije izgub električnega pogona z asinhronskim motorjem (EPAM) je bilo opravljenih že veliko raziskav. Pri skalarni metodi vodenja motorja je bilo opaziti pomanjkanje analiz, ki bi se osredotočile na problematiko premagnetenja pri nižjih navorih od nazivnega. V članku se bomo zato posvetili optimizaciji izgub EPAM s skalarno metodo glede na optimalni statorski magnetni pretok pri nižjih navorih od nazivnega.

2 Modeliranje izgub

2.1 Model asinhronskega motorja

Numerični model asinhronskega motorja smo izdelali v programu Matlab na podlagi enofazne nadomestne sheme asinhronskega motorja v stacionarnem stanju [1], ki ga prikazuje slika 1. Pri tem so rotorski parametri reducirani na statorsko stran.



Slika 1. Enofazna nadomestna shema asinhronskega motorja.

Izdelani numerični model motorja pri osnovni frekvenci upošteva izgube v statorskem in rotorskem navitju motorja, izgube v železu (histerezne in vrtinčne izgube) ter mehanske izgube (trenje in ventilacija), ki so predstavljene z enačbami (1)-(4) [2]. V modelu nismo upoštevali preostalih izgub, med katere se uvrščajo izgube zaradi višjih harmonskih komponent.

$$P_{Cu,s}(\underline{I}_s) = m \cdot R_s \cdot \underline{I}_s^2 \tag{1}$$

$$P_{Cu,r}(\underline{I}_r) = m \cdot R_r \cdot \underline{I}_r^2 \tag{2}$$

$$P_{core}\left(s, \underline{\psi}_m, f_s\right) = \left(1 + s \cdot \frac{m_r}{m_s}\right) \cdot k_{h,s} \cdot \underline{\psi}_m^v \cdot f_s + \left(1 + s^2 \cdot \frac{m_r}{m_s}\right) \cdot k_{e,s} \cdot \underline{\psi}_m^2 \cdot f_s^2 \tag{3}$$

$$\mathrm{P_{meh}(n)} = \mathrm{M_{dry}} \cdot \left(\frac{\pi}{30}\right) \cdot \mathrm{n} + \mathrm{B} \cdot \left(\frac{\pi}{30}\right)^2 \cdot \mathrm{n}^2 + k_{vent} \cdot \left(\frac{\pi}{30}\right)^3 \cdot n^3 \tag{4}$$

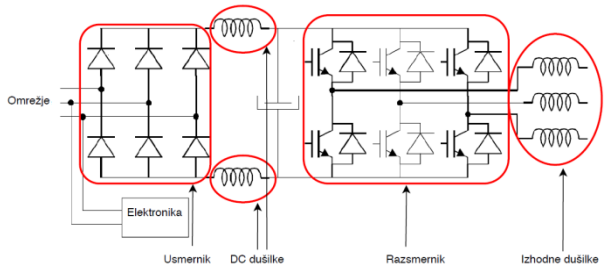
V predstavljenih enačbah predstavlja m število faz motorja, $\frac{m_r}{m_s}$ razmerje med maso rotorja in statorja motorja, $k_{h,s}$ konstanto histereznih izgub v statorju (W), $\underline{\psi}_m$ magnetni sklep, v magnetilni koeficient ter $k_{e,s}$ konstanto vrtinčnih izgub v statorju (W). Konstanti $k_{h,s}$ in $k_{e,s}$ sta pogojeni z materialom in konstrukcijo. M_{dry} predstavlja navor zaradi trenja v suhem okolju, B viskoznostni koeficient ter k_{vent} koeficient ventilacije.

V modelu motorja smo upoštevali spreminjanje glavne induktivnosti L_m od magnetilnega toka $\underline{I}_{mag}$, ki smo ga izračunali kot razliko med statorskim in rotorskim tokom, kar prikazuje enačba (5):

$$\underline{I}_{mag} = \underline{I}_s - \underline{I}_r \tag{5}$$

2.2 Model frekvenčnega pretvornika

Glede na izbrani model frekvenčnega pretvornika, ki ga prikazuje slika 2, smo razvili numerični model frekvenčnega pretvornika [2].



Slika 2. Model frekvenčnega pretvornika.

V modelu frekvenčnega pretvornika smo upoštevali sledeče izgube:

- prevodne izgube v usmerniku (6),
- izgube zaradi napajanja elektronike,
- prevodne izgube dušilke v enosmernem DC linku in izhodnih dušilkah (7) in (8),
- prevodne izgube v razsmerniku (9),
- preklopne izgube v razsmerniku (10).

$$P_{izg,usm} = 2 \cdot U_d \cdot \frac{P_{usm,izh}}{U_{dc}} \tag{6}$$

$$\mathrm{P_{izg,DC\text{-}dus}} = \mathrm{R_{DC\text{-}dus}} \cdot \left(\frac{\mathrm{P_{raz,vh}}}{\mathrm{U_{dc}}}\right)^2 \tag{7}$$

$$\mathrm{P_{izg,izh\text{-}dus}} = \mathrm{m} \cdot \mathrm{R_{izh\text{-}dus}} \cdot \underline{\mathrm{I}}_s^2 \tag{8}$$

$$P_{raz} = f(I_s, \varphi, m_i) \tag{9}$$

$$P_{sw,lin} = K_{sw} \cdot \underline{I}_s \cdot f_{sw} \tag{10}$$

V enačbah predstavlja m_i modulacijski indeks, φ kot med delovno in navidezno komponento moči ter K_{sw} empirično določeno konstanto preklapljanja tranzistorjev frekvenčnega pretvornika.

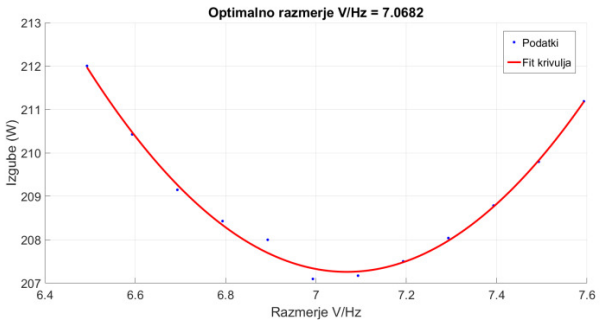
3 Določitev optimalnega magnetenja

Skalarni način vodenja EPAM z *U/f* metodo temelji na konstantnem statorskem magnetnem sklepu $\underline{\psi}_s$, ki je neposredno odvisen od razmerja statorske napetosti $\underline{U}_s$ in statorske frekvence f_s (11).

$$\underline{\psi}_s = \frac{\underline{U}_s}{f_s} \tag{11}$$

Enačbe (1)-(10) sestavljajo nelinearen sistem enačb, ki je za poljubno delovno točko (n, M) in napajalne pogoje (U, f) rešljiv z ustreznim numeričnim postopkom. Na ta način izračunamo tako rezultirajoče električne veličine (tokove) kot tudi izgube. Jasno je, da s spreminjanjem napajalnih pogojev pri nespremenjeni delovni točki vplivamo na iznos posameznih izgubnih komponent pogonskega sistema. V principu gre za prerazporeditev izgub iz magnetilne v rotorsko vejo in obratno: z zviševanjem magnetenja naraščajo izgube v železu, z zniževanjem pa izgube v rotorskem bakru.

Kot pomoč pri iskanju optimalnega napajalnega pogoja smo uvedli faktor k (razmerje *U/f*), ki ponazarja razmerje med statorsko napetostjo in frekvenco. Faktor k smo spreminjali od izbrane začetne pa do končne vrednosti s korakom 0,1. Za vsak k smo izračunali vrednost izgub motorja ter nato izrisali krivuljo na podlagi skupnih izgub. Faktor k, ki je zagotovil minimalne izgube (minimum krivulje na sliki 3), je optimalni k_{opt}, saj je bilo takrat izračunanih najmanj izgub v motorju. Primer iskanja za delovno točko (M = 8 Nm ter n = 900 min^{-1}) prikazuje slika 3.



Slika 3. Prikaz iskanja optimalnega k_{opt}.

Tabela 1 predstavlja izračunana optimalna razmerja V/Hz za različne delovne točke (n, M). Iz nje je razvidno, da se vrednost optimalnega razmerja *U/f* znižuje z zniževanjem navora. Z zniževanjem vrtilne hitrosti motorja se razmerje *U/f* rahlo povečuje, saj je potrebno pokriti padec napetosti na statorski upornosti, ki se močneje izrazi pri nižjih vrtilnih hitrosti. V numeričnem modelu EPAM smo namreč poenostavili enačbe tako, da sta vrednosti pritisnjene in inducirane statorske napetosti enaki.

Tabela 1. Izračunana optimalna razmerja *U/f*.

	2 Nm	6 Nm	10 Nm	14 Nm
300 vrt/min	4,6	7,4	8,8	9,6
600 vrt/min	4,0	6,7	8.0	8,7
900 vrt/min	3,7	6,3	7,6	8,3
1200 vrt/min	3,5	6,0	7,3	8,1
1500 vrt/min	3,3	5,7	7,1	7,9

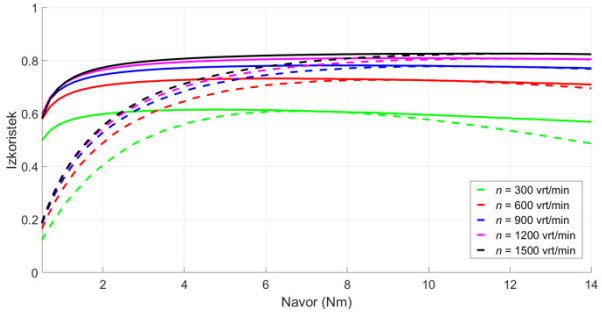
4 Rezultati simulacij in meritev

4.1 Simulacijski rezultati

V tem podpoglavju so predstavljeni simulacijski rezultati izkoristkov EPAM s konstantnim ter optimalnim razmerjem *U/f* za širok razpon obratovalnih točk. Simulirali smo delovanje 2,2 kW motorja ter frekvenčnega pretvornika Danfoss FC 302 [3].

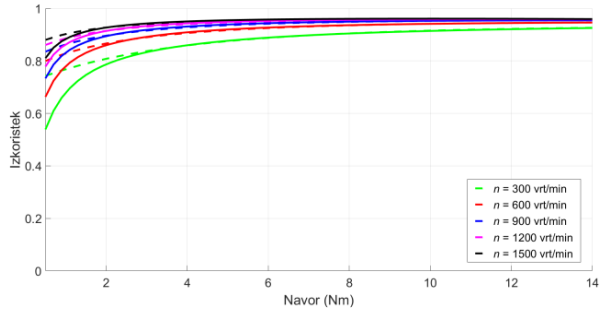
Slika 4 prikazuje simulirane izkoristke asinhronskega motorja pri optimalnem in konstantnem

razmerju *U/f*. Opazimo lahko, da se pri optimizaciji magnetenja motorja izkoristek izboljša predvsem pri nižjih navorih. Pri obremenitvah v bližini nazivnega navora optimalno magnetenje pričakovano takorekoč sovpada z nazivno vrednostjo. Izboljšanje izkoristka opazimo tudi pri nizkih vrtilnih hitrostih in visokih navorih (nad 10 Nm), zaradi izrazitejšega vpliva padca napetosti na statorski upornosti.



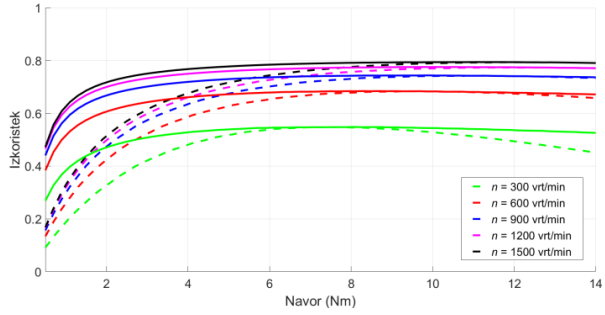
Slika 4. Izkoristki asinhronskega motorja v odvisnosti od navora za nekaj vrtilnih hitrosti: optimalni k_{opt} (polna črta), nazivni k_n (črtkana črta).

Slika 5 prikazuje simulirane izkoristke frekvenčnega pretvornika. Razvidno je, da vrednosti izkoristkov v večini delovnih točk ostajajo nespremenjeni, razen pri nizkih navorih (do 4 Nm). V tem področju so simulirani izkoristki z optimalnim razmerjem *U/f* manjši od nazivnega razmerja *U/f*, saj z zmanjšanjem izgub asinhronskega motorja zmanjšamo tudi izhodno moč iz frekvenčnega pretvornika. Posledično se izgube frekvenčnega pretvornika močneje izrazijo.



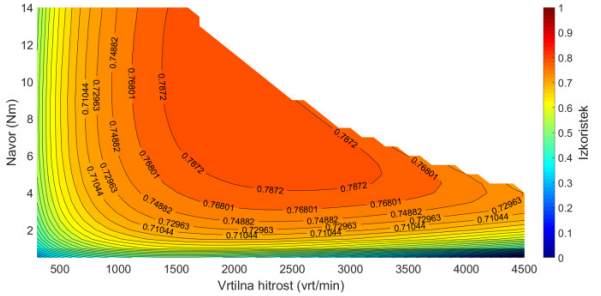
Slika 5. Izkoristki frekvenčnega pretvornika v odvisnosti od navora za nekaj vrtilnih hitrosti: optimalni k_{opt} (polna črta), nazivni k_n (črtkana črta).

Na sliki 6 so predstavljeni izkoristki EPAM, kjer je razvidno, da so poteki simuliranih izkoristkov ter ugotovitve podobne kot pri asinhronskem motorju.



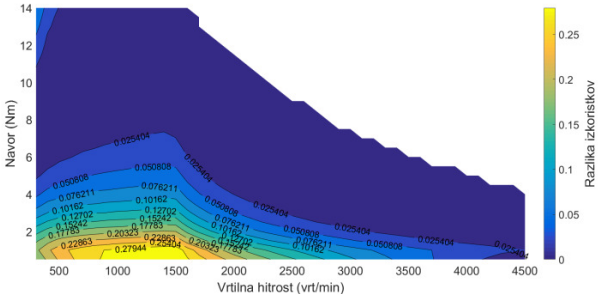
Slika 6. Izkoristki EPAM v odvisnosti od navora za nekaj vrtilnih hitrosti: optimalni k_{opt} (polna črta), nazivni k_n (črtkana črta).

Izkoristek EPAM smo ovrednotili tudi v področju slabljenja polja, kjer je za dosego nadnazivnih vrtilnih hitrosti potrebno magnetno polje že v osnovi znižati. Slika 7 prikazuje vrednosti izkoristkov EPAM pri optimalnem razmerju *U/f*. Pri tem je razvidno, da so najvišji simulirani izkoristki EPAM nahajajo v bližini nazivnih vrednosti motorja. Pri predstavljenih rezultatih je statorski tok omejen na nazivno vrednost.



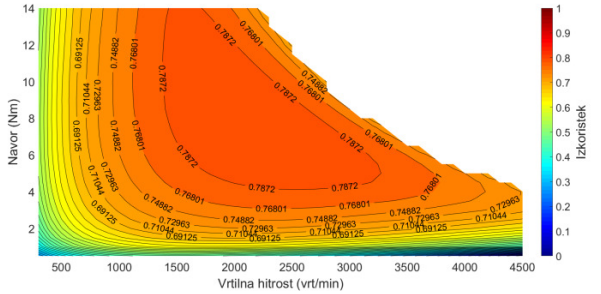
Slika 7. Prikaz izkoristkov EPAM pri optimalnem razmerju *U/f*.

Na sliki 8 je prikazano izboljšanje izkoristkov EPAM z nastavljenim optimalnim razmerjem *U/f*. Izkoristki se v primerjavi s konstantnim razmerjem *U/f* najbolj izboljšajo pri nizkih navorih ter vrtilnih hitrostih motorja do slabljenja polja (do 1500 vrt/min). V področju slabljenja polja je potrebno zmanjševati razmerje *U/f* zaradi napetostne omejitve frekvenčnega pretvornika. Zaradi tega sta razmerji, ki ga še dovoljuje sistem EPAM in optimalno razmerje *U/f* blizu ter je posledično izboljšanje izkoristka manj izrazito.



Slika 8. Razlika izkoristkov EPAM pri vodenju s konstantnim in optimalnim razmerjem *U/f*.

V nadaljevanju smo z možnostjo 25 % povečanja statorskega toka od nazivnega, kar je bilo pogojeno s konceptom delovanja pogona v področju slabljenja polja, razširili obratovalne točke EPAM. Slika 9 prikazuje vrednosti izkoristkov EPAM, kjer je bila dovoljena višja vrednost statorskega toka od nazivne.

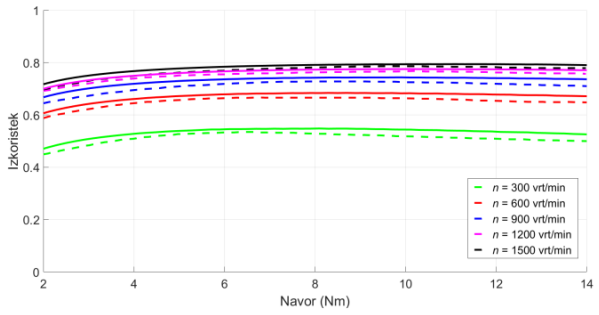


Slika 9. Prikaz izkoristkov EPAM pri dovoljenem višjem statorskem toku od nazivnega.

Takšen način delovanja EPAM je dovoljen za kratkotrajno obratovanje. Razvidno je, da smo pridobili dodatna področja delovanja EPAM, kjer je v področju višje vrednosti statorskega toka od nazivnega vrednost izkoristkov nekoliko nižja. Na sliki 9 niso prikazane vrednosti izkoristkov EPAM pri navorih višjih od nazivnega (14 Nm).

4.2 Primerjava simulacijskega in izmerjenega izkoristka pri optimalnem magnetenju

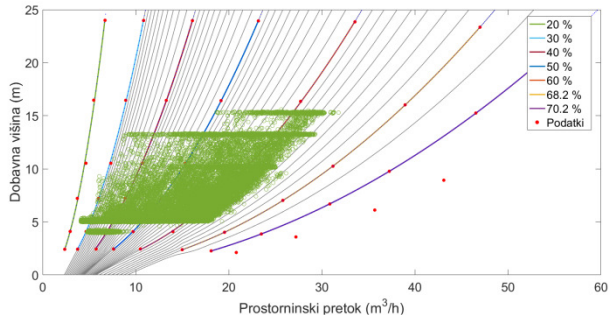
Pri meritvi [2] je bil uporabljen enak asinhronski motor in frekvenčni pretvornik kot v simulaciji. Slika 10 prikazuje izračunane in izmerjene izkoristke EPAM. Rezultata se dobro ujemata, saj znaša največje odstopanje 5 %. Do odstopanja je prišlo predvsem na račun neupoštevanja višjih harmonskih komponent ter deloma zaradi konstantno nastavljene temperature EPAM.



Slika 10. Izkoristki EPAM v odvisnosti od navora pri optimalnem k_{opt}. S polno črto so označeni simulacijski rezultati, s črtkano črto pa izmerjeni rezultati.

5 Analiza električnega pogona s črpalko

Idejo optimalnega magnetenja smo preizkusili na pogonu, ki ga sestavljata EPAM in črpalka – breme z izrazito kvadratno karakteristiko [4]. EPAM smo krmilili s skalarno metodo, pri čemer je bilo v prvem primeru razmerje *U/f* konstantno, v drugem pa optimalno. Na podlagi letne statistike obratovalnih točk (*H*, *Q*) črpalke, ki so na sliki 11 skupaj z njenimi izkoristki predstavljene z zelenimi krogi, smo izračunali vhodno električno moč P_{el}.



Slika 11. Letni cikel obratovanja pogona črpalke.

Pri vsaki obratovalni točki je pogon črpalke deloval 10 minut (1/6 h). Z enačbo (12) smo izračunali letno porabo električne energije pogona črpalke W_{letna} z optimizacijo magnetenja motorja in brez nje ter primerjali rezultate.

$$W_{letna} = \sum_{vzorec=1}^{52560} P_{el}(vzorec) \cdot \frac{1}{6}\, h \qquad (12)$$

Pri enačbi (12) smo z *vzorcem* ponazorili obratovalne točke pogona črpalke. Izračunali smo letno porabo energije pogona črpalke z optimizacijo, kar smo označili z $W_{letna,opt}$ in brez nje $W_{letna,kons}$:

$W_{letna,opt}$ = 5,1355 MWh
$W_{letna,kons}$ = 5,7396 MWh.

Z optimizacijo EPAM bi tako na letni ravni privarčevali okoli 0,6 MWh električne energije, kar predstavlja zmanjšanje porabe električne energije za 10,5 %. Iz tega sklepamo, da bi pri pogonih večjih močih zmanjšanje porabe električne energije pomembno pripomoglo k zmanjšanju stroškov obratovanja.

6 Zaključek

Optimalno magnetenje asinhronskega motorja, ki zagotavlja najmanjše izgube v motorju, hkrati sovpada tudi z minimumom skupnih izgub v celotnem pogonu. Zato v primeru enake stikalne frekvence močnostnega pretvornika zadostuje, da se osredotočimo zgolj na minimiziranje izgub v asinhronskem motorju. Pokazali smo, da z optimizacijo magnetenja motorja v največji meri izboljšamo izkoristke EPAM predvsem pri nizkih obremenitvah. Nasprotno je učinek prilagoditve magnetenja pri nizkih vrtilnih hitrostih (in enakih obremenitvah) bistveno manjši. Praktična vrednost prilagajanja magnetenja se najlepše pokaže v energijski bilanci na dolgi rok. Pri pogonu EPAM s črpalko smo z optimizacijsko metodo zmanjšali letno porabo električne energije za 10,5 %. V primeru pogonov črpalk (ali drugih bremen s kvadratno karakteristiko) večjih moči, so tako absolutni prihranki električne energije lahko zelo veliki.

Literatura

[1] M. Merljak, Zaviranje pretvorniško napajanega asinhronskega motorja, *Magistrsko delo*, Fakulteta za elektrotehniko, Univerza v Ljubljani, 2017.

[2] F. Abrahamsen, Energy Optimal Control of Induction Motor Drives, *Doktorska disertacija*, Institut for Energiteknik, Aalborg Universitet, 2000.

[3] Ž. Fabijan, Optimizacija izgub električnega pogona z asinhronskim motorjem, *Magistrsko delo*, Fakulteta za elektrotehniko, Univerza v Ljubljani, 2020.

[4] H. Lavrič, K. Drobnič, R. Fišer, "Energy Efficiency Assesment of Variable Speed Pump Drive in Industrial Cooling System", *IEEE CPE-POWERENG* 2020 *On-line Proceedings,* Setubal, Portugal, 8-10 July 2020, pp. 261-266.

Valovitost toka v 3-faznem 2-nivojskem razsmerniku

David Kovačič, Rastko Fišer, Klemen Drobnič

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija
E-pošta: dk1550@student.uni-lj.si

Current ripple in 3-phase 2-level inverter

Abstract. *3-phase 2-level voltage-source inverters are a standard choice for generating three-phase voltage from a DC voltage source. Because energy conversion should be as efficient as possible, a modulation technique known as Pulse Width Modulation (PWM) is employed, which generates a switching voltage waveform. However, the switching mode of operation inevitably produces the current ripple, which depends on a load type, switching frequency, and modulation technique and manifests itself in a higher part of the spectra. In this paper, we analyzed the instantaneous current ripple during one fundamental period. Using simulation we determined a relation between the amplitudes and angles of voltage vectors and current ripple in the complex plane. We found that operation with high modulation indices experiences six distinct areas with high current ripple per period.*

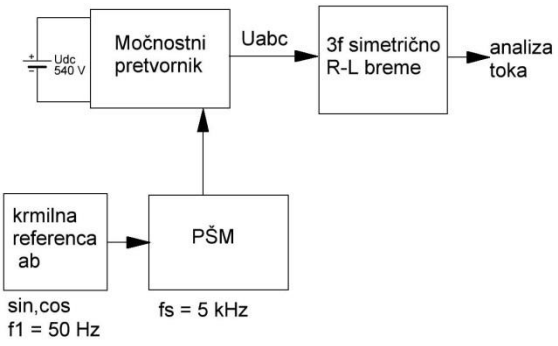
1 Uvod

3-fazni 2-nivojski razsmernik je topologija močnostnega pretvornika, ki jo najdemo v številnih aplikacijah močnostne elektrotehnike, predvsem v električnih pogonih in omrežno povezanih pretvornikih električne energije. V večini tovrstnih pretvornikov vodenje temelji na pulzno-širinski modulaciji (PŠM) z nosilnim signalom, kar zagotavlja enostavno implementacijo, konstantno stikalno frekvenco in dobro definiran frekvenčni spekter. Neizogibna posledica stikalnega načina delovanja je valovitost izhodnega toka, ki naj bo kar najmanjša, saj je prav to neposreden vzrok za nastanek višjeharmonskih izgub na bremenu. S prilagajanjem oblike nosilnega signala lahko vsaj delno vplivamo na velikost tovrstnih izgub, npr. z uporabo nezvezne PŠM pri relativno visokih modulacijskih indeksih [1]. V nadaljevanju privzamemo, da je močnostni pretvornik voden s PŠM s sinusno suboscilacijo, ki ima identični frekvenčni spekter kot sicer dobro znana modulacija s prostorskim vektorjem.

2 Močnostni pretvornik z bremenom

2.1 Generiranje napetosti

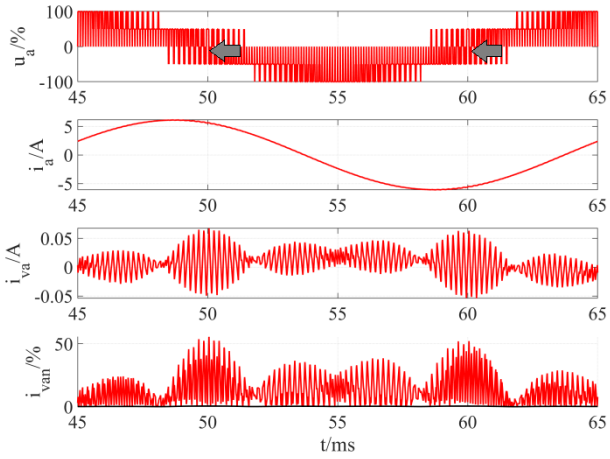
Na sliki 1 je prikazana konceptualna shema simetrično obremenjenega 3-faznega 2-nivojskega razsmernika vodenega s PŠM. Močnostni pretvornik, napajan iz enosmernega vira napetosti, služi za generiranje trifazne izhodne napetosti U_{abc}. Z modulacijskim indeksom, ki je omejen na vrednosti med 0 in 1, izberemo amplitudo krmilne reference, katera se nato primerja s visoko frekvenčnim trikotnim signalom. Ohmsko-induktivno breme je simetrično, nespremenljivo z upornostjo 20 Ω in induktivnostjo 0,15 H na vsaki veji in vezano v zvezdo.



Slika 1. Shema simetrično obremenjenega 3-faznega 2-nivojskega razsmernika

2.2 Analiza valovitosti toka

Zaradi nezveznega poteka napajalne napetosti u_a se v časovnem poteku toka i_a pojavi valovitost i_{va} (slika 2). Največja valovitost toka sovpada z najstrmejšim delom časovnega poteka želene napetosti (tam, kjer krmilni signal prečka ničlo – označeno s kazalcema). Zaradi simetrije faz *a*, *b* in *c* bomo od sedaj naprej obravnavali zgolj fazo *a*.



Slika 2. Primerjava valovitosti toka s potekom napetosti v eni periodi osnovnega harmonika

Iz poteka toka i_a in njegove valovitosti i_{va} (slika 2) vidimo, da tok vsebuje poleg osnovne harmonske

komponente tudi višje harmonike [2]. Frekvenčni spekter na sliki 3 nazorno pokaže frekvenčno vsebino faznega toka – njegov idealni potek bi vseboval zgolj osnovni harmonik, ki zaradi preglednosti ni izrisan. Iz spektra je razvidno, da se najbolj zastopani harmoniki pojavijo v okolici stikalne frekvence (5 kHz) in njenih celoštevilskih večkratnikov.

2.3 Ovrednotenje valovitosti toka

Časovni potek valovitosti i_{va} izločimo iz skupnega toka i_a z

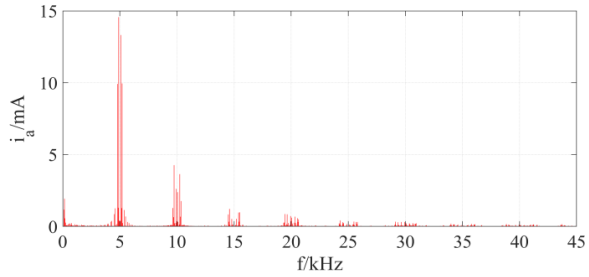
$$i_{va} = i_a - i_{a1}\,, \tag{1}$$

kjer je i_{a1} osnovni harmonik faznega toka. Največjo razliko v vršni-vršni vrednosti valovitosti v eni stikalni periodi definiramo kot

$$I_{vapp} = \max\,(i_{va}) - \min\,(i_{va}). \tag{2}$$

Valovitost i_{va} (slika 2) nato delimo s pravkar definirano vršno-vršno vrednostjo in dobimo normalizirano valovitost

$$i_{van} = \frac{|i_{va}|}{I_{vapp}}. \tag{3}$$



Slika 3. Frekvenčni spekter višjih harmonikov toka i_a

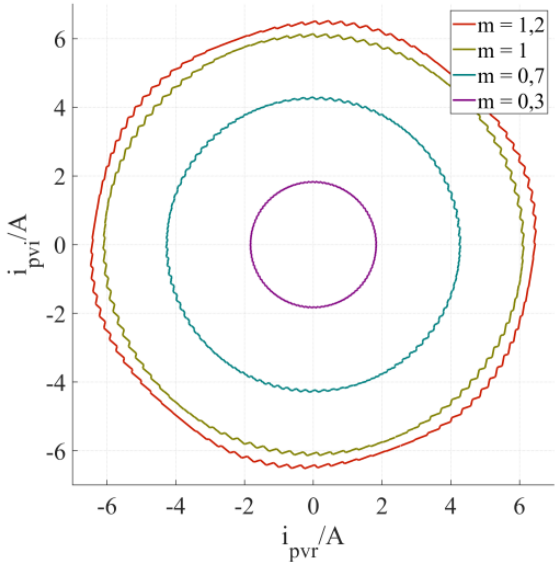
3 Analiza prostorskega vektorja

3.1 Analiza ovojnice prostorskega vektorja v kompleksni ravnini

Trifazni bremenski tok lahko v sistemih brez izvedenega ničelnega vodnika zapišemo s prostorskim vektorjem

$$\boldsymbol{i}_{pv} = \frac{2}{3}(i_a + \boldsymbol{a} i_b + \boldsymbol{a}^2 i_c) = i_{pvr} + \mathrm{j} i_{pvi}, \tag{4}$$

kjer je $\boldsymbol{a} = \exp(\mathrm{j}\cdot 2\pi/3)$. Časovni potek prostorskega vektorja pri posameznem modulacijskem indeksu m razkrije njegovo odvisnost od električnega kota. Na sliki 4 lahko vidimo, kako skupna valovitost faznih tokov vpliva na ovojnico prostorskega vektorja pri različnih vrednostih m. Idealne ovojnice, brez valovitosti toka, bi bile seveda krožnice. Razberemo lahko, da se valovitost spreminja z električnim kotom, in sicer šestkrat v eni osnovni periodi. V področju nadmodulacije ($m > 1$) se dodatno popači osnovna krožna oblika toka, saj se v toku takrat pojavijo harmonske komponente tudi v spodnjem delu spektra (5., 7., 11., 13.,…).

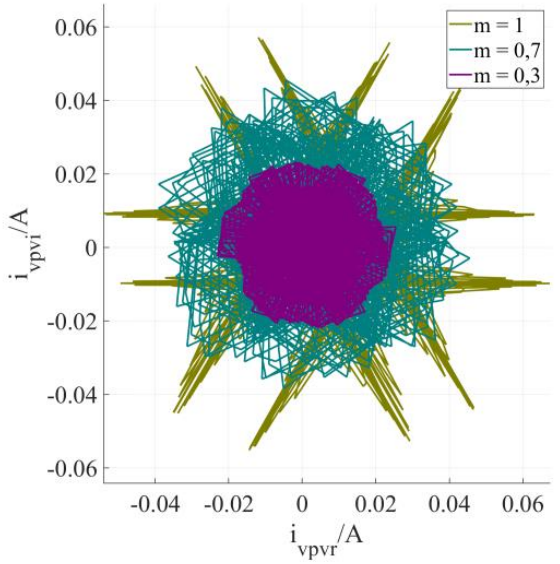


Slika 4. Potek ovojnic prostorskega vektorja toka z valovitostjo pri različnih modulacijskih indeksih

Iz faznih valovitosti sestavimo ovojnico prostorskega vektorja valovitosti

$$\boldsymbol{i}_{vpv} = \frac{2}{3}(i_{va} + \boldsymbol{a} i_{vb} + \boldsymbol{a}^2 i_{vc}) = i_{vpvr} + \mathrm{j} i_{vpvi}\,, \tag{5}$$

ki jo prikazuje slika 5.



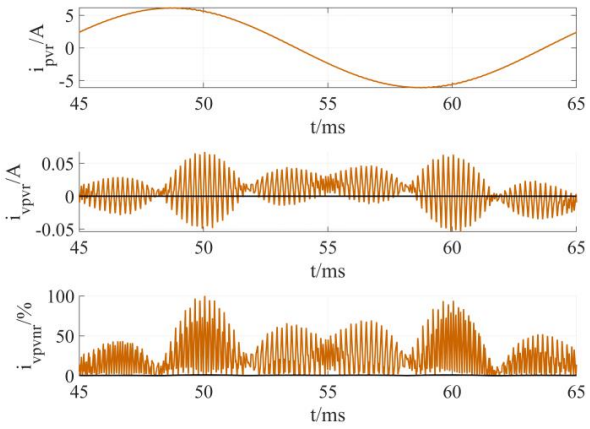
Slika 5. Projekcija valovitosti pri različnih modulacijskih indeksih

Razvidno je, da je valovitost odvisna od električnega kota. Največje izraženosti valovitosti sovpadajo z električnimi koti posameznih stikalnih

vektorjev. Opazimo tudi, da je prisotna podvojenost teh izraženosti ter da valovitosti ovojnice prostorskega vektorja s slike 4 ne sovpadajo z vrhovi stikalnih vektorjev.

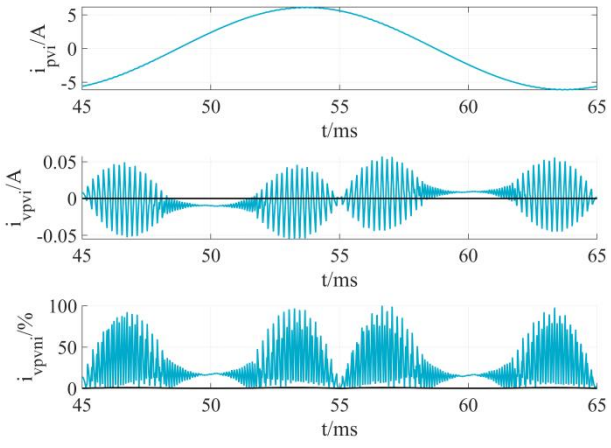
3.2 Analiza ovojnice prostorskega vektorja v kompleksno-časovnem prostoru

V kompleksno-časovnem prostoru obravnavamo imaginarni in realni del signala ločeno. Na sliki 6 lahko opazimo, da je izraženosti toliko, kot je stikalnih vektorjev (šest), in da je prisoten rahel odmik od abscisne osi.



Slika 6. Realna komponenta prostorskega vektorja, njena valovitost, in njena vršno-vršno ovrednotena valovitost

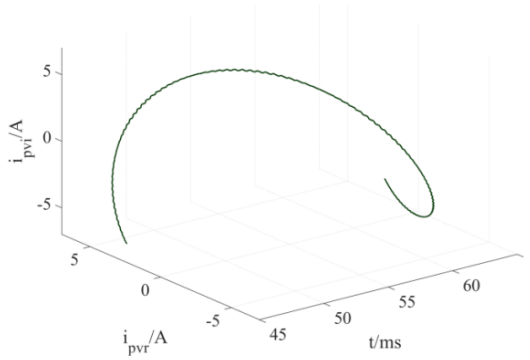
Tudi na sliki 7 lahko iz časovnega poteka valovitosti vidimo manjši odmik valovitosti od abscise, le da je tokrat prisoten pri okoli 50 ms in 60 ms. Zaradi projekcije na imaginarno ravnino so sedaj vidna štiri izražena področja. Preostali dve sta prikriti, ker sta pravokotni na imaginarno ravnino.



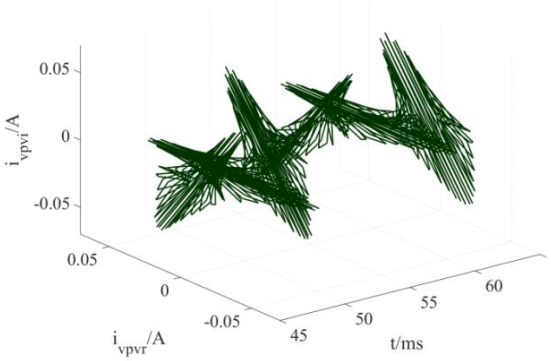
Slika 7. Imaginarna komponenta prostorskega vektorja, njena valovitost in njena vršno-vršno ovrednotena valovitost

3.3 Predstavitev ovojnice prostorskega vektorja v kompleksno-časovnem prostoru

Za boljšo predstavo vpliva enosmernih odmikov, kakor tudi šestih področij izraženosti valovitosti, lahko časovni potek ovojnice prostorskega vektorja i_{pv} in njegove valovitosti i_{vpv} prikažemo v kompleksno-časovnem prostoru (slika 8). S slike 9 lahko razberemo, da bi bile izraženosti na sliki 5 popolnoma poravnane s stikalnimi vektorji, če časovni potek valovitosti ne bi imel odmikov od središčnice. Tako pa pride do podvojenosti.



Slika 8. Časovni potek prostorskega vektorja z valovitostjo za eno periodo

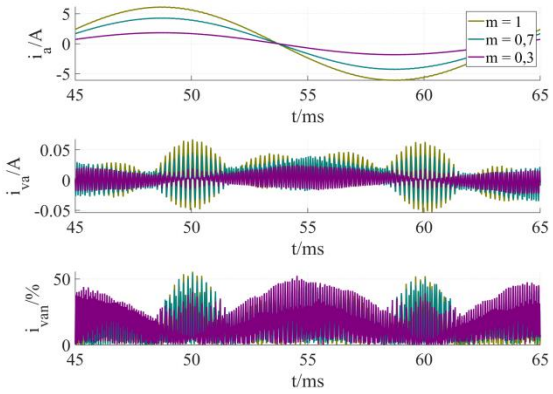


Slika 9. Časovni potek valovitosti v periodi prostorskega vektorja

4 Prikaz odvisnosti valovitosti od modulacijskega indeksa in kota

4.1 Prikaz valovitosti od kota pri posameznih modulacijskih indeksih

Pri spreminjanju modulacijskega indeksa v območju linearne modulacije ($m < 1$) je najopaznejša sprememba v amplitudah toka, kot prikazuje potek i_a na sliki 10 [2].

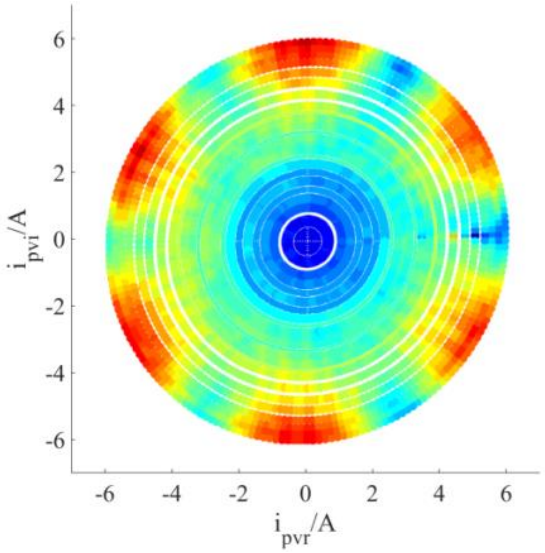


Slika 10. Časovni potek toka i_a, valovitosti i_{va} in njene normirane vrednosti i_{van} pri različnih modulacijskih indeksih

S slike 10 lahko razberemo, da amplitude valovitosti toka i_a (i_{va}) niso linearno odvisne od modulacijskega indeksa. Prav tako je z manjšanjem modulacijskega indeksa šest »izraženosti« vse manj izrazitih. Vpliv spreminjanja modulacijskega indeksa na prostorski vektor prikazuje slika 5, kjer vidimo, da se poveča »izraženost« valovitosti pri linearni modulaciji predvsem izven področja usmeritve stikalnih vektorjev.

4.2 Prikaz odvisnosti valovitosti od modulacijskega indeksa in kota

Z uporabo barvne lestvice (rdeča barva predstavlja največjo valovitost, medtem ko modra predstavlja najmanjšo valovitost) za velikost valovitosti omogočimo »zvezen« prikaz njene odvisnosti od modulacijskega indeksa in kota na sliki 11, ne da bi izgubili informacijo o njeni »izraženosti«.



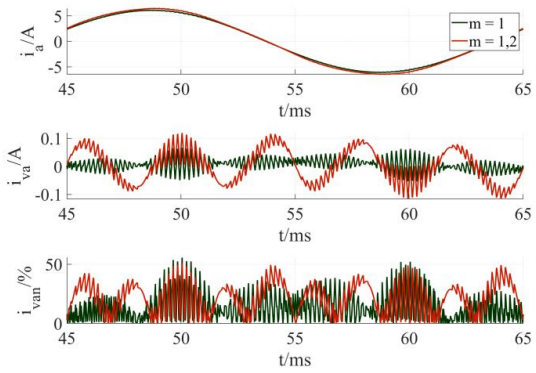
Slika 11. Barvni prikaz odvisnosti valovitosti od modulacijskega indeksa in kota prostorskega vektorja v linearnem področju modulacije

Iz razporeditve barv opazimo, da največja valovitost pri posameznih modulacijskih indeksih sovpada s sečišči »izraženosti« valovitosti med sabo. Ta sečišča so med seboj zamaknjena za 60° in so pri kotih 30°, 90°, 150°, 210°, 270°, 330° [2].

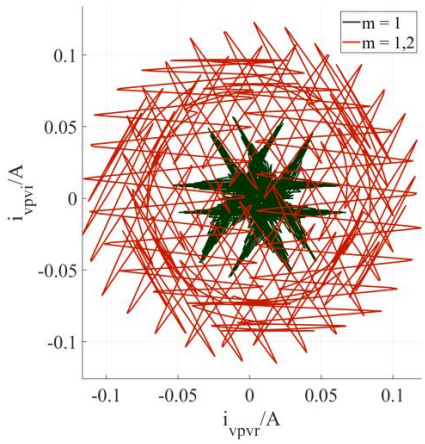
4.3 Nadmodulacija

Pri nadmodulaciji ostane izhod močnostnega pretvornika dlje časa odprt. Posledično pride do nesorazmerno večjih valovitosti v primerjavi z linearno modulacijo. Vpliv nadmodulacije na fazni tok i_a prikazuje slika 12, kjer ima valovitost sedaj prisoten izrazit peti harmonik, okrog katerega oscilirajo preostale komponente valovitosti. Njegova posledica je pretirana »izraženost« valovitosti pri nadmodulaciji, ki močno vpliva na obliko ovojnice prostorskega vektorja, kot je razvidno s slike 4.

S slike 13 lahko razberemo, da je vpliv petega harmonika na valovitost pri modulacijskem indeksu 1,2 bistveno bolj izrazit od vseh ostalih komponent skupaj.



Slika 12. Časovni potek toka i_a, valovitosti i_{va} in njene normirane vrednosti i_{van} pri modulacijskih indeksih 1 in 1,2



Slika 13. Projekciji valovitosti na kompleksno ravnino pri modulacijskih indeksih 1 (zelena) in 1,2 (rdeča)

5 Zaključek

Ugotovili smo, da se valovitost toka ne spreminja zvezno z modulacijskim indeksom, kakor tudi ne zvezno s kotom prostorskega vektorja. Pri nadmodulaciji je to nadvse očitno. Vendar se moramo tudi v področju linearne modulacije zavedati, da navkljub nekolikšnemu zmanjšanju valovitosti v smeri stikalnih vektorjev, pridelamo povečano valovitost na drugih območjih. Problemu pojava valovitosti se s spreminjanjem modulacijskega indeksa torej ne moremo izogniti.

Literatura

[1] Holmes D. G., Lipo T. A., *Pulse Width Modulation for Power Converters*, Wiley-IEEE Press, Hoboken, ZDA, 2003.

[2] Grandi G., Loncarski J., Dordevic O., "Analysis and Comparison of Peak-to-Peak Current Ripple in Two-Level and Multilevel PWM Inverters", *IEEE Trans. Industrial Electronics*, vol. 62, no. 5, May 2015, pp. 2721-2730.

Odpravljanje negativne sofazne napetosti pri unipolarno napajanem odštevalnem vezju z operacijskim ojačevalnikom

Marko Petkovšek, Peter Zajec

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, SI-1000 Ljubljana
E-pošta: marko.petkovsek@fe.uni-lj.si

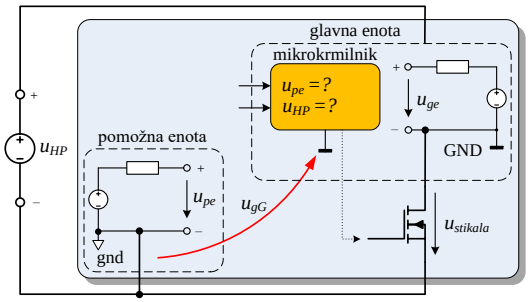
Elimination of a Negative Common-Mode Voltage of a Single-Supplied Operational Ampflifier

Abstract. *A possible approach for voltage measurement in an electronic device with multiple subunits that do not share a same reference (ground) potential is presented in the paper. The analysed method is based on a modified difference operational amplifier (OP amp) circuit with a single supply voltage. This calls for a special attention during the circuit parameter selection phase, since only a positive voltage is allowed not only on the output, but also on either of the OP amp inputs. With the proposed procedure, the common-mode voltage (CMV), present at both inputs of the signal conditioning circuit is kept within the boundaries that are in case of a rail-to-rail (RRIO) OP amp defined with the supply voltage range.*

1 Uvod

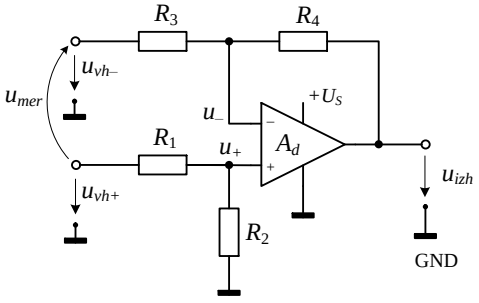
Delovanje sodobnih elektronskih naprav je v veliki meri pogojeno z informacijo o ključnih veličinah, ki jih največkrat zajamemo kar v obliki napetosti. Pri tem je posebno pozornost treba nameniti ustrezni obdelavi merjenega signala, tako zaradi različnih napetostnih nivojev, kot tudi zaradi tega, ker napetosti v kompleksni napravi nimajo skupne referenčne točke. Poenostavljeno shemo take naprave, ki jo sestavljata glavna enota z mikrokrmilniškim vezjem in pomožna enota, kaže slika 1. S slike je razvidno, da enoti nimata skupnega referenčnega potenciala, saj sta njuna negativna priključka povezana preko močnostnega stikala, ki v primeru napake ali nedovoljenih stanj odklopi močnostni del od njegove napajalne napetosti u_{HP}. Taka postavitev stikala (low-side switch [1]), ki je prisotna predvsem pri kompleksnejših topologijah (DC/DC ali DC/AC pretvorniki) v nizkonapetostnih avtomobilskih aplikacijah, ima za posledico, da je v primeru izklopljenega stikala napetostna razlika u_{gG} med enotama lahko precejšnja, predvsem pa je po predznaku negativna. Slednje predstavlja dodaten izziv – sploh v primeru unipolarno napajanega merilnega vezja, ki je posledica (tudi) cenovne optimizacije, predvsem pa uporabljenega mikrokrmilnika s spremljajočo periferijo (A/D in D/A pretvorniki, gonilniki, merilniki toka…).

Poleg nivojske prilagoditve v napravah močnostne elektronike večji izziv predstavljajo sevalne motnje zaradi hitrih preklopov močnostnih polprevodniških stikal .



Slika 1. Shema sodobne elektronske naprave.

Posledica sevalnih motenj na celotno merilno progo od zajema ključnih veličin za delovanje naprave do mikrokrmilnika s krmilnim algoritmom v njem je lahko nestabilno ali pa celo popolnoma nepredvidljivo delovanje naprave, čemur se želimo seveda v čim večji meri izogniti [2]. Prva izbira za merjenje napetosti v sodobnih napravah močnostne elektronike bi torej lahko bilo kar klasično odštevalno (diferenčno) vezje, ki je podano na sliki 2. Poleg preproste zasnove in cenenosti je glavna odlika diferenčnega vezja vezana na visoko stopnjo dušenja sofazne napetosti $u_{CM,OP}$ (vezano na podatek CMRR; common-mode rejection ratio), ki je definirana kot srednja vrednost napetosti na vhodih u_+ in u_-. Ob predpostavki, da je diferenčno ojačenje ojačevalnika zelo veliko ($A_d \to \infty$), postane diferenčna napetost ($u_{DM} = u_+ - u_- \to 0$) enaka 0, to pa pomeni, da je sofazna napetost enaka kar $u_{CM,OP} = u_+$, torej napetosti na + vhodu ojačevalnika. Glede na kataloške podatke splošnih operacijskih ojačevalnikov, pa tudi bolj namenskih diferenčnih ali celo instrumentacijskih ojačevalnikov [3, 4] je razpon sofazne napetosti omejen na območje napajalne napetosti, z nekaj redkimi izjemami, ki dopuščajo prekoračitev območja za maksimalno 0,5 V nad zgornjo (npr. +3,3 V) oz. pod spodnjo (GND) napajalno napetostjo [5, 6].



Slika 2. Odštevalno vezje.

Za zanesljivo delovanje naprave s slike 1 je ključnega pomena pomeriti ne samo napetosti u_{ge} glavne enote, kjer se nahaja tudi mikrokrmilnik in kar lahko storimo z vezjem s slike 2, temveč tudi napetosti pomožne enote u_{pe} in napajalne napetosti močnostnega dela u_{HP} in to v vseh obratovalnih stanjih naprave. A za merjenje napetosti u_{pe} in u_{HP} vezje s slike 2 ni primerno, saj je poleg merjene napetosti (u_{mer}) treba upoštevati še napetostno razliko u_{gG} med obema enotama. V najslabšem primeru lahko u_{gG} zavzame vrednost napetosti u_{HP}, po predznaku pa je negativna ($u_{gG} = -60$ V) in tako prispeva k bistveni spremembi – tudi v področje negativnih vrednosti – sofazne napetosti $u_{CM,OP}$, kar pa ni dovoljeno. V nadaljevanju bo zato analizirano delovanje modificiranega odštevalnega vezja, s katerim je mogoče kompenzirati negativno sofazno napetost na vhodih operacijskega ojačevalnika.

2 Merjenje napetosti z modificiranim odštevalnim vezjem

Negativni sofazni napetosti $u_{CM,OP}$ na vhodih operacijskega ojačevalnika se lahko izognemo z dodatnim virom referenčne napetosti, kot ga kaže slika 3 [7]. Izhodna napetost vezja u_{izh} je definirana kot:

$$u_{izh} = \frac{R_2}{R_1+R_2}\left(1+\frac{R_4}{R_3}\right)u_{vh+} + \\ + \frac{R_1}{R_1+R_2}\left(1+\frac{R_4}{R_3}\right)u_{REF} - \frac{R_4}{R_3}u_{vh-}, \quad (1)$$

kar lahko ob enakosti upornosti $R_1 = R_3$ in $R_2 = R_4$ zapišemo preprosteje:

$$u_{izh} = \frac{R_2}{R_1}(u_{vh+} - u_{vh-}) + u_{REF} = \frac{R_2}{R_1}u_{mer} + u_{REF}. \quad (2)$$

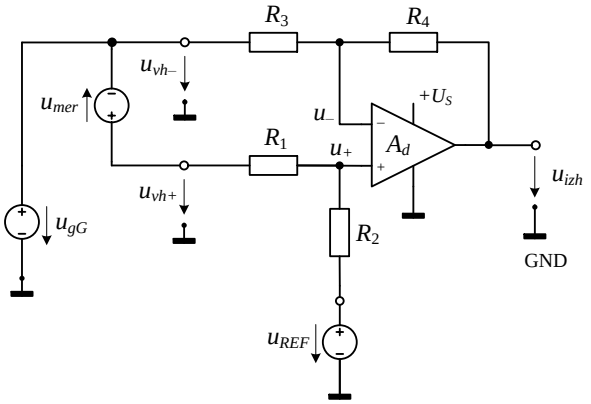
S primerno izbiro ojačenja (razmerje upornosti) in referenčne napetosti u_{REF} v modificiranem vezju lahko torej poskrbimo, da sofazna napetost:

$$u_{CM,OP} = u_+ = \frac{R_2}{R_1+R_2}u_{vh+} + \frac{R_1}{R_1+R_2}u_{REF}, \quad (3)$$

na vhodih operacijskega ojačevalnika v nobenem izmed obratovalnih pogojev naprave (vezano na u_{gG}) ne bo negativna.

V prvem koraku izračunamo vrednost ojačenja G, in sicer tako, da bo za maksimalni razpon $\Delta u_{mer,max}$ merjene napetosti (u_{pe} ali u_{HP}) na vhodu vezja dosežen maksimalni razpon napetosti na izhodu – upoštevajoč seveda maksimalno izhodno napetost $u_{izh,max} = 3{,}3$ V (enako pozitivni napajalni napetosti) in največjo (sicer negativno) vrednost napetosti na vhodu u_{vh-}, na katerega je priključena negativna sponka merjene napetosti, torej $u_{gG,max}$:

$$\frac{R_2}{R_1} = \frac{u_{izh,max}}{\Delta u_{mer,max} - u_{gG,max}} = G. \quad (4)$$



Slika 3. Modificirano odštevalno vezje.

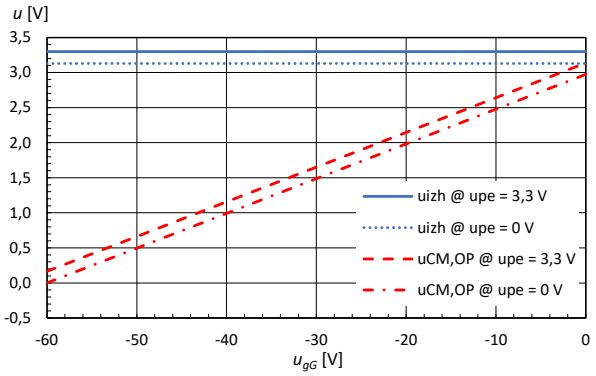
V drugem koraku sledi izračun referenčne napetosti upoštevajoč (2) in pogoja iz (3), da mora biti $u_+ \geq 0$ tudi, ko je merjena napetost enaka 0, hkrati pa je za vhod u_{vh-} ponovno upoštevan najbolj neugoden primer, ko je $u_{gG} = u_{gG,max}$:

$$u_{REF} = -u_{gG,max}\frac{R_2}{R_1}. \quad (5)$$

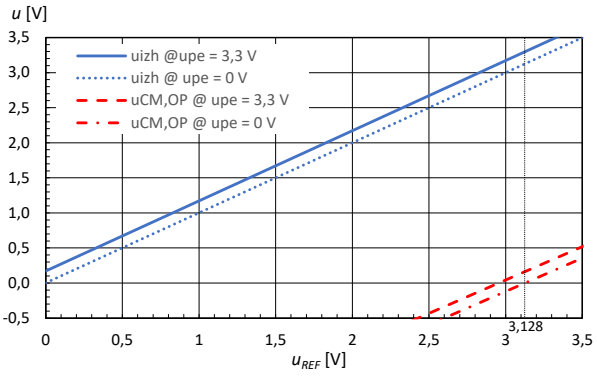
3 Rezultati

V nadaljevanju so podani simulacijski rezultati za primer merjenja napetosti u_{pe} v območju od 0 V do 3,3 V in napetosti u_{HP} v območju od 0 V do 60 V. Posledično je glede na sliko 1 napetostna razlika med potencialoma gnd in GND v najslabšem primeru (izklopljeno stikalo) enaka $u_{gG,max} = -60$ V. S to napetostjo je definiran tudi najbolj negativni potencial (torej – 60 V), ki je v primeru izklopljenega močnostnega stikala pripeljan na vhod vezja za merjenje napetosti in je eden izmed ključnih pogojev za izračun vrednosti elementov merilnega vezja. Pri tem je predpostavljeno, da je napajanje operacijskega ojačevalnika izvedeno z unipolarno napetostjo (+ 3,3 V in GND) in da je OP ojačevalnik tipa RRIO (rail-to-rail input and output), torej, da deluje v skladu s pričakovanji tudi, ko je napetost na njegovem vhodu ali izhodu na obeh skrajnih mejah napajalne napetosti.

Za merjenje napetosti u_{pe} je glede na (4) in (5) izračunana vrednost napetostnega ojačenja enaka $G = 0{,}0521$, referenčne napetosti pa $u_{REF} = 3{,}128$ V. Kaj to pomeni v praksi, je razvidno s slike 4. Namreč, izhodna napetost vezja u_{izh} glede na (2) ne more biti nižja od izračunane referenčne napetosti u_{REF}, saj je v primeru $u_{pe} = 0$ V izhodna napetost $u_{izh} = u_{REF}$, po drugi strani pa je maksimalna izhodna napetost v primeru, ko je $u_{pe} = 3{,}3$ V, omejena z višino napajalne napetosti ($u_{izh} = 3{,}3$ V). Celoten razpon vhodne napetosti $\Delta u_{pe,max} = 3{,}3$ V je torej preslikan na relativno nizek razpon izhodne napetosti $\Delta u_{izh,max} = 0{,}172$ V. Vzrok temu je seveda »obremenjenost« vezja z napetostjo u_{gG}.



Slika 4. Izhodna napetost vezja u_{izh} in sofazna napetost na vhodu ojačevalnika $u_{CM,OP}$ pri merjenju napetosti u_{pe} v odvisnosti od napetosti u_{gG}.

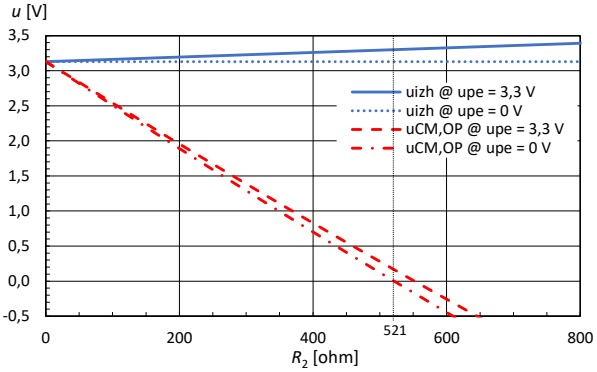


Slika 5. Izhodna napetost vezja u_{izh} in sofazna napetost na vhodu ojačevalnika $u_{CM,OP}$ pri merjenju napetosti u_{pe} v odvisnosti od referenčne napetosti u_{REF}.

Na sliki 4 sta poleg potekov obeh skrajnih vrednosti izhodne napetosti u_{izh} (za u_{pe} pri 3,3 V in 0 V) podana tudi poteka sofazne napetosti $u_{CM,OP}$ na vhodih operacijskega ojačevalnika v odvisnosti od višine napetosti u_{gG}. Razvidno je, da se sofazna napetost na vhodu op. ojačevalnika spreminja glede na višino napetosti u_{gG}, pri čemer je njena minimalna vrednost (0 V) dosežena, ko je $u_{gG} = -60$ V in je merjena napetost $u_{pe} = 0$ V.

Tako višina referenčne napetosti kot vrednost ojačenja G imata seveda velik vpliv na delovanje vezja. Glede na nizek razpon izhodne napetosti s predhodne slike bi lahko hitro (a napačno) sklepali, da lahko z nižjo referenčno napetostjo in višjim ojačenjem razpon izhodne napetosti razširimo. A temu ni tako, kar je razvidno s slike 5, na kateri je za izračunano vrednost napetostnega ojačenja $G = 0{,}0521$ in najneugodnejši primer, ko je $u_{gG} = u_{gG,max} = -60$ V, podan potek izhodne napetosti u_{izh} in napetosti $u_{CM,OP}$ za obe skrajni vrednosti merjene napetosti u_{pe} v odvisnosti od napetosti u_{REF}. Izbrana vrednost u_{REF} mora seveda rezultirati v pozitivni $u_{CM,OP}$ tudi v primeru minimalne vrednosti merjene napetosti $u_{pe} = 0$ V, kar je izpolnjeno pri $u_{REF} \geq 3{,}128$ V (upoštevajoč (5) izračunana vrednost $u_{REF} = 3{,}128$ V je na x osi dodatno označena). Razvidno je, da je pri nižjih vrednostih referenčne napetosti $u_{CM,OP}$ negativna (nedovoljeno!), pri višjih pa pozitivna, hkrati pa pri izbrani vrednosti referenčne napetosti (in ojačenja G) izhodna napetost u_{izh} pri maksimalni vhodni napetosti ($u_{pe} = 3{,}3$ V) ravno doseže vrednost napajalne napetosti. Razlika med potekoma $u_{CM,OP}$ za obe skrajni vrednosti merjene napetosti $u_{mer} = u_{pe}$ je relativno majhna (0,165 V), kar je ob upoštevanju (3) in $u_{vh+} = u_{gG} + u_{mer}$ ter predpostavljene številske podatke seveda logično.

Vpliv ojačenja G oziroma razmerja upornosti na potek izhodne napetosti ter na razmere v vezju je podan na sliki 6. Pri tem je upoštevana (predhodno izračunana) referenčna napetost $u_{REF} = 3{,}128$ V in nejneugodnejši primer, ko je $u_{gG} = u_{gG,max} = -60$ V. Z računsko določenim ojačenjem (4) mora sofazna napetost $u_{CM,OP}$ na vhodih operacijskega ojačevalnika biti pozitivna tudi pri minimalni vhodni merjeni napetosti $u_{pe} = 0$ V. Kot je videti s slike 6, zavzame napetost $u_{CM,OP}$ pozitivne vrednosti za vse vrednosti ojačenj G, ki so manjša od 0,0521, kar v primeru izbrane upornosti $R_1 = 10$ kΩ pomeni za vse upornosti R_2, ki so manjše od 521 Ω.

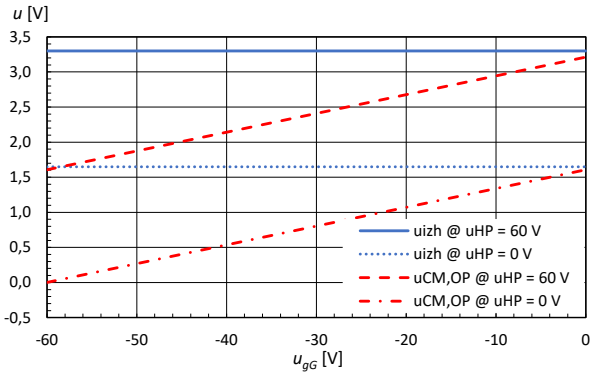


Slika 6. Izhodna napetost vezja u_{izh} in sofazna napetost na vhodu ojačevalnika $u_{CM,OP}$ pri merjenju napetosti u_{pe} v odvisnosti od ojačenja (upornosti R_2).

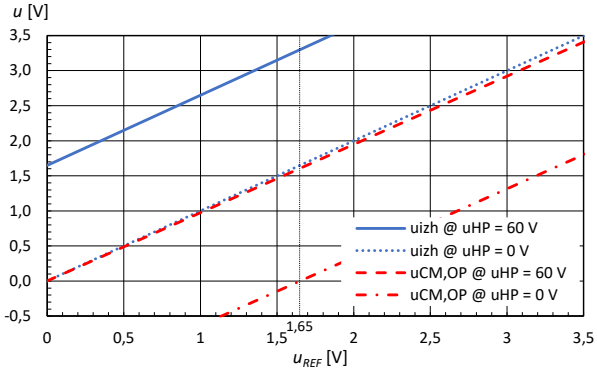
A glede na predhodno podano željo po večjem ojačenju, ki ga lahko dosežemo ravno z večjo upornostjo R_2, morebitno zmanjševanje vrednosti upornosti ni smiselno. Podobno kot prej lahko tudi zdaj ugotovimo, da pri izbrani vrednosti upornosti izhodna napetost u_{izh} pri maksimalni vhodni napetosti ($u_{pe} = 3{,}3$ V) ravno doseže maksimalno vrednost napajalne napetosti.

Kot je bilo prikazano za merjenje napetosti u_{pe}, lahko napravimo tudi za meritev napetosti močnostnega dela u_{HP}. Upoštevajoč (4) in (5) je izračunana vrednost napetostnega ojačenja modificiranega diferenčnega vezja z dodatno referenčno napetostjo enaka $G = 0{,}0275$, referenčna napetost pa $u_{REF} = 1{,}65$ V. Za navedena parametra je na sliki 7 podan potek izhodne napetosti u_{izh} za obe skrajni vrednosti merjene napetosti u_{HP} (za u_{HP} pri 60 V in 0 V) v odvisnosti od napetosti u_{gG}. Kot je razvidno, je razpon izhodne napetosti $\Delta u_{izh,max}$ tokrat absolutno gledano precej večji (1,65 V), saj je glede na (2) pri merjeni napetosti 0 V izhodna napetost enaka $u_{REF} = 1{,}65$ V, pri maksimalni merjeni napetosti $u_{HP} = 60$ V pa zavzame maksimalno vrednost napajalne napetosti (3,3 V). Prav tako je večja (1,606 V) tudi razlika med potekoma sofazne napetosti $u_{CM,OP}$ za obe skrajni vrednosti merjene napetosti.

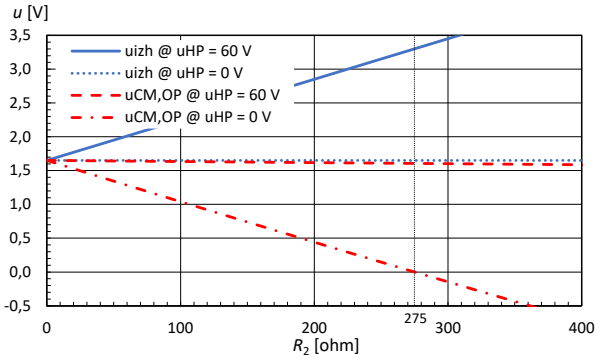
Vpliv referenčne napetosti u_{REF} in ojačenja G na potek izhodne u_{izh} in sofazne napetosti $u_{CM,OP}$ na vhodih op. ojačevalnika pri merjenju napetosti u_{HP} je podan na slikah 8 in 9.



Slika 7. Izhodna napetost vezja u_{izh} in sofazna napetost na vhodu ojačevalnika $u_{CM,OP}$ pri merjenju napetosti u_{HP} v odvisnosti od napetosti u_{gG}.



Slika 8. Izhodna napetost vezja u_{izh} in sofazna napetost na vhodu ojačevalnika $u_{CM,OP}$ pri merjenju napetosti u_{HP} v odvisnosti od referenčne napetosti u_{REF}.



Slika 9. Izhodna napetost vezja u_{izh} in sofazna napetost na vhodu ojačevalnika $u_{CM,OP}$ pri merjenju napetosti u_{HP} v odvisnosti od ojačenja (upornosti R_2).

Če merjeni napetosti u_{pe} in u_{HP} ne bi bili obremenjeni z napetostjo u_{gG} (oz. ko je $u_{gG} = 0$ V), bi sofazna napetost $u_{CM,OP}$ bila vedno pozitivna, zato bi glede na (2) lahko uporabili precej večje ojačenje G ($G_{pe} = 1$ in $G_{HP} = 0{,}055$ – enako za meritev u_{ge}) ter kar klasično diferenčno vezje oz. predlagano z $u_{REF} = 0$ V. Na ta način bi seveda na izhodu v obeh primerih dobili maksimalni razpon $\Delta u_{izh,max} = 3{,}3$ V (namesto 0,172 V oz. 1,65 V) za predviden maksimalni razpon vhodnih napetosti $\Delta u_{pe,max} = 3{,}3$ V in $\Delta u_{HP,max} = 60$ V.

4 Zaključek

V prispevku je osvetljena tematika diferenčnega merjenja napetosti z modificiranim odštevalnim vezjem z unipolarnim napajanjem. S predlaganim postopkom za izračun parametrov vezja je mogoče izvesti meritev napetosti tako, da na vhodih operacijskega ojačevalnika ni prisotne negativne sofazne napetosti. Kot je razvidno iz simulacijskih rezultatov, je delovanje modificiranega odštevalnega vezja odvisno od dveh ključnih parametrov, in sicer ojačenja G in referenčne napetosti u_{REF}. Njun vpliv na delovanje vezja je prikazan na primeru merjenja dveh različno visokih napetosti (3,3 V in 60 V). Ker je pri predlaganem postopku za izračun parametrov vezja treba upoštevati tudi za pojav negativne sofazne napetosti »odgovorno« napetostno razliko u_{gG} med obema enotama s slike 1, je uporaben razpon izhodne napetosti za obe merjeni napetosti precej nižji, kot bi bil v primeru, ko bi napetosti bili referirani proti skupni referenčni točki glavne enote (GND) in bi torej bila napetost $u_{gG} = 0$ V.

Zaradi precej nenavadnih vrednosti izračunanih parametrov – predvsem referenčne napetosti, ki jo je težko realizirati z na trgu (tudi cenovno) dostopnimi komponentami [8] - [10] – je posebno pozornost pri praktični realizaciji merilnega vezja treba nameniti izbiri komponent, a slednje presega tematiko tega članka.

Literatura

[1] Texas Instruments. “Basics of Power Switches,” Application report [Online]. Available: https://www.ti.com/lit/an/slva927a/slva927a.pdf. [June 20, 2020].

[2] B. Deutschmann, G. Winkler, P. Kastner, “Impact Of Electromagnetic Interference On The Functional Safety Of Smart Power Devices For Automotive Applications,” Elektrotechnik & Informationstechnik, vol. 135, no. 4-5, pp 352-359, 2018.

[3] Texas Instruments. “INA117” [Online]. Dosegljivo: https://www.ti.com/lit/ds/symlink/ina117.pdf. [20.6.2020].

[4] Analog Devices. “AD628” [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/AD628.pdf. [20.6.2020].

[5] ON Semiconductor. “NCV5230” [Online]. Dosegljivo: www.onsemi.com/pub/Collateral/NE5230-D.pdf. [20.6.2020].

[6] Texas Instruments. “TLV9001” [Online]. Dosegljivo: http://www.ti.com/lit/ds/symlink/tlv9001.pdf. [20.6.2020].

[7] M. Petkovsek, P. Zajec, “Extending the Input Common-Mode Voltage Range of Single-Supply OP Amps,” Proceeding of the I2MTC conference 2020, May 25-27 2020, Dubrovnik, Croatia.

[8] Texas Instruments. “LM4140 High Precision Low Noise Low Dropout Voltage Reference,” [Online]. Dosegljivo: https://www.ti.com/lit/ds/symlink/lm4140.pdf [20.6.2020].

[9] Analog Devices. Precision Series Sub-Band Gap Voltage Reference [Online]. Dosegljivo: https://www.analog.com/media/en/technical-documentation/data-sheets/ADR130.pdf [30.6.2020].

[10] Y. Liu, C. Zhan, L. Wang, J. Tang, G. Wang, “A 0.4-V Wide Temperature Range All-MOSFET Subthreshold Voltage Reference With 0.027%/V Line Sensitivity,” IEEE Trans. on Circuits and Systems II: Express Briefs, vol. 65, no. 8, pp. 969–973, Aug. 2018.

Merilna tehnika

Measurement

Sledljivost in točnost čelnih termometrov

Igor Pušnik [1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: igor.pusnik@fe.uni-lj.si

Accuracy of forehead thermometers

Abstract
In major part of the world instruments for body temperature measurements (BTM) are not required legally to be traceable to the national or international temperature standards. Special type of thermometers are non-contact thermometers that are used for body temperature measurement. Such measurements could be perfomed by thermal imagers (general or special purpose) or non-contact clinical thermometers (tympanic or ear thermometers, forehead thermometers).

Traceability to national or international standards are prerequisite for accurate measurements with known measurement uncertainty. Another issue is a measurement technology which requires certain knowledge and training of users. Depending on an instrument used the appropriate measurement protocol has to be developed and followed in order to achieve reliable and accurate measurement results with known uncertainty which is inevitably larger than the measurement uncertainty provided by the laboratory calibration of the same instrument.

In the light of recent pandemic Covid-19 it seems that technologies for non-contact body temperature measurement, although relatively simple and easy to use from users perspective, represent a great challenge to temperature experts which have to provide traceability and relevant guidance how to use various instruments in order to provide accurate and reliable results.

The article describes the case of forehead thermometers. In the LMK we studied their performance in terms of accuracy and measurement uncertainty. Results of 5 different types of forehead thermometers, which are used in the University Clinical Center of Ljubljana, were far from specifications given by manufacturers, which claim the accuracy of ±0,2 °C.

1 Uvod

V večini držav na svetu ni zakonskih zahtev za preverjanje sledljivosti instrumentov za merjenje telesne temperature. Posebna so brezkontaktni termometri, ki se uporabljajo za merjenje telesne temperature. Takšne meritve lahko izvajamo s pomočjo termovizijskih kamer (splošnega ali posebnega namena) ali z brezkontaktnimi kliničnimi termometri (timpaničnimi ali ušesnimi termometri, čelnimi termometri). Sledljivost do nacionalnih ali mednarodnih standardov je predpogoj za točnost meritev z znano merilno negotovostjo.

Drugo vprašanje je merilna tehnologija, ki zahteva določeno znanje in usposabljanje uporabnikov. Glede na uporabljeni instrument je treba razviti in upoštevati ustrezen merilni protokol, da bi dosegli zanesljiva in točna merjenja z znanimi negotovostmi, ki so neizogibno večje od merilne negotovosti, ki jo zagotavlja laboratorijska kalibracija istega instrumenta.

Glede na nedavno pandemijo Covid-19 se zdi, da tehnologije za brezkontaktno merjenje telesne temperature, čeprav so razmeroma enostavne in enostavne za uporabo z vidika uporabnikov, predstavljajo velik izziv za temperaturne strokovnjake, ki morajo zagotoviti sledljivost in ustrezne napotke za uporabo različnih instrumentov, da bi zagotovili točne in zanesljive rezultate.

Članek opisuje primer petih različnih čelnih termometrov, ki jih uporabljajo v Univerzitetnem kliničnem centru v Ljubljani, in katerih laboratorijsko točnost ter merilno negotovost smo preverili v LMK. Rezultati še zdaleč niso ustrezali specifikacijam proizvajalcev, ki navajajo točnost reda ± 0,2 ° C.

2 Zagotavljanje sledljivosti meritev

Nacionalni meroslovni sistem je primarnega pomena za vsako državo in pogojuje uspešen razvoj gospodarstva, novih tehnologij, znanosti in razvoja družbe v celoti. Temeljne naloge na tem področju so vzdrževanje merilne sledljivosti nacionalnih etalonov za posamezne veličine mednarodnega sistema merskih enot (SI) na mednarodno raven ter prenos merilne sledljivosti do uporabnikov za potrebe merjenj posameznih sektorjev; kot so znanost, okolje, zdravstvo, prehrana, kmetijstvo, promet, telekomunikacije, industrija in storitve, trgovina, davčna dejavnost, pravosodni organi, policija. Osnovna rezultata tega dela sta predvsem zagotavljanje sledljivih meritev v Republiki Sloveniji ter prenos znanja oziroma strokovna pomoč pri realizaciji najzahtevnejših merjenj za široko področje uporabnikov [1].

Na področju termodinamične temperature je LMK nosilec nacionalnega etalona, ki v Republiki Sloveniji zagotavlja najvišji hierarhični nivo točnosti merjenj za vse vrste termometrov, tudi termometre za merjenje telesne temperature. Celo več, LMK je po rezultatih mednarodnih primerjav med nacionalnimi laboratoriji med najboljšimi v Evropi in svetu, kar potrjujejo njegove merilne zmogljivosti, ki jih za nacionalne laboratorije potrjuje in objavlja Mednarodni urad za mere in uteži BIPM v Parizu [2]. Nacionalni etaloni zagotavljajo sledljivost in mednarodno primerljivost meritev vseh veličin po vsem svetu in v posamezni državi, s tem pa

tudi zaupanje v merilne rezultate med posameznimi državami. Predvsem je pomembno razumeti, da sledljivost meritev ne pomeni zgolj sledljivosti merilnega instrumenta, temveč predvsem sledljivost merilnega rezultata, kar vključuje tudi merilno metodo z vsemi vplivnimi parametri.

Prvi element, ki mora zagotoviti sledljivost postopka merjenja, je merilni instrument, ki mora biti sledljiv do nacionalnega ali mednarodnih etalonov. To lahko zagotovijo samo usposobljeni (nacionalni ali akreditirani) kalibracijski laboratoriji.

Drugi element sledljivosti je postopek merjenja z izvajalcem le-tega. Postopek mora biti preverjen in s tem dovolj skladen s tehničnimi zahtevami, ki zagotavljajo ciljno točnost.

Trenutno imamo po vsem svetu težave s sledljivostjo obeh elementov merilnega sistema za merjenje telesne temperature. Večinoma niso sledljivi niti instrumenti, ki uporabljajo za merjenje telesne temperature (čelni termometri, ušesni termometri, termovizijske kamere), niti so usklajeni merilni protokoli, da bi omogočali točna in zanesljiva merjenja. Z namenom zagotavljanja sledljivosti merjenja telesne temperature po celem svetu je bila junija 2020 ustanovljena posebna delovna skupina v okviru BIPM, ki bo za merjenje telesne temperature analizirala obstoječe merilne protokole po svetu, evidentirala težave in dobre prakse ter predlagala najboljše možne protokole z morebitnimi različicami, ki so potrebni v različnih specifičnih okoliščinah. Predlagani protokoli bodo sčasoma postali del mednarodnih standardov, ki se nanašajo na merjenje telesne temperature.

3 Brezkontaktni merilniki telesne temperature

Za brezkontaktno merjenje telesne temperature uporabljamo ušesne ali timpanične termometre (Slika 1), čelne termometre ter namenske in splošne termovizijske kamere (Slika 2). Ušesni termometri kljub brezkontaktnemu načinu merjenja zahtevajo stik s človekom in celo pol-invazivni postopek merjenja, ker se vrh termometra vstavi v ušesni kanal. Princip delovanja pri vseh termometrih je merjenje sevalne temperature določene površine telesa (timpanična membrana, čelo, senca, očesni kot). Instrumenti zaznavajo toploto v obliki infrardečega sevanja, ki ga na različne načine pretvorijo v odčitek temperature. Prikazana temperatura običajno ni dejansko izmerjena temperatura na merjeni površini, temveč preračunana temperatura v temperaturo drugega dela telesa, kar mora biti jasno navedeno na termometru samem.

Zaradi brezkontaktnega principa merjenja in brez stika, so določene naprave koristne pri merjenju telesne temperature, ko želimo preprečiti oziroma omejiti širjenje nalezljivih bolezni, kar je primer trenutne pandemije Covid-19.



Slika 1: Ušesni termometer



Slika 2: Namenska termovizijska kamera

3.1 Čelni termometri

Čelni termometri se delijo na termometre v obliki traku s tekočimi kristali (Slika 3) in termometre, ki merijo infrardeče sevanje nad senco oziroma na čelu (temporal artery thermometer, Slika 4). Prvi so bili razviti in patentirani leta 1972, vendar zaradi omejitev pri točnosti nikoli niso dosegli širše uporabe [3], slednji pa je bil patentiran leta 1998 [4].



Slika 3: Čelni termometer s tekočimi kristali

Meritev telesne temperature dobimo s skeniranjem senzorja toplotnega sevanja čez del stranskega čela nad temporalno arterijo. Najvišja izmerjena temperatura se upošteva za izračun notranje temperature telesa v odvisnosti od temperature okolice in izmerjene površinske temperature. Funkcija vključuje uteženo razliko temperature površine in okolice, pri čemer se utež spreminja glede na ciljno temperaturo do najmanjše v območju od 35,6 °C do 37,8 °C. Senzor sevanja meri površino pod kapico za kompenzacijo emisivnosti, ki je oddaljena od kože s krožno membrano z nizko toplotno prevodnostjo.



Slika 4: Čelni termometer (temporal artery thermometer)

Ostale vrste čelnih termometrov, ki ne merijo po istem principu kot patentirani termometer za čelno merjenje, so bistveno cenejše, vendar kljub navedbam proizvajalcev glede točnosti, jo le stežka kateri doseže. Večina teh nizkocenovnih termometrov je proizvedena na Kitajskem [5].

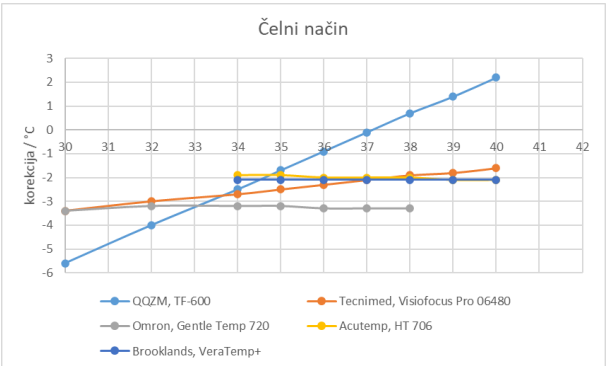
4 Rezultati umerjanja čelnih termometrov

V LMK smo izvedli umerjanje in s tem preverjanje točnosti petih različnih čelnih termometrov, ki se uporabljajo v UKC Ljubljana. Termometri so bili umerjeni v čelnem načinu, štirje pa tudi v površinskem načinu merjenja, kar omogoča bolj zanesljivo preverjanje točnosti detektorja, ki je za oba načina merjenja seveda isti. Kot vir referenčne temperature je bilo uporabljeno črno telo z emisivnostjo 0,998 in razširjeno merilno negotovostjo 0,2 °C.

4.1 Čelni način merjenja

Tabela 1. Korekcije termometrov v čelnem načinu merjenja

T_{ref}	QQZM	Tecnimed	Omron	Acutemp	Brooklands
30	-5,6	-3,4	-3,4		
32	-4	-3	-3,2		
34	-2,5	-2,7	-3,2	-1,9	-2,1
35	-1,7	-2,5	-3,2	-1,9	-2,1
36	-0,9	-2,3	-3,3	-2	-2,1
37	-0,1	-2,1	-3,3	-2	-2,1
38	0,7	-1,9	-3,3	-2	-2,1
39	1,4	-1,8		-2,1	-2,1
40	2,2	-1,6		-2,1	-2,1
41					
$U_{k=2}$	**0,6**	**0,8**	**0,7**	**1,0**	**1,0**

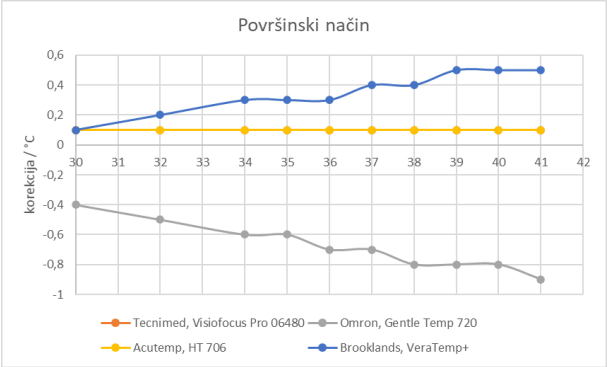


Slika 5: Korekcije termometrov v čelnem načinu merjenja

4.2 Površinski način merjenja

Tabela 2. Korekcije termometrov v površinskem načinu merjenja

T_{ref}	QQZM	Tecnimed	Omron	Acutemp	Brooklands
30		0,1	-0,4	0,1	0,1
32		0,1	-0,5	0,1	0,2
34		0,1	-0,6	0,1	0,3
35		0,1	-0,6	0,1	0,3
36		0,1	-0,7	0,1	0,3
37		0,1	-0,7	0,1	0,4
38		0,1	-0,8	0,1	0,4
39		0,1	-0,8	0,1	0,5
40		0,1	-0,8	0,1	0,5
41		0,1	-0,9	0,1	0,5
$U_{k=2}$		**0,3**	**0,3**	**0,3**	**0,3**



Slika 6: Korekcije termometrov v površinskem načinu merjenja

Kot lahko vidimo, se rezultati umerjanja (korekcije) v ušesnem načinu merjenja pri istih pogojih, kajti umerjanje je bilo izvedeno sočasno, zelo razlikujejo. Hkrati se razlikujejo tudi merilne negotovosti in zanimivo je, da ima najmanjšo negotovost čelni termometer, ki ima daleč največja odstopanja, ki so povrhu vsega še zelo spremenljiva glede na referenčno temperaturo. Žal ta termometer nima površinskega načina merjenja, da bi lahko preverili karakteristiko senzorja. Pri ostalih termometrih je to bilo možno in dva izmed njih sta imela korekcije v celotnem območju merjenja 0,1 °C, kar pomeni, da detektor zaznava stabilno sevanje in temu primerna je konstantna vrednost preračuna površinske temperature. Vsi termometri v površinskem načinu merjenja so imeli negotovost 0,3 °C, kar potrjuje, da je vsak izmed termometrov v tem načinu merjenja bistveno boljši kot v čelnem načinu merjenja.

5 Razprava

Pri merjenju telesne temperature večina naslavlja točnost meritev, ki je navedena v standardu ISO 80601-2-56: 2017 Medicinska električna oprema - Del 2-56: Posebne zahteve za osnovno varnost in bistvene zmogljivosti kliničnih termometrov za merjenje telesne temperature [6], kjer se zahteva točnost merilne naprave v laboratorijskih pogojih ±0,2 °C. Iz izkušenj vemo, da nekateri ušesni termometri in večina čelnih termometrov dosegajo te stopnje točnosti. Da ne omenjam, da klinične

raziskave za večino čelnih termometrov in termovizijskih kamer, ki merijo različne dele obrazne površine (čelo, kateri koli obrazni del, očesni kot), niso bile izvedene ali vsaj ne predstavljene.

Pri merjenju določenega dela obraza in izračunu telesne temperature je treba najprej določiti korelacijo med različnimi merilnimi mesti na ali v telesu. Doslej je ta vidik znanstveno slabo raziskan za korelacijo s čelom. Zato se izračuni v čelnih termometrih razlikujejo za več kot 2 ° C, ki na drugi strani kažejo veliko boljše rezultate v površinskem načinu merjenja. Očitno je glavna težava teh termometrov drugačen pristop pri izračunavanju telesne temperature na podlagi izmerjene temperature na čelu.

Drugo zelo pomembno vprašanje je tudi nastavitev emisivnosti čelnega termometra ali termovizijske kamere. Čeprav velja skupno mnenje, da je emisivnost človeške kože znaša 0,97 do 0,98, je treba pri umerjanju upoštevati razliko emisivnosti v primerjavi s črnim telesom. Namreč razlika emisivnosti 0,03 pri sobni temperaturi 23 °C (temperaturo detektorja upoštevamo enako 23 °C), temperaturi črnega telesa 37 °C in valovni dolžini detektorja 8 µm do 14 µm zahteva korekcijo temperature črnega telesa +0,41 °C. Brez upoštevanja točnih vrednosti v merilnem sistemu je povsem nemogoče pričakovati točnost ±0,2 ° C, ki je v večini primerov zelo pretirana. Da ne omenjam, da skoraj vsi ponudniki namenskih sistemov termovizijskih kamer za merjenje telesne temperature navajajo točnost ±0,3 °C ali bolje. Hkrati pa za te naprave ne nudijo kalibracijskega certifikata, ki bi ga izdal akreditiran kalibracijski laboratorij kot je LMK [7].

6 Zaključek

Na osnovi tehnologije sevalnih termometrov za merjenje temperature na čelu, ki je prisotna že nekaj desetletij in še vedno ne dosega zahtevanih a vendarle navedenih točnosti, imamo veliko težav, ker se večina uporabnikov zanaša na specifikacije proizvajalcev in verjame navedbam o njihovi točnosti. Zaradi tega merjenja telesne temperature ne samo da niso sledljiva, verjetno imamo celo veliko število lažno negativnih rezultatov. Najbolj enostavna rešitev v danem trenutku je uporaba kontaktnih termometrov za merjenje telesne temperature, ki so precej bolj zanesljivi in povečini dosegajo točnost ±0,2 °C, kar pa pri preprečevanju prenosa nalezljivih bolezni ni najbolj enostaven postopek, kajti terja dodatne ukrepe razkuževanja, s čimer se podaljša čas merjenja in nekoliko povišajo stroški. Izboljšanje tehnologije merjenja z večino čelnih termometrov, da bi dosegali sprejemljivo točnost, bo zahtevalo precej vlaganj v razvoj in bo zagotovo podražilo njihovo trenutno ceno. Podobna ugotovitev velja za namenske termovizijske kamere, katerih cena že sedaj presega ceno čelnih termometrov za stokrat in več.

Literatura

[1] https://www.gov.si/teme/zagotavljanje-sledljivosti-meritev/
[2] https://www.bipm.org/kcdb/
[3] https://en.wikipedia.org/wiki/Liquid_crystal_thermometer
[4] https://patents.google.com/patent/US6292685B1/en
[5] https://www.alibaba.com/showroom/forehead-thermometer.html
[6] https://www.iso.org/standard/67348.html
[7] http://www.slo-akreditacija.si/acreditation/univerza-v-ljubljani-fakulteta-za-elektrotehniko-2/

Automatic ski-jump distance measurement with convolutional neural networks and computer vision

Matjaž Kukar[1], **David Nabergoj**[1]

[1] *University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenija*
E-mail: matjaz.kukar@fri.uni-lj.si, david.nabergoj@student.uni-lj.si

Abstract. *In ski jumping, video-assisted distance measuring is used only at the top-level competitions (world cup, continental cup). For smaller competitions it is prohibitively expensive and for this purpose we have developed a cost-effective system using commercially available equipment (a surveillance camera and an ordinary laptop computer). We further support distance-measuring umpires by introducing fully automatic distance measuring based on convolutional neural networks and computer vision techniques. We test our system on smaller ski jumping hills and show that while the system cannot completely replace the human operator, it can significantly speed-up the distance measuring process. Preliminary experiment on large hills confirm our experiences and will require only minor modifications for using multiple cameras.*

1 Introduction

In Slovenia, there has been a significant increase in ski jumping popularity in recent years, a consequence of excellent competitive results. At the primary level clubs reported doubled numbers of youngsters. This increased the burden on ski jumping coaches, organizers and professional staff in competitions. In our previous work [1, 2] we focused primarily on supporting distance umpires who have a demanding, exposed role, and their mistakes can significantly influence the competition outcomes. Our aim was to develop a system for supporting video-assisted distance measuring on smaller hills with accessible hardware requirements (a single-camera video system and a laptop).

In this paper we upgrade the system using convolutional neural networks and computer vision techniques in order to provide automatic distance measurements with reasonable accurracy. We evaluate the automatic measurements in ski jumping competitions on small hills in regional competitions (Cockta Cup) with respect to official results and show some directions for future development and use on larger hills.

2 Methods and materials

Ski jump distance is defined as a distance between the edge of the jumping ramp and a point where both ski jumper's skis have touched the ground with full surface [3, article 432.1]. The middle point between both legs is used when the legs are apart (e.g., Telemark landing style). FIS requires a jump distance accuracy of 0.5 m. There are several special cases [4] and even on the smallest competition hills it is difficult to manually measure the exact jump distance, as landing speeds exceed 10 m/s, and the angles between the landing slope and landing trajectories of ski jumpers are often minute [5]. Existing video distance measuring systems (Swiss Timing, Ewoxx) are basically video recorders and provide no automatic aid for umpires.

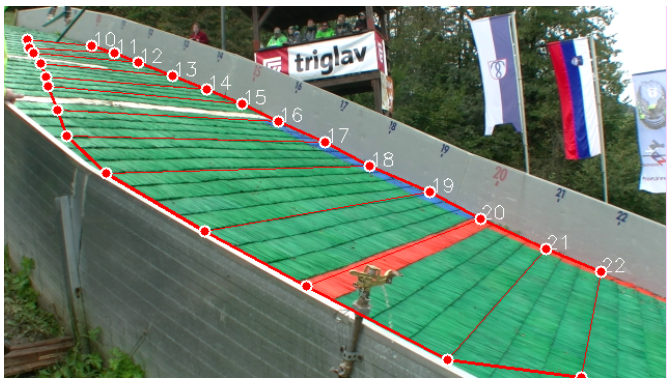


Figure 1: The measurement grid. Each measurement line is precisely calibrated to the particular hill and corresponds to a distance (in meters).

2.1 Automatic distance measurement system

The automatic distance measurement system analyses a sequence of frames where the ski jumper is visible. The ski jumper's location is determined for each frame, the frames without ski jumper are ignored (i.e. before of after the jump). The landing frame is the first frame in the sequence where the jumper has already landed. A measurement grid is overlaid on top of each frame. It contains lines annotated with distances (in meters) used to estimate the distance for each pixel in the frame. The main steps in the system are: determining the landing frame, finding the jumper's feet (a pixel) in this frame, and determining the distance based on the feet point and the measurement grid. The measurement grid is shown in figure 1. The distance measurement system is part of the full system, which helps the distance umpire.

2.2 Motion detection

The input to the distance measurement system is a sequence of frames where the jumper is visible. However, the jumper is only visible when jumping and near the landing area. Anything that occurs between jumps is not relevant to the system. We must first extract the relevant subsequence from a typically longer sequence of frames. We do so using an algorithm that receives a sequence of RGB images as input and outputs a vector $\mathbf{v}$ of equal length that tells us when a jumper is visible. The i-th element of $\mathbf{v}_i$ is equal to 1 if the jumper is visible in the i-th frame, and 0 otherwise. The algorithm works by using background subtraction [6] to determine the frames where motion occurs. A binary mask is generated for each frame, the white pixels in the mask correspond to the area with motion. The largest contour in the mask typically corresponds to the jumper. If it is not large enough (a parameter, dependent on

the camera distance), we conclude that the jumper is not in the frame. We use a median filter to remove some noise from this contour and finally extract the smaller image of the jumper from the axis-aligned bounding box of the processed contour [7]. Visualization of the process is shown in figure 2.

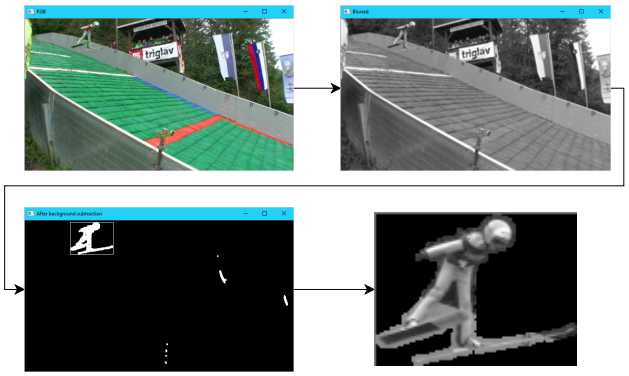


Figure 2: The jumper localization process. The input to the algorithm is a RGB image. It is converted to grayscale and blurred using a Gaussian filter. A motion mask is obtained using background subtraction. The largest moving object in the mask corresponds to the jumper. Finally, the jumper cutout is returned.

After the jump we get a sequence of boolean values, corresponding to the jumper being in the frame. The jump is therefore a subsequence of frames where the corresponding boolean value is true, meaning that the jumper is visible. There might be some frames in between the jump where the jumper is not detected. We allow some such frames in the jump, otherwise the complete jump might not be correctly extracted. We utilize a threshold $t = 2$, which allows for at most t consecutive frames without the detected ski jumper in the jump sequence.

2.3 Landing detection

The next part of the measurement pipeline is detecting the landing when it occurs. We use a convolutional neural network that is able to receive a frame of the jump segment as input. It activates the output 0 when the jumper in the frame is in the air and 1 when he is on the ground. 0 and 1 correspond to classes *Air* and *Ground*, respectively. We assign the landing to the first frame in the sequence classified as *Ground*. The architecture of the neural network is shown in figure 3.

2.4 Feet point calculation

Once we know the landing frame, we can process it using computer vision methods to find the precise jump distance. The procedure for this is two-fold: we first find the jumper's *feet point* and then use the measurement grid to calculate the actual distance. We find the feet point utilizing the mask generated in the motion detection phase. First we use a median filter afterwards to remove noise. The *feet point* is then calculated as the intersection of the line going through the jumper's body and the line through their skis (figure 4. We can to do this without a significant loss of accuracy because of the constrained position of the camera. The lines are found by using the RANSAC [8] procedure to fit two linear models on to motion mask. The models are selected if their lines are sufficiently vertical (corresponding to the line going through the body) or horizontal (corresponding to the line going through the skis). The use of RANSAC is justified since the motion mask contains many

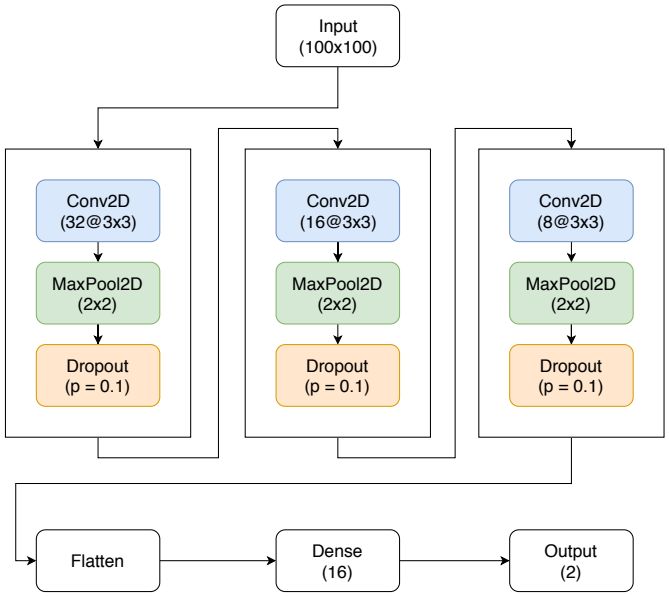


Figure 3: The CNN architecture. The input is a 100x100 grayscale image and the output is a vector with two elements whose values correspond to the probabilities of classifying the image as *Air* or *Ground*.

outlier pixels which are irrelevant to the lines (e.g., pixels corresponding to the jumper's hands are irrelevant to the line going through the body).



Figure 4: Feet point approximation. The blue lines correspond to the skis and the body of the jumper. The red circle represents the approximated feet point, which is computed as the intersection of the two lines.

2.5 Automatic measurement using measurement lines

Once the feet point is calculated, we use the measurement grid to find the measurement lines before above and below it. Each measurement line has a distance associated with it (e.g. 20 meters). We use linear interpolation with respect to the closest two measurement lines to calculate the precise distance at the *feet point*. Formally, the distance is calculated based on measurement line distances p_1 and p_2 as well as the euclidean distances d_1 and d_2 from the *feet point* to these lines along the

grid direction vector:

$$x = \frac{d_2}{d_1 + d_2} \cdot p_1 + \frac{d_1}{d_1 + d_2} \cdot p_2 \quad (1)$$

The values d_1 and d_2 correspond to the difference between T, P_1, and P_2. These points and lines are visualized in figure 5. The *grid direction vector* is based on *local direction vectors*. A *local direction vector* is computed based on two neighboring measurement lines, which are represented with their left and right points L_1, R_1 and L_2, R_2 as follows:

$$\mathbf{v_{1,2}} = \frac{\overrightarrow{L_1L_2} + \overrightarrow{R_1R_2}}{2} \quad (2)$$

The grid direction vector is computed as the component-wise weighted sum of local direction vectors. The grid direction vector should be more similar to local direction vectors, which correspond to measurement lines that are close to the feet point than to those further away. This reduces the degree to which imprecisely placed measurement lines affect the final distance. The values d_i and d_{i+1} denote the distance from the measurement lines i and $i+1$ to the feet point. The weights are calculated as follows:

$$w_{i,i+1} = e^{\left(-\frac{d_i + d_{i+1}}{2}\right)} \quad (3)$$

$$w'_{i,i+1} = \frac{w_{i,i+1}}{\sum_{i=1}^{n-1} w_{i,i+1}} \quad (4)$$

The weights are greater for local direction vectors, which are closer to the feet point. We use them to compute the grid direction vector:

$$\mathbf{v} = \sum_{i=1}^{n-1} w'_{i,i+1} \cdot \mathbf{v_{i,i+1}} \quad (5)$$

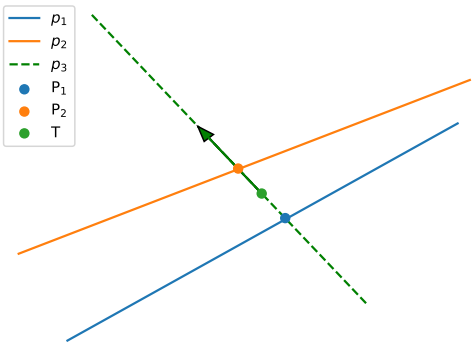


Figure 5: Visualization of the points and lines, used in the precise distance computation. The lines p_1 and p_2 correspond to measurement lines with their associated distances (e.g. 10 and 12 meters). The point T corresponds to the feet of the jumper. The measurement grid direction vector and the point T are used to describe the line p_3. The intersections of lines p_1 and p_2, as well as p_2 and p_3 are points P_1 and P_2, which use in the linear interpolation to obtain a more precise distance measurement.

2.6 Data

For 330 ski jumps recorded in junior competitions within PKP [1] and ŠIPK projects [2], we acquired the official results (measured by umpires using the eyes-only manual method). Ski jumps were recorded on smaller hills (HS 20-30m) and on average consisted of 36 frames recorded in HD resolution (1280×720 pixels) at 50 FPS. In order not to obstruct umpires' view, the camera was placed lower than the umpires. Each official measurement was further augmented by a ski-jumping professional coach using manual video measurement. For testing purposes the data was split to folds in a stratified manned, so that all frames belonging to a ski jump were assigned to the same fold (either for training or for testing).

3 Results

3.1 Landing frame detection results

The CNN by itself performed landing prediction relatively well. It was compared with a naive method which always predicted the landing to be in the middle of the sequence. The prediction results on an independent testing set are shown in a confusion matrix in table 1. Performing a 10-fold cross validation on the CNN results in an average accuracy of 0.922. There are two

Predicted class \ Actual class	*Air*	*Ground*
Air	995	56
Ground	122	1308

Table 1: Confusion matrix for the CNN predictions.

neurons in the last layer of the network, each outputting a value between 0 and 1. With normalization we get probabilities P_{Air} and P_{Ground} which correspond to the two possible classes. The landing is determined at the first frame where $P_{Air} < 0.5$. We can observe that the model often output probabilities of about 0.5 for frames around the true landing frame (figure 6).

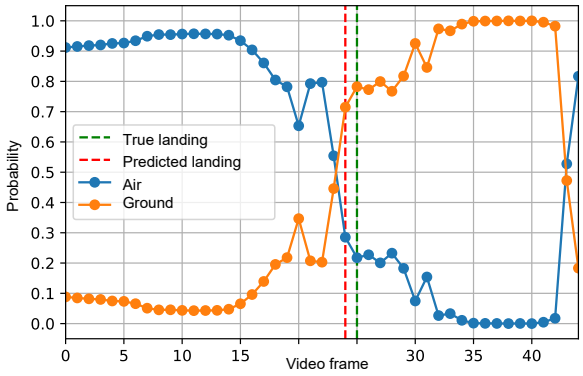


Figure 6: Sequence of predicted class probabilities on a test video. The model is initially very certain that the jumper is in the air. The value P_{Ground} increases as P_{Air} decreases. The landing is predicted as soon as P_{Air} first drops below 0.5.

3.2 Determining the landing distance

We considered several scenarios in order to evaluate important aspects of the system (table 2). First we evaluated the landing distance computation procedure independently of the landing frame predictions. This meant using the actual landing frames in the evaluation, which yielded the mean absolute error $MAE_{dist} = 0.404$ m. The error is larger when the jumper lands at the top of the landing area since the measurement lines are placed more densely together, a consequence of the camera placement very close to the hill and lower than umpires.

Jump distance	*MAE*
On true landing frame (testing set)	0.404
With landing frame detection (testing set)	0.586
With landing frame detection, 10-fold CV	0.946
With landing frame detection and bias correction, 10-fold CV	0.785

Table 2: Mean absolute error for jump distance predictions in meters. The first row correspond to predictions based on the actual (true) landing frames. The second row corresponds to full pipeline predictions, based on predicted landing frames.

The full measurement pipeline consists of the landing frame detection (CNN) and the landing distance determination. Due to misclassification of landing frames the $MAE_{pipeline} = 0.586$ m is a higher. Both numbers were obtained on an independent testing set without problematic jumps.

We also performed a 10-fold cross validation using all 330 jumps and achieved the total error of $MAE_{CV} = 0.946$ m. The predicted distances (figure 8) are mostly too short, indicating a systematic error (bias). On a typical laptop (without utilizing the GPU), the entire prediction procedure takes about 0.84 seconds. A systematic error (bias) in the system is caused be-



Figure 7: Measurement system in action at the HS=109 m hill in Kranj (bottom) and HS=21 m hill in Mengeš (top), Slovenia. The laptop and the camera are connected using a PoE switch.

cause the camera is placed too low on the side of the landing hill. As in World Cup competitions, we can specifically focus on jumps that are long enough (at least 17 meters in our case). By accounting for the median error of 0.51 meters to these predictions, the MAE_{CV} is reduced to 0.785 meters. By only analyzing the jumps where the CNN correctly predicts the landing, we achieve the *MAE* of 0.25 meters.

4 Conclusion

Our evaluation shows great potential for automatic ski jumping distance measuring. Even in current limited configuration we can measure distances with 1 m precision (at most 2 frames off, which requires very little human intervention) in reasonable conditions on the relevant part of the hill, both on small and

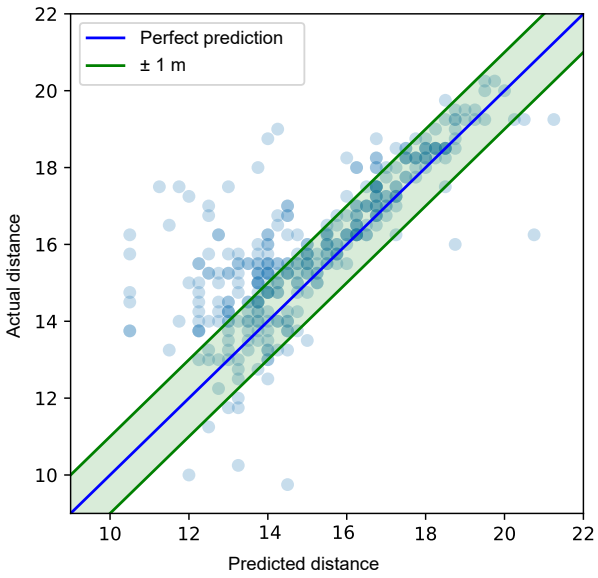


Figure 8: Distribution of the predicted and the true jump distances. The diagonal (blue line) indicates perfect predictions. Points within the green area refer to jumps where the prediction error is at most 1 meter. Points above the diagonal denote too short predictions.

large hills. While this is more than FIS-required 0.5 m, it is still a very useful addition. There is considerable interest from ski jumping clubs and Ski Association of Slovenia (SAS). For use on larger hills, slight modification of software are planned in order to allow for several cameras. The landing detection subsystem needs further testing under non-optimal (rain, snow) and artificial lighting conditions. We plan to achieve these aims in future partnership with SAS as we applied for co-founding from the Slovenian Foundation for Sports.

References

[1] T. Ciglarič et al. "Video meritve dolžin smučarskih skokov". In: *Zbornik šestindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2017* (2017), pp. 337–340.

[2] M. Kukar. "Evaluation and Prospects of Semi-Automatic Video Distance Measurement in Ski Jumping". In: *Proceedings of the 21st International Multiconference - IS 2018* (2018), pp. 62–65.

[3] *FIS: The International Ski Competition Rules (ICR)*. `https://fis-ski.com`. Accessed: 9. 8. 2019.

[4] *FIS: Guidelines to Video Distance Measurement of Ski Jumping 2011*. `https://fis-ski.com`. Accessed: 9. 8. 2019.

[5] N. Sato, T. Takayama, and Y. Murata. "Early Evaluation of Automatic Flying Distance Measurement on Ski Jumper's Motion Monitoring System". In: *2013 IEEE 27th International Conference on Advanced Information Networking and Applications*. IEEE. 2013, pp. 838–845.

[6] Z. Živković. "Improved adaptive Gaussian mixture model for background subtraction." In: *ICPR*. 2004, pp. 28–31.

[7] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[8] M. A. Fischler and R. C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". In: *Commun. ACM* 24.6 (1981), pp. 381–395. ISSN: 0001-0782. DOI: `10.1145/358669.358692`.

Optimizacija ojačevalnika za praktični šumni termometer

Rok Tavčar[1], Samo Beguš[1], Jovan Bojkovski[1]

[1]Univerza v Ljubljani, Fakulteta za Elektrotehniko, Laboratorij za meritve in kakovost
E-pošta: Rok.Tavcar@fe.uni-lj.si

Optimization of amplifier for practical Johnson noise thermometer

Abstract. *Over last decade the interest of making practical primary thermometer is increasing. One of most promising primary thermometry is based on Johnson noise in resistor. The largest obstacle with Johnson noise thermometer is measuring noise voltage, which is in range of few microvolts RMS for practical resistors. In this article is presented one of possibility for measuring Johnson noise and its optimization.*

1 Uvod

Poznamo več vrst primarnih merilnikov temperature, mednje uvrščamo plinski termometer, akustični termometer, termometer celostnega sevanja, šumni termometer itd. Vsi našteti termometri potrebujejo za delovanje laboratorijske pogoje in dolge merilne čase. V zadnjem času pa poteka razvoj primarnih termometrov, ki so bolj praktični za industrijsko uporabo. Eden izmed potencialnih praktičnih primarnih termometrov je tudi šumni termometer [1,2]. Tak termometer deluje na osnovi termičnega šuma, ki se generira v vsakem uporu, ki je nad temperaturo absolutne ničle. Električno napetost ki je posledica termičnega gibanja, lahko opišemo z:

$$\overline{V_T^2} = 4k_B TR\Delta f \qquad (1)$$

kjer je k_B Boltzmannova konstanta, T temperatura, R upornost in Δf pasovna širina. Za upor 5 kΩ pri temperaturi 20 °C, ki mu merimo šumno napetost s pasovno širino 1 MHz, dobimo efektivno vrednost šuma $V_T = 9\ \mu V$. Ta napetost je primerljiva z vhodnim šumom najboljših nizko šumnih ojačevalnikov. Pri zahtevi, da je negotovost izmerjene temperature manjša od 0,1 °C, mora biti negotovost izmerjene vrednosti šuma manjša od 1,3 nV.

Obstoječi primarni termometri na podlagi Johnsonovega šuma se običajno uporabljajo v povezavi s kvantnim psevdonaključnim generatorjem napetostnega šuma (QVNS) [3,4] na temperaturnem območju od 0,3 K do 1100 K. Taka kombinacija omogoča izjemno majhne negotovost meritve temperature, vendar je za to potrebno hladiti QVSN s tekočim helijem, meritev pa traja po več dni [5,6]. Tak način ni primeren za uporabo v industriji, zato je potrebno uporabiti drug vir napetostnega šuma za primerjavo z izmerjenim Johnsonovim šumom.

2 Metode

Vse metode za merjenje šumne napetosti uporabljajo korelator, s katerim se izloči šum uporabljenih ojačevalnikov. Pri tem je predpostavljeno, oziroma je s pravilno izdelavo ojačevalnika zagotovljeno, da je šum ojačevalnikov med seboj nekoreliran. Izhod korelatorja je:

$$V_k = \sum V_1 * V_2^* \qquad (2)$$

kjer je V_1 spekter izhoda prvega ojačevalnika in V_2 spekter izhoda drugega ojačevalnika.

Iz enačbe 2 ni mogoče ugotoviti absolutne vrednosti šumne napetosti, ker ni mogoče poznati točnega ojačanja uporabljenih ojačevalnikov. Za določitev absolutne vrednosti šumne napetosti se uporabljata dva načina.

Pri prvem načinu se izmenično meri šumno napetost pri neznani temperaturi in se jo primerja s šumno napetostjo pri znani temperaturi, izmerjeni z istim parom ojačevalnikov. To se stori z dodatnim uporom, ki je na znani temperaturi in deluje kot umetni izvor belega šuma z znano močjo. Neznano temperaturo se izračuna z:

$$T = T_0 \left(\frac{V_n}{V_0}\right)^2 \cdot \frac{R_0}{R_n} \qquad (3)$$

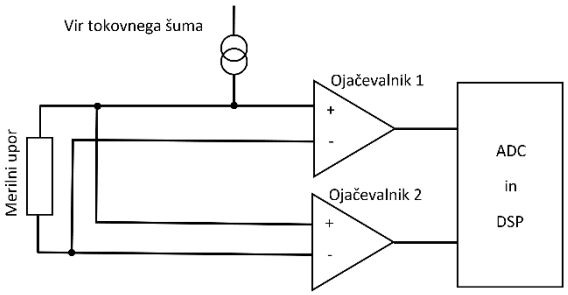
kjer je T_0 temperatura na dodatnem upor, R_0 je upornost dodatnega upora, V_0 je izhod korelatorja ob merjenju šumne napetosti na dodatnem uporu, V_n je izhod korelatorja ob merjenju šumne napetosti na merilnem uporu in R_n je upornost merilnega upora.

Prednost tega načina je bolj enostavna izvedba, saj potrebuje dva upora in rele z dvema sistemoma kontaktov (DPDT). Slabost tega načina je neenaka impedanca v povezavah do ojačevalnikov in počasno merjenje, ker je potrebno šumno napetost izmeriti na dveh uporih.

Pri drugem načinu se v upor inducira znan tok z znano amplitudo in z znanim spektrom, nato pa se primerja izmerjena šumna napetost in izmerjena inducirana napetost na uporu. Prednost tega načina je enaka impedanca in hitrejše merjenje šumne napetosti, ker ne prihaja do preklopov med upori. Slabost tega načina pa je potreba po signalnem generatorju, ki mora biti sinhroniziran z enoto za zajem podatkov in zahtevna obdelava signalov.

2.1 Merjenje ojačenja z metodo induciranja toka v merilni upor

Na sliki 1 je prikazana shema termometra na podlagi Johnsonovega šuma z metodo induciranega toka. Temperaturo merimo z merjenjem šumne napetosti na merilnim uporu. Šumno napetost se ojači z dvema ojačevalnikoma, nato pa digitalizira z analogno digitalnim pretvornikom (ADC). Korelator in naknadna obdelava signalov je izvedena na digitalnem signalnem procesorju (DSP). Vir tokovnega šuma je uporabljen za kalibracijo obeh ojačevalnikov.

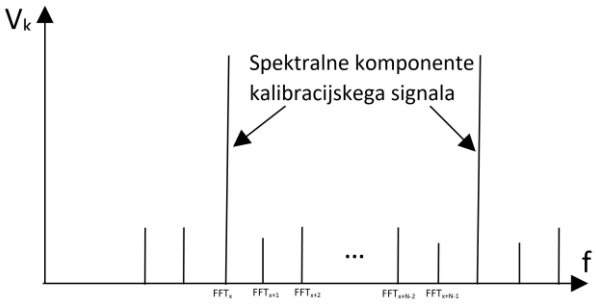


Slika 1: Poenostavljena shema merilnika Johnsonovega šuma z metodo induciranja toka v merilni upor

Tokovni šum je sestavljen iz vsote sinusnih signalov določenih frekvenc, enake amplitude in naključne faze. Tak signal izgleda na spektru kot glavnik (slika 2) in je podoben Gaussovemu šumu. Taka oblika kalibracijskega signala je uporabljena zato, ker omogoča meritev ojačenj ojačevalnikov in dušenja signala zaradi povezovalnih žic v odvisnosti od frekvence. Dušenje signala je izrazito pri višjih frekvencah (1 MHz) zaradi kapacitivnosti povezovalnih kablov in vhodne kapacitivnosti ojačevalnika.

Da je kalibracijski signal samo v enem razdelku FFT spektra, je poskrbljeno s sinhronizacijo vira tokovnega šuma in ADC. Prav tako je potrebno sinhronizirati periodo kalibracijskega signala in izračun spektra s FFT. Izbire frekvenc oziroma period sinusnih signalov v tem signalu so omejene, saj mora biti v eni periodi kalibracijskega signala celo število period posameznega sinusnega signala.

Na sliki 2 je prikazan izhod korelatorja. Vidne so spektralne komponente kalibracijskega signala (razdelek FFT_x), ki se ponavljajo s periodo N.



Slika 2: Prikaz izhoda korelatorja, na spektru so vidne spektralne komponente (razdelek FFT_x), ki prikazujejo kalibracijski signal, ki se ponavlja vsakih N razdelkov. Med njimi so spektralne komponente merjene šumne napetosti

Amplituda kalibracijskega signala v spektru mora biti veliko višja od amplitude napetostnega šuma za pravilno delovanje termometra. V primeru, da je vir kalibracijskega signala napetostni vir, ki tvori z merilnim uporom napetosti delilnik, lahko temperaturo izračunamo iz:

$$T = \frac{(V_{cal}^2 - (NX-1)4k_B T_{cal} R_{cal} \Delta f) R}{(NX-1)4k_B R_{cal}^2 \Delta f} \qquad (4)$$

kjer je V_{cal}^2 efektivna napetost kalibracijskega signala, N perioda kalibracijskega signala v FFT spektru, R_{cal} upornost kalibracijskega upora, T_{cal} temperatura kalibracijskega upora R_{cal}, Δf pasovna širina enega razdelka FFT spektra in X je razmerje med kalibracijskim signalom in izmerjenim šumom in je enako:

$$X = \frac{V_x}{\sum_{i=1}^{N-1} V_{x+i}} \qquad (5)$$

kjer je V_x vrednost FFT spektra izhoda korelatorja, kjer je prisoten kalibracijski signal, V_{x+i} so vrednosti FFT spektra izhoda korelatorja med dvema kalibracijskima signaloma.

3 Optimizacija ojačevalnika

Glavni pogoj za pravilno delovanje termometra je zagotovitev nekoreliranosti šuma med obema ojačevalnikoma. Poleg tega pa je pomembno, da ima ojačevalnik čim večjo pasovno širino, čim večje ojačenje in majhno porabo energije.

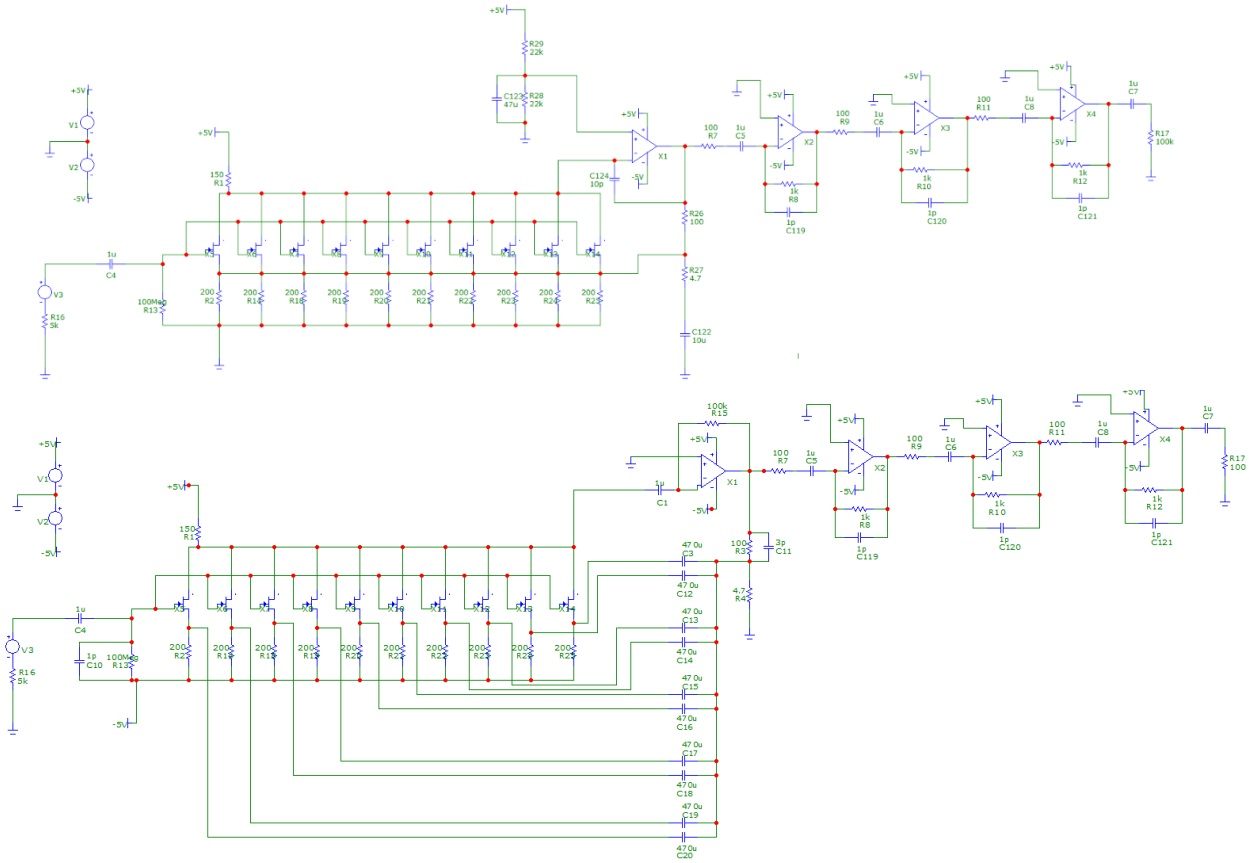
Največji prispevek h koreliranosti šuma obeh ojačevalnikov ima njihov tokovni šum, ker le ta ustvari padec napetosti na uporu, ki jo merimo z obema ojačevalnikoma. Napetostni šum ojačevalnikov nima vpliva na koreliranost signalov, vendar podaljšuje čas meritev oziroma povečuje negotovost izmerjene temperature.

Večja pasovna širina in večje ojačenje zvišata efektivno vrednost napetosti šuma in s tem izboljšata ločljivost sistema in zmanjšata negotovost izmerjenega signala. Majhna poraba je pomembna zato, ker sta oba ojačevalnika napajana s svojima baterijama, da ne prihaja do koreliranosti signala zaradi istega napajalnega vira.

Iz teh zahtev je razvidno, da uporaba tudi najboljših (nizko šumnih) operacijskih ojačevalnikov ali instrumentacijskih ojačevalnikov ni dovolj za izpolnitev zahtev [7]. Pomanjkljivosti operacijskih ojačevalnikov se lahko odpravi z izdelavo JFET predojačevalnika za operacijski ojačevalnik, z izdelavo diskretnega JFET ojačevalnika ali z izdelavo bipolarnega tranzistorskega predojačevalnika.

3.1 JFET predojačevalnik

Glavna prednost JFET pred bipolarnim tranzistorjem je zelo nizek tokovni šum [8], ki je običajno pod 1 $fA/\sqrt{Hz}$, kar omogoča vzporedno vezavo večjega števila tranzistorjev in s tem zmanjšanje njihovega napetostnega šuma. Dodatna prednost je tudi višja vhodna upornost. Ti dve lastnosti povesta, da je za vhodne signale, ki imajo upornost večjo od nekaj kΩ, JFET predojačevalnik boljša izbira, saj ima nižji šum. Na sliki 3 je prikaz dveh simuliranih vezij. Na sliki ni vidno filtriranje napajanja.



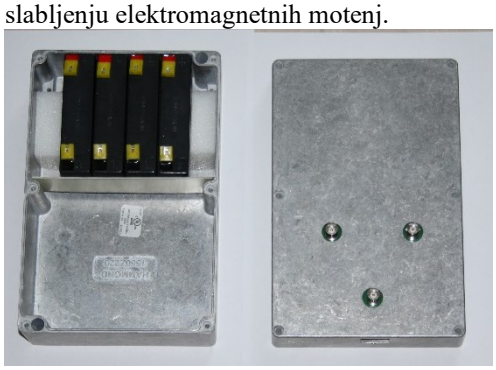
Slika 3: Shema ojačevalnika z JFET predojačevalnikom različice 1 (zgoraj) in različice 2 (spodaj)

Obe vezji sta zasnovani modularno, kar pomeni, da je mogoče spreminjati število uporabljenih tranzistorjev, glede na zahteve napetostnega in tokovnega šuma.

Merilni upor je simuliran kot brezšumni upor zaporedno z virom napetostnega šuma. Sam JFET predojačevalnik je sestavljen iz desetih JFET, kar zmanjša vhodni napetosti šum za trikrat ($\sqrt{10}$) na $0{,}5\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ in iz operacijskega ojačevalnika za zagotovitev dovolj velikega odprtozančnega ojačenja. Predojačevalniku sledijo še trije operacijski ojačevalniki, ki imajo skupno ojačenje 1000. Ojačenje predojačevalnika je nastavljeno na 22,3 z uporoma R_3 in R_4. Celotni ojačevalnik ima ojačenje torej 22300. Glavna razlika med predojačevalnika je v načinu povratne zanke. Pri prvi verziji ima povratna zanka tudi enosmerno komponento, kar izboljša stabilnost predojačevalnika, ampak tudi oteži izdelavo, ker je potrebno uporabiti tranzistorje s podobnimi karakteristikami. To je zahtevno opravilo, ker so proizvodne tolerance največkrat zelo široke (tudi do 300 %). Druga različica omogoča nastavite delovne točke za vsak tranzistor posebej, vendar je povratna zanka brez enosmerne komponente.

3.2 Ohišje ojačevalnika in merilnega upora

Ker so velikosti merjenih napetosti majhne, je potrebno nameniti veliko pozornost preprečevanju in zmanjševanju motenj iz okolice. To je še posebno problematično zato, ker je merilni upor oddaljen od preostanka vezja za pol metra. To je potrebno zato, da je možno meriti temperature, višje od tistih, ki bi uničile ojačevalnik. Glavni poudarek je na preprečevanju in slabljenju elektromagnetnih motenj.



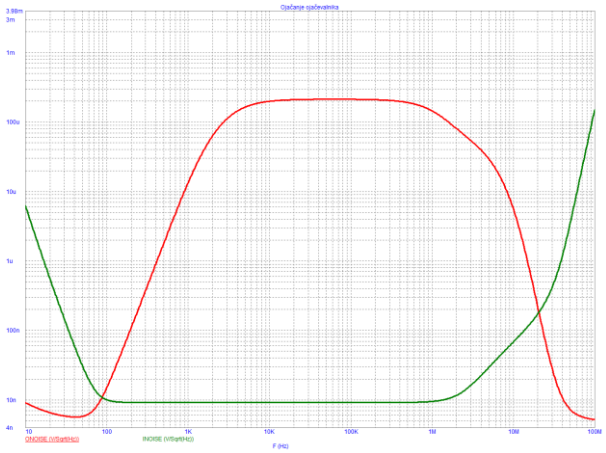
Slika 4: Notranjost ohišja z akumulatorji (levo) in pokrov s priključki (desno)

Ohišje ojačevalnika je narejeno iz aluminija debeline 5 mm, v katerega so narejene odprtine za izhodni in vhodni signal, drugače pa je popolnoma zaprto. Za povezavo med merilnim uporom in ojačevalnikom je uporabljen kabel z dvojnim oklopom in posukanimi žicami za povezavo z ojačevalnikoma. Prednost dvojnega oklopa je, da lahko z ustrezno vezavo ustvarimo tudi slabljenje magnetnega polja, kar samo z enim oklopom ni mogoče doseči.

4 Rezultati

Na sliki 5 je prikazan rezultat simulacije vezja iz slike 3 zgoraj. Z zeleno barvo je označen vhodni šum, z rdečo pa izhodni signal. Pri frekvencah do 100 Hz je na vhodnem signalu opaziti porast šuma, kar je posledica tokovnega

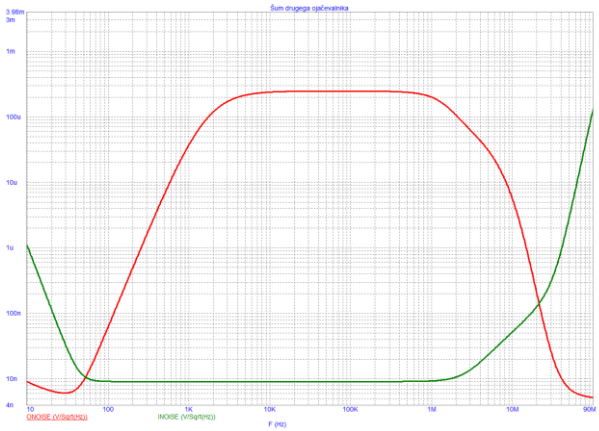
šuma JFET, ki imajo izrazit 1/f šum pri nizkih frekvencah.



Slika 5: Spekter napetostnega šuma na vhodu vezja (zelena) in spekter šuma na izhodu vezja (rdeča) verzije 1.

Skupna poraba vezja je 280 mW. Pasovna širina ojačevalnika je od 4,7 kHz do 1,0 MHz pri ojačenju 87,3 dB.

Na sliki 6 je prikazan rezultat simulacije drugega vezja.

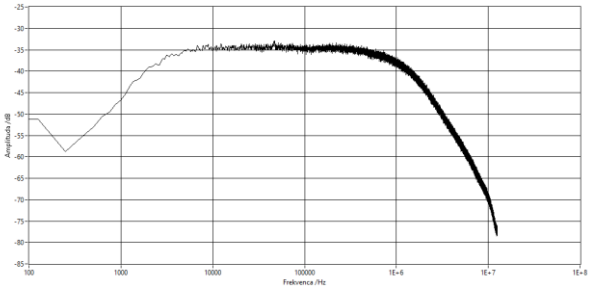


Slika 6: Spekter napetostnega šuma na vhodu vezja (zelena) in spekter šuma na izhodu vezja (rdeča) verzije 2.

Skupna poraba vezja je 398 mW, Pasovna širina ojačevalnika je od 3,1 kHz do 1,2 MHz pri ojačenju 88,7 dB.

Kot boljša izbira se obnese druga različica, saj ima večje ojačenje pri večji pasovni širini. Dodatna prednost je, da tako vezje zaduši Millerjev pojav na vhodnih tranzistorjih, ker operacijski ojačevalnik drži invertirajoči vhod na virtualni zemlji.

Na sliki 7 je prikazana preliminarna meritev delovanja drugega vezja. Meritev je bila izvedena z Ettus USRP N200 pri vzorčni frekvenci 25 MHz. Graf je povprečje meritev pridobljenih v 10 min.



Slika 7: Preliminarna meritev delovanja drugega vezja.

Iz slike 7 je razvidno, da je delovanje izdelanega ojačevalnika v skladu z rezultati simulacije.

5 Zaključek

Praktični termometer na podlagi Johnsonovega šuma meri napetosti v območju nekaj µV s pasovno širino 1 MHz. Predstavljeni ojačevalnik z JFET predojačevalnikom zadošča potrebam meritev Johnsonovega šuma z uporabo korelatorja. Ojačevalnik ima ojačenje 86 dB in pasovno širino od 3 kHz do 1,2 MHz, pri porabi 398 mW. Skupni izhodni šum ojačevalnika je 47 $\mu V/\sqrt{Hz}$ pri tem ojačenju in vhodni upornosti 5 kΩ.

Literatura

[1] P. Bramley, D. Cruickshank, and J. Pearce, “The Development of a Practical, Drift-Free, Johnson-Noise Thermometer for Industrial Applications,” International Journal of Thermophysics, vol. 38, no. 2, p. 25, Dec. 2016

[2] L. Callegaro, V. D’Elia, M. Pisani, and A. Pollarolo, An absolute Johnson noise thermometer. 2009

[3] S. W. Nam, S. P. Benz, P. D. Dresselhaus, C. J. Burroughs, Jr., W. L. Tew, D. R. White, and J. M. Martinis, “Progress on Johnson noise thermometry using a quantum voltage noise source for calibration,” IEEE Trans. Instrum. Meas., vol. 54, no. 2, pp. 653–657, Apr. 2005.

[4] S. P. Benz, J. M. Martinis, S. W. Nam, W. L. Tew, and D. R. White, “A new approach to Johnson noise thermometry using a Josephson quantized voltage source for calibration,” in Proc. TEMPMEKO, 2001, pp. 37–44.

[5] S. P. Benz, J. Qu, H. Rogalla, D. R. White, P. D. Dresselhaus, W. L. Tew, and S. W. Nam, “Improvements in the NIST Johnson noise thermometry system,” IEEE Trans. Instrum. Meas., vol. 58, no. 4, pp. 884–890, Apr. 2009.

[6] D. R. White, S. P. Benz, J. R. Labenski, S. W. Nam, J. F. Qu, H. Rogalla, and W. L. Tew, “Measurement time and statistics for a noise thermometer with a synthetic-noise reference,” Metrologia, vol. 45, no. 4, pp. 395–405, Aug. 2008

[7] P. Horowitz, W. Hill, The Art of Electronics 3E, 2015

[8] J. Vojtěch, Design of ultra low noise amplifiers, 2018

Avtomatizacija sistema za umerjanje merilnikov napetostnega razmerja

Ana Mandeljc[1], Miha Hiti[2], Gaber Begeš[1]

[1] Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška c. 25, 1000 Ljubljana
[2] Zavod za gradbeništvo Slovenije (ZAG), Laboratorij za metrologijo, Dimičeva 12, 1000 Ljubljana
E-pošta: ana.mandeljc@fe.uni-lj.si

Automation of system for calibration of voltage ratio meters

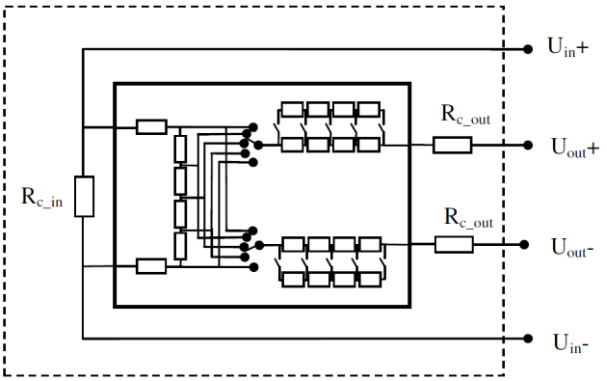
Abstract. *Our objective was to develop a control circuit in combination with a suitable drive for changing the position of the switches for a voltage ratio meter calibration system. It has two rotary switches, each switch allowing nine positions. At the same time, it was also necessary to establish communication between the computer and the voltage ratio meter, so that the computer could automatically obtain and store the measured values for each of the selected combination of switches.*

1 Uvod

Za umerjanje merilnikov napetostnega razmerja lahko kot referenčne etalone uporabljamo simulatorje mostičev z uporovnimi lističi. Ti omogočajo izbor različnih napetostnih delilnih razmerij, s čimer v izbranem območju merilnika lahko ugotovimo odstopanje prikaza od referenčnih vrednosti. Pri umerjanju simulatorjev nacionalni meroslovni inštituti in umerjevalni laboratoriji dosegajo najboljše vrednosti negotovosti umerjanja od 10 nV/V oziroma 20 nV/V navzgor, kar pa v primeru nizkih merjenih vrednosti še vedno pomeni preveliko relativno negotovost umerjanja za mnoge končne uporabnike. Negotovost umerjanja merilnikov napetostnega razmerja pri nizkih vrednostih lahko bistveno izboljšamo z uporabo simulatorja, ki preverja linearnost merilnika po postopku kombinatornega umerjanja [1].

V postopku kombinatornega umerjanja uporabljamo več med seboj neodvisnih etalonov, katerih vrednosti lahko seštevamo in preko tega tvorimo nove referenčne vrednosti. S tem dobimo večje število referenčnih vrednosti, kar nam omogoči ovrednotenje pogreška linearnosti instrumenta. Tega ovrednotimo na podlagi porazdelitve razlike med izračunanimi ter izmerjenimi vsotami združenih etalonov, pri čemer ni nujno, da poznamo vrednosti osnovnih etalonov. Za izvedbo kombinatornega umerjanja lahko uporabimo simulator na osnovi uporovnega napetostnega delilnika, ki ob konstantni vhodni in izhodni upornosti omogoča prosto združevanje posameznih izhodnih napetosti, tako na posameznem uporu kot tudi poljubni kombinaciji zaporedno vezanih uporov v delilniku [1].

Električno vezje simulatorja (Slika 1) sestoji iz napetostnega delilnika, ki ima med napetostne napajalne sponke zaporedno vezane upore, stikalnega vezja s stikali, ki so priključena v vozlišča prej omenjene uporovne verige in dveh kompenzacijskih uporovnih vezij, ki sta zaporedno vezani v izhodni veji delilnika [1].



Slika 1: Skica vezja v simulatorju [1].

S pomočjo kompenzacijskih uporov v kompenzacijskih vezjih, izhodna upornost napetostnega delilnika ostaja nespremenjena, ne glede na aktivirane upore v delilniku. Ob premiku para stikal za izbiro vozlišč v napetostnem delilniku, se aktivirata tudi odgovarjajoča para stikal v kompenzacijskih vezjih, s čimer se vrednosti ustreznih uporov v teh vezjih zmanjšajo na polovično vrednost v vsaki veji. S tem se ohrani konstantna vsota upornosti v izhodni zanki. Razporeditev elementov zagotavlja tudi simetrično konfiguracijo vezja. Vse to nam omogoča izbor padcev napetosti na posameznem uporu in poljubni kombinaciji zaporedno vezanih uporov delilnika [1].

Vezje omenjenega simulatorja je sestavljeno iz 8 osnovnih uporov, kar pomeni 36 možnih kombinacij uporov, posledično pa tudi 36 vrednosti delilnega razmerja. Te vrednosti so razporejene preko celotnega območja uporabe simulatorja, zaradi česar lahko enakomerno preverimo linearnost instrumenta.

Meritev, ki vključuje preverjanje vseh 36 možnih kombinacij uporov, se trenutno opravlja ročno. To je zaradi velikega števila potrebnih preklopov stikal in časa, potrebnega za ustalitev izmerjene vrednosti med posamezno meritvijo, utrujajoče in monotono opravilo za operaterja. Ta mora stikala preklapljati po vnaprej določenem protokolu, izmerjene vrednosti pa prav tako beležiti ročno, kar v meritev vnaša možnost napak. Dotikanje stikal lahko vpliva tudi na dvig temperature v simulatorju, hkrati pa je ta izpostavljen še drugim negativnim zunanjim vplivom okolja kot so vlaga, tresljaji in podobno.

Prav zaradi tega avtomatizacijo postopka preverjanja linearnosti merilnikov napetostnega razmerja vidimo kot smotrno rešitev za razbremenitev operaterja, zmanjšanje števila potencialnih napak in posledično izboljšanje kakovosti opravljenih meritev. Avtomatizacija postopka bi prav tako omogočila, da simulator lahko zapremo v komoro, s čimer bi izločili vpliv motečih zunanjih faktorjev. Nenazadnje pa avtomatizacija postopka pomeni konstantne ter ponovljive časovne intervale preklopov stikal ter zajema izmerjenih vrednosti, kar omogoči uporabo tudi bolj kompleksnih merilnih protokolov in večjo primerljivost rezultatov.

Namen članka je predstaviti postopek avtomatizacije sistema za preverjanje linearnosti, ki je del postopka za umerjanje merilnikov napetostnega razmerja. Podoben problem je že obravnaval Schäck, ki pa je za aktivacijo stikal, namesto para rotacijskih stikal z več stikalnimi ploščami, uporabil releje [2].

2 Merilni sistem in postopek avtomatizacije

2.1 Merilni sistem

Merilni sistem sestoji iz merilnika napetostnega razmerja z resolucijo ranga nV/V, simulatorja pretvornika z uporovnimi lističi, krmilnika ter računalnika. Simulator in merilnik sta med seboj povezana s 6 žilnim kablom. Da bi preklapljanje stikal simulatorja lahko avtomatizirali, smo se odločili za uporabo 2 dvofaznih koračnih motorjev, proizvajalca Sanyodenki, model 103H7126-0740. Osnovni korak motorja je 1.8°. Razlika med sosednjima položajema stikala je 30°, kar pomeni, da mora motor za en preklop stikala opraviti 17 korakov. Za krmiljenje motorjev smo uporabili 4 kanalno krmilno ploščico K8097, proizvajalca Velleman. Po uspešni pritrditvi vseh komponent na ploščico, ta preko USB povezave z računalnikom omogoča krmiljenje do 4 koračnih motorjev s pomočjo programske opreme, dostopne na spletu. Merilnik in računalnik med seboj komunicirata preko pošiljanja SCPI ukazov. Merilni sistem, z izjemo računalnika, je prikazan na Sliki 2.

Koračna motorja smo na nastavka za stikala pritrdili tako, da smo nanju privili kovinski puši, ki sta na drugi strani povezani z osema motorjev. Da se ohišji motorjev med vrtenjem osi ne bi premikali, smo ju privili na skupno kovinsko ploščico. Tak način povezave osi motorjev z nastavki za stikala omogoča enostavno namestitev in snemanje, zato manipulacijo s stikali zlahka spreminjamo iz ročne v avtomatizirano in obratno.

2.2 Program

Izkazalo se je, da že obstoječa programska oprema ne ustreza našim potrebam, saj smo morali za vsak premik motorja posebej poslati ukaz preko uporabniškega



Slika 2: Merilni sistem - merilnik napetostnega razmerja (zgoraj), simulator pretvornika z nameščenima koračnima motorjema (spodaj desno) in krmilnik (spodaj levo).

vmesnika. Zato smo se odločili razviti svoj program za krmiljenje motorjev.

Najprej smo s pomočjo programske opreme »Device Monitoring Studio«, proizvajalca HDD Software Ltd., zajeli ukaze, ki jih je že obstoječi program preko uporabniškega vmesnika pošiljal krmilni ploščici ter jih analizirali. Ugotovili smo, da posamezen ukaz sestoji iz 56 heksadecimalnih vrednosti – 4 končnih, 4 namenjenih preverjanju pravilnosti zapisa ukaza ter po 12 vrednosti za vsakega izmed 4 motorjev, ki podajajo informacijo o številu korakov, hitrosti ter smeri vrtenja. Ker smo uporabljali le 2 motorja, smo morali vrednosti za neuporabljena motorja nastaviti na 0. Na podlagi teh ugotovitev smo oblikovali ukaze, preko katerih krmilnik lahko sočasno ali posamično zavrti enega ali oba motorja hkrati, vsakega v poljubno smer, pri poljubni hitrosti in številu korakov.

V nadaljevanju smo v razvojnem okolju PyCharm, specifično namenjenega programiranju v jeziku Python, oblikovali program, ki je omogočal komunikacijo ter izmenjavo krmilnih ukazov med računalnikom in krmilnikom. Kasneje smo program dopolnili tako, da omogoča še komunikacijo z merilnikom, uvoz celotnega seznama ukazov za premike stikal iz tekstovne datoteke ter beleženje izmerjenih vrednosti v novo tekstovno datoteko.

V končni obliki program deluje tako, da najprej preveri ali sta stikali simulatorja v začetnih položajih – na mestih A0 in B0. Nato odpre serijska vrata za komunikacijo s krmilnikom in merilnikom, s slednjim preko pošiljanja konfiguracijskih ukazov omogoči komunikacijo ter preveri njegovo identifikacijo, uvozi datoteko z ukazi in ustvari datoteko za shranjevanje izmerjenih vrednosti. Datoteka z ukazi ima v vsaki vrstici zapisan en ukaz oziroma željen položaj stikal, v obliki »A1B4« – to pomeni stikalo A na položaju 1 in stikalo B na položaju 4, kar aktivira verigo uporov od R2 do R4. Program za vsako vrstico v datoteki izračuna razliko v položajih stikal glede na predhodni in trenutni ukaz. Glede na vrednost razlike, program pošlje krmilniku ustrezno število ukazov za vrtenje posameznega stikala v levo ali desno. Namreč, vsak izmed širih možnih ukazov – stikalo A levo/desno in stikalo B levo/desno, spremeni položaj stikala le za eno mesto v levo ali desno, kar posnema ročno upravljanje stikal. Na podlagi empiričnih opazovanj smo ugotovili, da je, glede na izbrano hitrost in kot zasuka, potrebno med dvema ukazoma pustiti 0,1 s časa, da krmilna ploščica uspe izvesti predhodni ukaz ter sprejeti novega. Po tem, ko sta stikali na novih položajih, program izvede premor, da se vrednost na merilniku ustali. Nato izbere ustrezen kanal na merilniku, ga prebere, iz prebranih podatkov izlušči izmerjeno vrednost in jo zapiše v novo vrstico tekstovne datoteke, namenjene shranjevanju. Ko program prebere prazno vrstico v tekstovni datoteki z ukazi, to smatra kot konec datoteke, zapre vrata za serijsko komunikacijo ter zaključi z izvajanjem.

3 Meritve

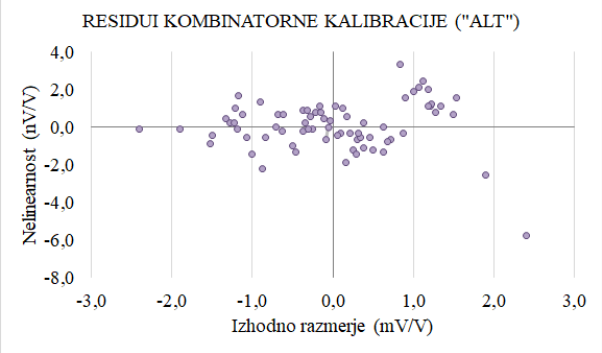
Da bi preverili uspešnost avtomatizacije postopka preverjanja linearnosti, smo s sistemom izvedli več meritev. Pri tem smo uporabili tri različne protokole – protokol s pozitivno polariteto (»POZ«), pri katerem se stikalo A premakne v naslednji položaj po tem, ko se stikalo B premakne preko vseh možnih položajev; protokol z obratno polariteto (»NEG«), pri katerem je gibanje stikal ravno obratno kot pri prvem protokolu (»POZ«) in rezultira v negativnih vrednostih napetostnega razmerja; ter izmenični protokol (»ALT«), pri katerem se zaporedno izmenjata položaja stikal A in B za pozitivno in negativno polariteto za vsako kombinacijo. Tako postopek preverjanja linearnosti merilnika zajema meritve izhoda simulatorja z merilnikom napetostnega razmerja pri vseh možnih kombinacijah.

Meritve so se, poleg izbora merilnega protokola, med seboj razlikovale tudi glede na vrsto in frekvenco uporabljenega filtra na merilniku. Predvsem frekvenca filtra močno vpliva na hitrost ustalitve merjene vrednosti, zato moramo temu primerno prilagoditi čas trajanja premora izvajanja programa pred odčitanem vrednosti. Zaradi tega smo v začetnih meritvah najprej preverjali, kolikšen čas je potreben za ustalitev merjene vrednosti pri posamezni kombinaciji vrste in frekvence filtra. Pri končnih meritvah, katerih rezultate smo primerjali tudi z ročno izvedbo merilnega protokola, smo izbrali Bessel filter, frekvence 0,1 Hz ter s trajanjem premora pred zajemom vrednosti 30 s.

Med izvajanjem meritev smo občasno opazili problem nepopolno opravljenega premika stikala pri protokolu »ALT« – meritve s takšno napako smo izključili iz nadaljnje obravnave. Problem je mehanske narave ter ga odpravimo s ponovno zategnitvijo vijakov – ker pri tem protokolu prihaja do večjega števila preklopov in večjih sprememb v položaju stikal nasploh, se občasno zgodi, da puša v manjši meri zdrsne z gredi motorja, ker se stiki med vijaki zrahljajo.

4 Rezultati

Ker pravih vrednosti osnovnih delilnih razmerij ne poznamo, pogrešek pri teh vrednostih določimo glede na regresijsko funkcijo vseh meritev. Pri ostalih vrednostih pa pogrešek kazanja določimo kot odstopanje kazanja vsote delilnih razmerij od izračunane vsote osnovnih delilnih razmerij za vsako točko. Standardno negotovost rezultata meritve linearnosti predstavlja standardna deviacija ostankov linearne regresije [1, 3]. Rezultate meritve nelinearnosti merilnika, torej odstopanje od »linearnega fita«, pri protokolu »ALT«, prikazuje Slika 3. Tem boljši rezultat pomeni tem manjše odstopanje in raztros vrednosti od 0 nV/V.



Slika 3: Nelinearnost umerjanega merilnika - protokol »ALT«.

Standardne negotovosti vseh meritev so zbrane v Tabeli 1 na naslednji strani. Kot zadovoljiv rezultat smatramo vrednosti standardne negotovosti ranga 10^{-6} in manj. Tak kriterij smo izbrali zato, ker je 10^{-6} razločljivost prikaza trenutno najboljšega komercialno dosegljivega merilnika napetostnega razmerja za mostiče z uporovnimi lističi HBM DMP41 [4], s katerim smo tudi sami izvajali meritve. Prav tako je laboratorij akreditiran za umerjanje z vrednostmi negotovosti $3x10^{-6}$ za ročni postopek kombinatornega umerjanja, zato mora biti standardna deviacija kombinatornega umerjanja ranga 10^{-6} ali manj. Standardna negotovost ranga 10^{-5} bi bila prevelika, in zanjo niti ne bi rabili kombinatornega postopka.

Iz Tabele 1 je razvidno, da je izbira dovolj dolgega premora pred zajemom vrednosti ključnega pomena za uspešnost meritev – pri uporabi obeh vrst filtrov je pri tem nižji frekvenci potreben tem daljši čas za ustalitev merjene vrednosti. Pri frekvenci 4 Hz tako zadošča 5 s premor, pri 1 Hz moramo počakati vsaj 10 s, pri 0,2 Hz

Tabela 1: Rezultati meritev.

	Filter		Premor (s)	Protokol	Standardna negotovost (nV/V)		Zadovoljiva negotovost
	Vrsta	Frekvenca (Hz)			+ polariteta	- polariteta	
1.	Butterworth	4,0	5	A-POZ	3,5	-	☑
2.	Butterworth	1,0	5	A-POZ	213,4	-	☒
3.	Butterworth	1,0	10	A-POZ	2,7	-	☑
4.	Butterworth	0,2	20	A-POZ	3,5	-	☑
5.	Butterworth	0,1	30	A-POZ	3,6	-	☑
6.	Butterworth	0,1	30	A-NEG	-	3,2	☑
7.	Bessel	0,1	20	A-POZ	106,7	-	☒
8.	Bessel	0,1	30	A-NEG	-	1,4	☑
9.	Bessel	0,1	30	A-ALT	3,1	2,4	☑
10.	Bessel	0,1	40	A-ALT	2,1	7,2	☑
11.	Bessel	0,1	60	A-ALT	2,5	2,9	☑
12.	Bessel	0,1	30	A-POZ	1,8	-	☑
13.	Bessel	0,1	30	A-POZ	1,6	-	☑
14.	Bessel	0,1	30	A-POZ	3,6	-	☑
15.	Bessel	0,1	30	A-POZ	1,5	-	☑
16.	Bessel	0,1	30	A-POZ	1,6	-	☑
17.	Bessel	0,1	30	A-NEG	-	0,8	☑
18.	Bessel	0,1	30	A-NEG	-	1,0	☑
19.	Bessel	0,1	30	A-NEG	-	1,7	☑
20.	Bessel	0,1	30	A-NEG	-	1,0	☑
21.	Bessel	0,1	30	A-NEG	-	1,2	☑
22.	Bessel	0,1	30	A-ALT	1,9	4,5	☑
23.	Bessel	0,1	30	A-ALT	2,1	5,4	☑
24.	Bessel	0,1	30	A-ALT	2,5	1,2	☑
25.	Bessel	0,1	30	A-ALT	2,0	1,6	☑
26.	Bessel	0,1	30	A-ALT	1,9	0,9	☑
27.	Bessel	0,1	30	R-POZ	3,0	-	☑
28.	Bessel	0,1	30	R-NEG	-	0,6	☑

20 s ter pri 0,1 Hz vsaj 30 s, da lahko zajamemo ustaljeno vrednost. Pri primerjavi rezultatov, pridobljenih z avtomatiziranimi (A-) ter ročnimi (R-) meritvami, ni opaziti razlik, vrednosti negotovosti pa so prav tako zadovoljive, zato proces avtomatizacije postopka preverjanja linearnosti sistema smatramo kot uspešnega.

5 Zaključek

Dosegli smo zastavljene cilje, poleg uspešne avtomatizacije sistem deluje enako dobro, kot če z njim upravljamo ročno. Dodana vrednost pri tem je seveda razbremenitev operaterja in možnost implementacije kompleksnejših ter dlje časa trajajočih merilnih protokolov – denimo, za izvedbo protokola »ALT«, pri uporabi 30 s premorov pred zajemom vrednosti, program potrebuje 40 minut.

Vidimo kar nekaj možnosti za izboljšavo. Program bi lahko nadgradili tako, da v določenem časovnem intervalu zajame večje število vrednosti ter na podlagi dovolj majhne razlike med njimi sam določi, da je vrednost na merilniku ustaljena in primerna za zajem. Poleg tega mora biti lista ukazov pripravljena vnaprej in je bila v našem primeru spisana ročno. Tako možnost izboljšave vidimo tudi v samodejnem generiranju liste ukazov ali pa možnosti, da program v naključnem vrstnem redu sam izbere vsako izmed možnih kombinacij stikal.

Literatura

[1] M. Hiti, “Simulator mostička z uporovnimi lističi za kombinatorno kalibriranje merilnih ojačevalnikov,” *Zbornik štiriindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2015,* 21. - 23. september 2015.

[2] M. M. Schäck, “Long term proven and optimized high-precision 225 Hz carrier frequency technology in a modern and universal data acquisition system,” *Journal of Physics: Conference Series*, 1065(4), str. 042035, 2018.

[3] D. R. White, M. T. Clarkson, P. Saunders in H. W. Yoon, “A general technique for calibrating indicating instruments,” *Metrologia*, vol. 45, str. 199-210, 2008.

[4] M. M. Schäck, “High-precision measurement of strain gauge transducers at the physical limit without any calibration interruptions,” *IMEKO 22nd TC3, 12th TC5 and 3rd TC22 International Conferences*. 2014.

Orodje za oddaljeno testiranje spletnih uporabniških vmesnikov

Jure Trilar[1], Vid Čermelj[1], Emilija Stojmenova Duh[1]

[1]Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška c. 25, Ljubljana
E-pošta: jure.trilar@fe.uni-lj.si

Remote online user interface testing tool

Abstract. *As the Covid-19 outbreak became an inevitable factor in our interface user-testing activities, we have devoted time and knowledge to implement an unmoderated remote online interface testing tool that corresponds with our requirements. In this article we describe inception, practical aspects of implementation, typical usage scenario and future prospects in development of such a tool. Special attention was given to producing mirrored copies of websites behind login forms to be included in interfaces to be tested with the tool.*

1 Uvod

Pandemija bolezni covid-19, ki se je v Sloveniji pojavila v drugem četrtletju 2020, nas je prisilila v razmislek, kako opraviti že načrtovane aktivnosti testiranja uporabniških vmesnikov, ki so do sedaj običajno potekala v laboratorijskem okolju skupaj z uporabniki. Ob ukrepih za preprečevanje okužb je prišlo spoznanje, da neposredni stiki z zunanjimi udeleženci testiranj v bližnji prihodnosti ne bodo mogoči zaradi spremenjenih družbenih navad in možnosti ponovitev valov okužb. Poleg tega pa neposredni stik z določenimi demografskimi skupinami, ki nas še posebej zanimajo, kot so npr. starejši, verjetno ne bo mogoč tudi v daljšem časovnem obdobju. Ne glede na to situacijo je potreba po izvedbi oddaljenih testov, ki omogočajo preprostejšo in udobnejšo udeležbo, evidentna zaradi po vsebini različnih, po strukturi pa podobnih in ponavljajočih se testiranj uporabniških vmesnikov [1].

Smiselna je investicija časa in znanja v to, da se vzpostavi infrastruktura za učinkovito oddaljeno testiranje uporabniških vmesnikov, ki bo obenem reševala še vrsto drugih težav, kot sta avtomatska obdelava anketnih podatkov brez ročnega vnašanja in natančno merjenje uporabniške izkušnje s standardnimi pristopi [2]. Pri tem obstaja tudi potencial za doseganje večjega vzorca testirancev. Vsekakor to prinaša določene metodološke zadržke, saj na primer ne bi mogli zaznati drugih, verbalnih in neverbalnih, signalov, prisotnih pri testiranju v laboratorijskem okolju, vendar so zadržki uravnoteženi s pomembnimi prednostmi takega merjenja uporabnosti, uporabniške izkušnje in obremenitev pri uporabi.

2 Izhodišča in izvedba orodja

V okviru raziskovalnega dela želimo testirati scenarije uporabe na posnetku portala eDavki [3] v dveh (originalnem in prilagojenem) uporabniških vmesnikih. Portal eDavki je javnosti dobro znana rešitev, ki ima pomembno vlogo v e-življenju in e-poslovnosti slovenskih državljanov in za katero verjamemo, da bo podala relevantne rezultate. Primerjava rezultatov bo pokazala uporabnost koncepta družinsko usmerjenega razvoja (angl. family-centered design), ki ga razvijamo [4, 5].

Pri merjenju uporabnosti, uporabniške izkušnje in obremenitve pri opravljanju nalog se poslužujemo standardiziranih načinov merjenja (čas opravljanja nalog, uspešnost opravljanja nalog) in standardnih vprašalnikov (SUS [6], UEQ [7], NASA-TLX [8]), ki se uporabljajo na tem področju. Kombinacija vseh teh mehanizmov za merjenje ni bila združena v katerem od komercialnih orodij, ki so že na voljo [9]. Poleg tega nekatera obstoječa orodja sicer podpirajo večjezičnost, podpora za slovenski jezik pa ni navedena pri nobenem, zato smo se odločili za lastno izvedbo orodja.

Zaradi posebnih potreb glede zadostne možnosti prilagoditev in omenjenih tehničnih omejitev glede potreb testiranja [10] uporabniškega vmesnika portala eDavki smo identificirali nekaj izhodišč:

– Nemoderirano testiranje na daljavo. Udeleženci samostojno rešujejo naloge, brez osebe, ki vodi testiranje. Pri tem je ključno, da ne morejo zaiti izven »območja testiranja«, obenem pa se ne smejo znajti v brezizhodni situaciji, ki jim bi onemogočala nadaljevanje oz. zaključek testiranja, v nasprotnem primeru je treba to zaznati in ustrezno označiti v podatkih.

– Testiranje temelji na reševanju nalog in standardnih meritvah, ki se nanašajo na uspešnost in čas reševanja naloge. Udeleženci prejmejo navodila, nato rešujejo serijo nalog. Naloge konkretnega testiranja si v vsakem od vmesnikov sledijo linearno in obsegajo prijavo v zaledni del portala eDavki, iskanje vloge za obročno plačilo davka v največ treh mesečnih obrokih za fizične osebe in njeno oddajo. V snovanju scenarija smo predvidevali, da uporabniki niso že vnaprej seznanjeni s postopkom, kot je to npr. možno pri bolj popularni vlogi za popravek dohodnine. Izboljšave so na mestih, ki niso predmet redne uporabe, temveč jih uporabniki obiščejo redko, npr. enkrat letno ali manj. S tem ni učinka učenja in je večja potreba po še bolj intiuitivni uporabi. Zavedamo se, da izbrani scenariji le delno naslovijo testiranje koncepta družinsko usmerjenega razvoja.

– Interaktivnost in prilagodljivost vmesnikov. Testirani posnetki vmesnikov morajo delovati kot v originalu. Omogočati morajo polno uporabniško izkušnjo, obenem pa mora obstajati možnost prilagoditve delovanja za testiranje posebnih mehanizmov prikaza in interakcije. V prvi iteraciji je bila rešitev razvita za namizne računalnike, z vmesniki zaslona, miške in tipkovnice, v prihodnosti pa bi bila pomembna izboljšava testiranje z mobilnimi napravami.

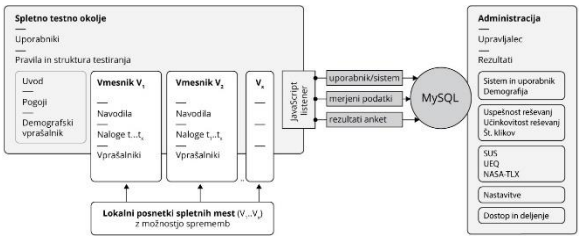
– Upoštevanje dobrih praks in standardnih postopkov testiranja uporabniških vmesnikov. Primer: izničiti je treba učinek učljivosti [2], zato se testirani vmesniki in naloge, kjer je to potrebno, pojavijo v naključnem vrstnem redu. Za testiranje uporabljamo standardizirane metrike s področja raziskovanja uporabniške izkušnje.
– Zasebnost in upoštevanje aktualnih pravil uporabe osebnih podatkov sta obvezna.
– Po-uporabljivost orodja za druge raziskovalce in druge projekte.

Navedena izhodišča so služila za prvotno načrtovanje okolja za oddaljeno testiranje in pomenijo prvotne okvire za izvedbo. To niso kritične omejitve in orodje bo v prihodnosti prilagojeno za morebitne nove potrebe.

2.1 Izvedba

Prototip orodja je izdelan s spletnimi jeziki HTML5, CSS3, Javascript in PHP z bazami MySQL, to je kombinacija, ki ustreza splošnim oblikam komercialnih gostovanj v Sloveniji. Orodje je sestavljeno iz treh funkcionalnih komponent (Slika 1):

– Uporabniški del je okolje, namenjeno testnim uporabnikom, ki jih povabimo k reševanju z enotnim URL-naslovom. V tem okolju uporabniki dobijo navodila za reševanje nalog, ki se prikazujejo na različnih testnih vmesnikih v naključnem vrstnem redu. Testni vmesniki, posnetki obstoječih spletnih strani ali namenski prototipi so »vstavljeni« v poseben element na spletni strani. Ob upoštevanju varne povezave (https) in povezanih priporočil [11, 12] glede uporabe zunanjih virov posameznih brskalnikih dobimo možnosti manipuliranja HTML-strukture z vsiljeno kodo Javascript. To je ključno za spremljanje ciljnih akcij, ki so praviloma v obliki klika na v nastavitvah določen element vmesnika. O uporabnikih zberemo splošne demografske podatke in poskušamo zagotoviti istovetnost posameznega reševalca skozi t. i. *device/browser fingerprint*, brez uporabe piškotkov brskalnika. To storimo za primer poskusov zlorabe in neprimernega vplivanja na rezultate.
– Posnetki vmesnikov in prototipni vmesniki so pomembni za kvalitetno izvedbo testiranja, zato temu namenjamo samostojno poglavje. V testno okolje so integrirani kot samostojne enote, ki omogočajo, da uporabnik vidi ter uporablja vmesnike na način, identičen originalom, kadar gre za posnetek obstoječe spletne strani oz. aplikacije, ali pa dejanski prototip brez večjih omejitev testiramo v istem okolju. Možno je testiranje več vmesnikov. Treba je opozoriti, da je kvaliteta posnetka odvisna od zunanjega orodja, ki ga uporabljamo za ta namen. Različna orodja namreč ponujajo različne pristope za zajem spletnih strani, portalov ali spletnih aplikacij.
– Administrativni del je na voljo le lastniku infrastrukture. Prikazuje vse zbrane podatke o reševanju nalog, izračunane rezultate vprašalnikov, omogoča nastavitve opravljanja nalog, shranjevanja podatkov in parametrov posameznih testnih vmesnikov. V začetni iteraciji še ni na voljo mehanizem za upravljanje z uporabniki. Prav tako še niso na voljo vse načrtovane funkcionalnosti. V prihodnosti bo tako na voljo tudi priporočilno orodje, ki zbrane podatke posameznih vmesnikov primerja glede na določene pristope k zasnovi uporabniških vmesnikov. Dodatne funkcije in izpisi rezultatov se bodo dodajali glede na bodoče potrebe.



Slika 1. Shema zasnove

2.2 Namestitev

Testno okolje je sestavljeno iz najbolj pogoste konfiguracije slovenskih komercialnih ponudnikov gostovanj – t. j. Apache, PHP, MySQL na strežniški strani. Testno okolje je moč vzpostaviti tudi na lastni infrastrukturi.

Testne vmesnike v obliki posnetkov obstoječih strani ali prototipov prenesemo v relevantno podmapo v direktorijski strukturi. Namestimo lahko enega ali več testnih vmesnikov – če jih je več, uporabnik testira serijo vmesnikov v naključnem vrstnem redu zaradi izogibanja učinku priučljivosti.

Posebno pozornost je treba nameniti nastavitvam, ki jih trenutno shranjujemo v naboru parametrov. Nastaviti je treba pot do posameznega vmesnika v direktorijski strukturi. Določimo serije ciljnih akcij – praviloma gre za klike na določene elemente v testiranih vmesnikih. V testnih vmesnikih moramo ročno poimenovati takšne elemente z oznakami razredov (angl. class) ali identifikatorji (angl. ID). Pri tem določimo tudi opis s ciljno akcijo povezane naloge in polje shranjevanja v bazi MySQL. Nastavimo tudi čas, po izteku katerega se uporabniku omogoči preskočiti nalogo, in pot nadaljevanja testiranja v naslednjem koraku. Čas nastavimo glede na lastno oceno kompleksnosti vmesnika in zgornje meje, za katero menimo, da začne pri uporabniku povzročati velike težave z upravljanjem vmesnika. Vsekakor ima uporabnik še vedno možnost zaključiti ciljno akcijo naloge; če se izgubi ali meni, da je naloga pretežka, pa jo preskoči in v bazo se doda zapis o neuspešnosti opravljanja naloge.

3 Posnetki obstoječih uporabniških vmesnikov

Ključna komponenta za izvedbo testiranja obstoječih uporabniških vmesnikov so t. i. posnetki spletnih strani (angl. website mirrors), ki omogočajo enako uporabniško izkušnjo kot original, vendar delujejo popolnoma samostojno. Omogočajo testiranje kopij vmesnikov, novih funkcij na teh vmesnikih v testnem okolju in testiranje samostojnih prototipnih rešitev – ter neposredno primerjavo med različnimi vmesniki, kot na

primer pri t. i. testiranju A/B [13]. Za pridobitev posnetkov obstoječih spletnih strani je treba previdno izbrati pravi način izvedbe, saj hitro naletimo na vrsto tehničnih in drugih izzivov. Za avtomatiziran zajem spletnih vmesnikov praviloma uporabljamo t. i. pajke, ki jih lahko naredimo sami, lahko pa se poslužimo že izdelanih rešitev.
Pri izbiri orodja oziroma pisanju programa za pridobivanje posnetkov spletnih strani je treba posebno pozornost nameniti robustnosti in vljudnosti spletnega pajka.
Robustnost: Med prehajanjem po spletnih straneh se lahko pajek ulovi v past (angl. spider trap). S tem se lahko zaplete v neskončno indeksiranje istih strani znotraj neke domene. Vse pasti niso nujno zlonamerne, do zapletanja lahko pride tudi v primeru, ko spletna stran ni bila pravilno oblikovana.
Vljudnost: Spletni strežniki imajo različne politike glede spletnih pajkov. V njih so zapisana različna priporočila, ki jih je treba pri obisku in zajemu teh strani upoštevati. Tudi če spletne strani nimajo definirane politike za spletne pajke, je priporočljivo, da se držimo dobrih praks [14].

3.1 Dobre prakse uporabe spletnih pajkov

Pri uporabi spletnih pajkov obstajajo dobre prakse, ki se jih držijo vsi največji iskalniki in so implementirane tudi v večini programov za ekstrakcijo podatkov s spleta. Po potrebi jih je mogoče vklopiti ali izklopiti. Razdelimo jih lahko v tri sklope pravil:
Spletni pajek se mora predstaviti kot spletni pajek in se ne sme pretvarjati, da je navaden uporabnik.
Spletne strani beležijo število obiskovalcev, poleg tega pa jim to omogoči nadzor nad pasovno širino, ki jo namenijo spletnih pajkom.
Spletni pajek mora spoštovati pravila datoteke robots.txt, ki določa, do katerih strani ima pajek dostop.
Upravljavci s tem sporočijo, na katerih straneh nočejo imeti spletnih pajkov, in tudi posredno potrdijo, da so spletni pajki na preostalih straneh v isti domeni dovoljeni. Ta datoteka pajku ne prepreči dostopa in jo lahko ignorira, vendar je pri dostopanju do takih spletnih strani treba biti pozoren, saj lahko vsebujejo podatke, ki jih je prepovedano hraniti (osebni podatki, avtorska dela). Če nas podatki na teh straneh ne zanimajo, je bolje, da jih ne shranjujemo.
Spletni pajek na posamezni spletni strani ne sme porabiti preveč pasovne širine.
To pomeni, da s posamezne strani ne prenaša več kot ene datoteke hkrati in da med posameznimi prenosi počaka nekaj sekund. Če shranjujemo podatke z različnih domen, je bolje, da vsakič obiščemo drugo spletno stran, ki jo izberemo z algoritmom krožnega izbiranja (angl. round-robin) [15, 16].

3.2 Izbira in uporaba orodja za posnetke uporabniških vmesnikov

Za potrebe izdelave posnetkov obstoječih spletnih strani smo iskali primerno obstoječe orodje [17]. Med vsemi, komercialnimi in odprtokodnimi, ki smo jih preizkusili, se je izkazalo, da našim potrebam najbolj ustreza odprtokodno orodje Httrack [18], ker ponuja najširši spekter možnih nastavitev pri tem opravilu, izdeluje uporabne posnetke, ki dobro delujejo v nepovezanem (*offline*) načinu, ter, najpomembnejše, omogoča snemanje strani, ki so dostopne le ob prijavi uporabnikov v zasebne dele strani, kar je ključno pri pridobivanju kopij spletnega portala eDavki.
Nabor možnosti pri upravljanju te rešitve je izredno dolg in podrobno opisan na straneh z dokumentacijo [19]. Posebej pa je treba omeniti možnost, da po prijavi z uporabniškim imenom in geslom v zasebne dele spletnih portalov izvozimo ustrezne piškotke. Te, shranjene v formatu Netscape, namestimo v ciljno mapo in orodje Httrack samodejno upošteva prijave, ki so dostopne prek piškotkov, ter naredi kopije zasebnih delov strani. Od upravljalca portala eDavki smo dobili dovoljenje za pridobitev posnetkov zasebnega dela uporabniškega vmesnika za potrebe raziskave, v okviru katere bomo v izvajali testiranje uporabnosti koncepta družinsko usmerjenega razvoja.

4 Uporaba orodja in prikaz rezultatov

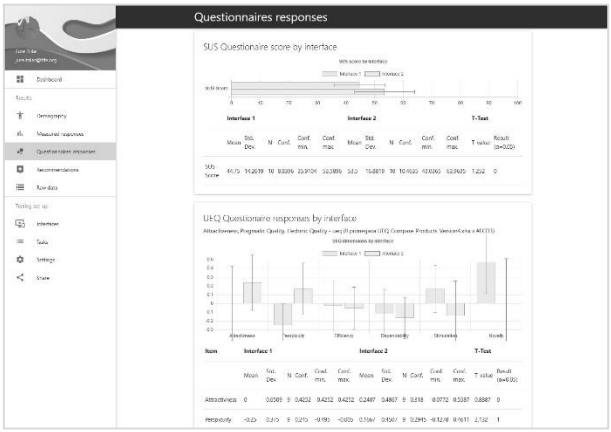
4.1 Tipični scenarij uporabe

Po namestitvi okolja in nastavitvi različnih parametrov vmesnika z enotnim URL-naslovom povabimo udeležence k sodelovanju. Vabilo diseminiramo v skladu s cilji raziskave in želenim ciljnim vzorcem.
Uporabnikom, ki sledijo povezavi, se v začetnih navodilih predstavijo cilji in namen raziskave, odgovorna oseba in kateri podatki se zbirajo. Po GDPR je shranitev podatkov o internetnem naslovu (angl. IP address) ter podpisu naprave in brskalnika (t. i. tehnika *browser/device fingerprinting*) dovoljena [20] le ob izrecnem dovoljenju testiranega uporabnika; čeprav se na prvi pogled zdi, da takšni podatki ne vsebujejo osebnih podatkov.
Testiranje oz. beleženje se začne po dovoljenju za zbiranje podatkov. Ob tem se označi čas začetka testiranja. Uporabnik nato reši preprost demografski vprašalnik, ki se nanaša na cilje raziskave.
Sledijo navodila za reševanje serije nalog. Opis trenutne naloge je prikazan v vmesniku. Naloga velja za uspešno rešeno, če uporabnik izvede ciljno akcijo – ki je lahko kjerkoli v strukturi vmesnika. Če od začetka reševanja naloge preteče arbitrarno določen čas, se uporabniku ponudi možnost, da trenutno nalogo preskoči, sistem pa to ustrezno zabeleži – na ta način poskušamo povečati možnost, da uporabnik nadaljuje s testiranjem, in s tem zagotovimo večji nabor podatkov.
Na koncu vsakega sklopa nalog v vmesniku sekvenčno sledijo standardni vprašalniki o uporabniški izkušnji, uporabnosti sistema in obremenitvi pri opravljanju nalog. Sledi še zahvala testnim uporabnikom za njihov prispevek k raziskavi in zabeležba v bazi o končanju testiranja.

4.2 Izpis rezultatov

Kumulativne rezultate testiranj izpišemo v administratorskem vmesniku in obsegajo primerjave merjenih vrednosti med različnimi testnimi vmesniki na nivoju nalog, kjer prikažemo povprečno uspešnost in

povprečen čas reševanja posamezne naloge ter dodatno še povprečno število klikov ob tem.
Prikažemo tudi izračunane vrednosti vprašalnikov, ki temeljijo na subjektivni oceni uporabniške izkušnje, uporabnosti in napora pri reševanju nalog (Slika 2). Prikažemo jih na način, ki je lasten vsaki metodi, npr. porazdelitev po komponentah uporabniške izkušnje ali pa rezultat na enotni lestvici uporabnosti.
Izpis rezultatov je prilagojen trenutnim potrebam raziskave vmesnikov, ki temeljijo na portalu eDavki.



Slika 2. Zaledni vmesnik in prikaz rezultatov

5 Zaključek

Pri našem dosedanjem delu je testiranje vmesnikov potekalo v laboratorijskem okolju. Zato je treba obravnavati določene zadrege, ki nastanejo v moderiranem in (za nas novem) nemoderiranem načinu v zvezi z vplivi na rezultat. Obstaja možnost, da udeleženci prilagodijo svoje odgovore (npr. prava starost, digitalne veščine), če je prisotna oseba, ki vodi testiranje. V nemoderiranem načinu ne moremo učinkovito zajeti nekaterih drugih signalov, besedne in nebesedne komunikacije udeležencev, ki nam daje globlji, kvalitativni vpogled v uporabnost, uporabniško izkušnjo in obremenitve pri uporabi sistemov ter nam kadarkoli omogoča testiranje na večjem vzorcu. V to je treba vključiti razmisleke o kombiniranem pristopu, kjer s sledenjem kurzorju miške ali z uporabo video spletne konference med reševanjem nalog poskusimo zaznati dodatne značilnosti uporabe testiranih vmesnikov. Primerjava rezultatov med moderiranim in nemoderiranim pristopom s pomočjo manjše skupine testirancev bi lahko pokazala ustreznost ali pomanjkljivosti orodja za oddaljeno testiranje. Dodatna izboljšava bi bil razmislek glede tega, kako vizualno motivirati uporabnike za reševanje.
Razumljivo je, da testno okolje, ki je bilo izvedeno, ni konkurenca obstoječim komercialnim rešitvam – vsekakor pa dobro služi našim, bolj akademskim, potrebam pri oddaljenem testiranju uporabniških vmesnikov in ga bomo prilagajali oz. dopolnjevali glede na cilje prihodnjih raziskav.

6 Literatura

[1] ISO 9241-11:2018 – Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts.
[2] T. Tullis, B. Albert, Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics, Morgan Kaufmann, Amsterdam, 2016.
[3] eDavki. Dosegljivo: https://edavki.durs.si/ [6. 7. 2020].
[4] J. Trilar, A. Kos, S. Jazbinšek, L. Jensterle, in E. S. Duh, ICT to Promote Well-Being within Families, *Sensors*, vol. 18, no. 9, p. 2760, 2018.
[5] J. Trilar, V. Zavratnik, V. Čermelj, B. Hrast, A. Kos, in E. S. Duh, Rethinking Family-Centred Design Approach Towards Creating Digital Products and Services, *Sensors*, vol. 19, no. 5, p. 1232, 2019.
[6] K. Finstad, The Usability Metric for User Experience, *Interacting with Computers*, vol. 22, no. 5, 2010, str. 323–327, doi:10.1016/j.intcom.2010.04.004.
[7] B. Laugwitz, T. Held, and M. Schrepp, Construction and Evaluation of a User Experience Questionnaire, *Lecture Notes in Computer Science HCI and Usability for Education and Work*, pp. 63–76, 2008.
[8] A. Cao, K. Chintamani, A. K. Pandya in R. D. Ellis, NASA TLX: Software for assessing subjective mental workload, *Behavior Research Methods*, vol. 41, no. 1, pp. 113–117, 2009.
[9] Justinmind, Best 15 usability testing tools for happy users. Dosegljivo: https://www.justinmind.com/blog/usability-testing-tools/ [6. 7. 2020].
[10] Nielsen Norman Group, Tools for Unmoderated Usability Testing. Dosegljivo: https://www.nngroup.com/articles/unmoderated-user-testing-tools/ [6. 7. 2020].
[11] MDN Web Docs, Same-origin policy. Dosegljivo: https://developer.mozilla.org/en-US/docs/Web/Security/Same-origin_policy [6. 7. 2020].
[12] Wikipedia, Cross-origin resource sharing. Dosegljivo:https://en.wikipedia.org/wiki/Cross-origin_resource_sharing [6. 7. 2020].
[13] Optimizely, A/B testing. Dosegljivo: https://www.optimizely.com/optimization-glossary/ab-testing/ [6. 7. 2020]
[14] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008, pp. 405–420.
[15] B. Ribeiro, R. Baeza-Yates, *Modern Information Retrieval*, Upper Saddle River, NJ: Pearson Higher Education, 2010, pp. 515–544.
[16] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*, OReilly Media, Incorporated, 2018, pp. 263–278.
[17] Top 15 Website Ripper or Website Downloader Compared. Dosegljivo: https://prowebscraper.com/blog/top-website-ripper-or-website-downloader-compared/ [6. 7. 2020].
[18] HTTrack Website Copier, Dosegljivo: http://www.httrack.com/ [6. 7. 2020].
[19] HTTrack Help, Dosegljivo: http://httrack.kauler.com/help/Home [6. 7. 2020].
[20] K. Szymielewicz., B. Budington, The GDPR and Browser Fingerprinting: How It Changes the Game for the Sneakiest Web Trackers, Electronic Frontier Foundation. Dosegljivo: https://www.eff.org/deeplinks/2018/06/gdpr-and-browser-fingerprinting-how-it-changes-game-sneakiest-web-trackers [6. 7. 2020].

Električno merjenje penetracije kaplje v porozni podlagi

Marion Thébault [1], Franco Policardi [2, 3]

[1] *Kompetenzzentrum Holz (Wood K Plus), Altenberger Straße 69, 4040 Linz, Austria, c/o: Wood Carinthian Competence Center (W3C), Klagenfurter Straße 87-89, 9300 Sankt Veit an der Glan, Austria*

[2] *LUCAMI, Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1000 Ljubljana, Slovenija*

[3] *Raziskovalni Inštitut za nove tehnologije in energetiko (R.I.N.T.E.) d.o.o., Stritarjeva 6/a, 4000 Kranj, Slovenia*

m.thebault@wood-kplus.at

Electrical measuring apparatus for the penetration of liquid droplets in a porous substrate

The study of liquid penetration in porous substrates is an important subject interesting different industries, since many manufactured products implies impregnation or printing processes. The ability of some liquids to enter specific materials, as for example papers, hardboards and tissues depends on several chemical and physical properties, resulting complex to characterize. In recent years, a special device was developed and patented [1] in Kompetenzzentrum Holz (Wood K Plus) in Linz, Austria, which describes the passage of a drop through a small piece of substrate via electrodes. In this article we present the equipment, its workflow, an example of the generated electrical signal and some possible interpretations. Different issues are still to be solved and future perspectives will be mentioned too.

1 Uvod

Študija vdora tekočine v porozne podlage je pomembna in zanimiva za različne industrije, saj mnogi industrijski procesi predvidevajo postopke impregnacije ali tiskanja. Sposobnost nekaterih tekočin, da vstopajo v določene materiale, na primer papirje, trde plošče in tkiva, je odvisna od več kemijskih in fizikalnih lastnosti, kar je posledično zapleteno karakterizirati. V zadnjih letih smo v Kompetenzzentrum Holz (Wood K Plus) v avstrijskem Linzu razvili in patentirali posebno napravo imenovano Penetrazionszeitmessungsaparat [1], ki omogoča opis prehoda kapljice skozi majhen kos podlage preko elektrod. V tem prispevku bomo predstavili opremo, njen potek dela, primer generiranega električnega signala in njegove možne interpretacije. Nekatera vprašanja je še potrebno rešiti in omenili bomo tudi prihodnje perspektive.

2 Problem

Na prodor tekočin skozi porozne podlage vplivajo različni parametri in pojavi, odvisno od sistema tekočina-substrat [2]-[6]. To področje je zelo pomembno za tiskanje papirja, za črnila, za industrijo tkanine in trdega kartona, saj mora tekočina vstopati le v pore površinske podlage in se v njej ne širiti [7], [8].

Izdelki iz industrije laminatov trdih plošč so v glavnem izdelani iz več prekrivnih papirjev, impregniranih s smolo, ki se strdi pod visokim in specifičnim tlakom in temperaturo [8]-[10]. Homogene in hitre impregnacijske zmogljivosti so v teh primerih zelo pomembne in neizogibne lastnosti [2].

Raziskovalci tovrstne težave že vrsto let preučujejo in trenutno obstajajo različne metode za označevanje vpijanja tekočine, vendar so večinoma optične [11]. J. A. Bristow je leto 1967 razvil neoptični test absorpcije tekočine v papirju za švedsko papirno industrijo z aparatom, imenovanim "bristowsko kolo" "*the Bristow Wheel*" [12]. Ta naprava meri dolžino impregnacije na vzorcu papirja za določen volumen tekočine z različnimi hitrostmi.

Vsekakor mora industrija najti kompromis pri nadzoru kakovosti med preprostostjo, hitrostjo, ustreznostjo informacij in natančnostjo, zato je bilo treba izboljšati to standardizirano metodo.

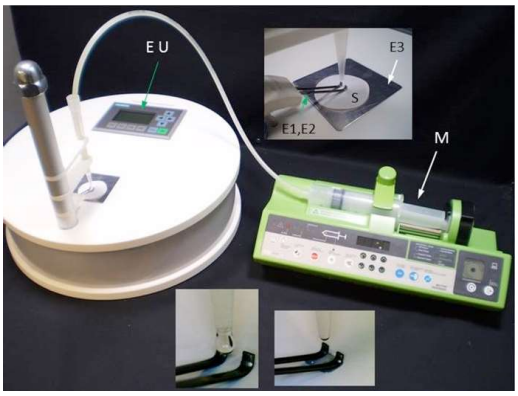
3 Razvoj aparature

Zaznavanje tekočin z uporabo elektrod je star koncept; različne uporabljene metode so počasi postale vedno bolj specifične in natančne na področju kromatografije [13], [14] in mikrofluidike [15], [16]. Posebna naprava, ki temelji na meritvah električne prevodnosti med zgornjo in spodnjo površino absorpcije in prodiranja tekočine v papir, je bila patentirana že leta 1960 [17]. Ta specifična aparatura je bila zanimiva zaradi svoje preproste uporabe, vendar je potrebovala izboljšave. V našem Kompetenzzentrum Holz (Wood K Plus) in Linz, Austria, smo razvili in patentirali napravo [1], ki odkriva, identificira in meri prehod kapljice skozi majhen kos podlage preko elektrod.

Aparat je sposoben meriti impregnacijo porozne podlage s prevodno kapljevino, tudi s tekočo smolo. Ta preprosta metoda je bila prvotno zasnovana za hitro in objektivno merjenje časa, ki ga potrebuje kapljica kapljevine oz. smole, da prodre na tanek list papirja. Razvoj časovne penetracije je še en pomemben parameter, ki ga naprava lahko zazna in prikaže. Napravo smo uporabili tudi za proučevanje prodiranja fenolnih smol v Kraft papir [18].

4 Aparatura

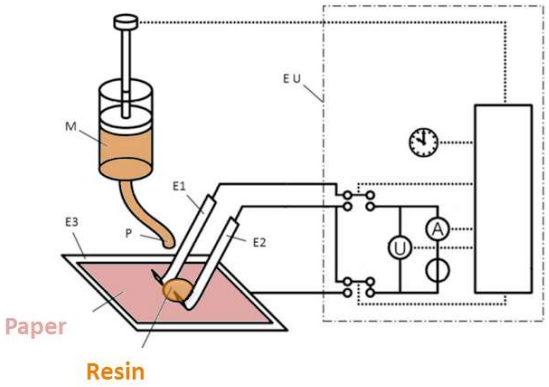
Naprava je trda valjasta podlaga, opremljena s tremi elektrodami. Dve enaki ukrivljeni elektrodi E1 in E2 sta zgoraj, med kateri odložimo kapljico kapljevine ročno, ali z uporabo avtomatskih sistemov, opremljenih z brizgalkami (M). Tretja elektroda (E3) je večja plošča, vodoravno pritrjena na nosilec oz podlago (S). Elektrode so povezane v elektronsko enoto (EU), vgrajeno v podlago in sestavljeno iz Siemensovega LOGO! 8 napajalnega sistema DC24V ter osnovnega modula 12/24RCEO. Meritve spremljamo preko vmesnika v računalniku, ki je z napravo povezan preko Ethernet kabla in s pomočjo programske opreme Siemens LOGO! Confort Version 7.1. Za depozicijo kapljice uporabljamo sistem rednega pritiska na brizgo, ki vsebuje tekočino (M).



Slika 1. Fotografija naprave opremljena s pipeto, s snovjo in z vzorcem lista papirja.

5 Delovanje aparature

Tri elektrode merijo dva tokovna signala (dve prevodnosti) na obeh straneh lista papirja (slika 2).

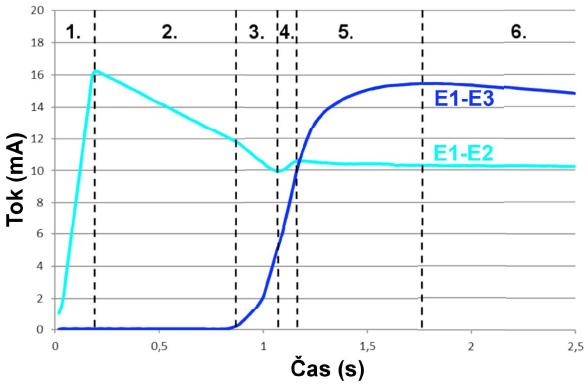


Slika 2. Shema naprave; EU elektronska enota; E1 in E2 zgornji elektrodi; E3 spodnja elektroda; R kapljica tekočine; S podlaga; M brizga s tekočino.

Prvi signal dobimo med zgornjima elektrodama E1 in E2 na isti strani papirja, na kateri kapne kaplja tekočine; drugi signal dobimo med zgornjo elektrodo E1 in spodnjo elektrodo E3. Oba električna signala sta časovno registrirana z ločljivostjo 2 ms in 0,02 mA.

6 Interpretacija signalov

Ko se kapljica tekočine dotika elektrod, se sprožita dva signala, katera lahko časovno interpretiramo tako, da ločimo različne korake.



Slika 3. Značilen dvojni električni signal, ki ga ustvarja kapljica npr. fenolne smole, ki prodira na kos papirja.

Slika 3 prikazuje primer dvojnega električnega signala med elektrodama E1-E2 (svetlo modro) in med elektrodama E1-E3 (temno modro), ustvarjenega s kapljico fenolne smole na posebnem Kraft papirju [18]. Glede na potek obeh signalov lahko identificiramo najmanj šest različnih časovnih segmentov.

Prvih 0,2 s prikazuje čas odlaganja kaplje med elektrodi E1-E2 (faza 1), kjer se tok zaradi električne poti, ki jo povzroči kaplja, nenadoma poveča do več kot 16 mA (papir je precej električno izolacijski material). Tok se potem začne zmanjševati (faza 2), ker se kaplja vodoravno širi in/ali prodira v debelino papirja navpično in s tem zmanjšuje volumen prevodne poti med E1 in E2. V tretjem koraku se pojavi tok med E1 in E3, kar pomeni, da je prevodna pot vzpostavljena, in sicer je tekočina dosegla spodnjo elektrodo; ta drugi signal se začne povečevati, ko se tekočina širi v material, ki ločuje elektrode (faza 3).

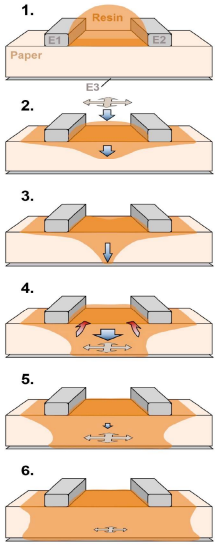
Zanimivo je omeniti, da se v fazi 3 tok med E1 in E2 hitreje zmanjšuje, kar bi pomenilo, da je tekočina z navlažitvijo spodnje elektrode ustvarila rahlo depresijo, ki tekočino potegne navzdol. Zmanjšani volumen kaplje povzroči opazno zmanjšanje toka. Nato se kompenzira, očitno s kapilarnostjo navzgor, zaradi česar se signal E1-E2 v določeni meri znova poveča, medtem ko se prevodna pot med E1 in E3 nenehno povečuje, ko se smola še naprej vpija v substrat (faza 4). Potem se tok E1-E2 stabilizira: kapljevina se na zgornji strani papirja ne širi več. Poleg tega se tok širjenja smole med E1 in E3 začne upočasnjevati in signal doseže največjo vrednost okoli 15,6 mA (faza 5). Nazadnje (faza 6), se tekočina še naprej vodoravno širi v papir, pri čemer v določeni meri ohranja prevodni volumen med E1 in E3, zaradi česar se signal še zmanjšuje.

Evolucija dvojnega signala se razlikuje od enega sistema tekočina-podlaga do drugega. Zgoraj predstavljene različne faze se lahko razlikujejo zaradi količine tekočine v naloženi kapljici, prevodnosti materiala, lastnosti površinske napetosti itd.. Prav tako je obnašanje tekočine odvisno od značilnosti podlage, kot so npr. njena makro, mezo in mikro poroznost, površinske napetostne lastnosti (polarne in disperzivne), njegova debelina itd.. Zaradi vsega omenjenega je ta metoda za sedaj omejena na primerjavo sistemov tekočina-podlaga iste narave.

Za pridobitev objektivnih numeričnih podatkov je mogoče časovne vrednosti izračunati iz točk prekrivanja krivulj. Zakasnitve med njimi sporočajo hitrost impregnacije; razlike med prevodnostjo med zgornjo in spodnjo stranjo so značilne za vodoravne in navpične širitve tekočine.

7 Postopek impregnacije

Razlaga 6 korakov impregnacije v sistemu papir-smola je prikazana na sliki 4. Poudarjamo, da je oblika meje tekočine v papirnem substratu le ilustrativna.



Slika 4. Razlaga različnih korakov širjenja tekočine v substratu. Oblika meje tekočine v papirnem substratu je le ilustrativna.

8 Aktualna vprašanja in perspektive

Glede na nekaj primerov meritev penetracije melamin-formaldehidne smole v dekorativnem papirju se pojavlja vprašanje ponovljivosti

preizkusov, predvsem kadar tekočine vpijajo porozne medije počasi. Številni razlogi lahko problematiko pojasnjujejo, predvsem vezani na dejstvo, da se med penetracijo pojavijo številni mikroskopski pojavi, medtem ko ta metoda predstavlja karakterizacijo v mezoskopskem merilu. Na eni strani papir ni mikroskopsko homogen: tekočina gre skozi debelino v funkciji lokalnih prepletov vlaken, ki jih poganjajo gravitacijske sile, kapilarnost in premiki [1], [19]. Poleg tega se namazana površina razlikuje med lokalnimi hrapavostmi, gladkostjo in poroznostjo verjetno od vzorca do vzorca. Tako v praksi kaplje iste tekočine ne zmočijo iste površinske podlage, čeprav sta obe iste narave. Po drugi strani bi lahko tekočina vsebovala nekaj delcev ali koloidov, kar se pogosto zgodi v smolah z določeno viskoznostjo. Njihov vpliv na prodor je še vedno težko določiti.

Nekatere karakteristike prevodnosti so torej odvisne od narave tekočine in na primer elektrode naprave ne morejo zaznavati čistega metanola. V zvezi s kontrolo količine tekočine za merjenje so tudi velikosti kapljic odvisne od njihove hidrofilnostne zmožnosti z zrakom in z brizgo, ki jih spusti. Lahko tudi štejemo, da zgornje elektrode pritiskajo na vzorec in s tem v določeni meri spremenijo hidrofilnost s tekočino ter lokalno poroznostjo.

9 Sklep

Aparatura predstavljena v tem članku, temelji na dveh sočasnih meritvah električnih tokov, ki jih povzročajo kapljice tekočine med prodiranjem v porozno matriko substrata, nameščena med tremi elektrodami. Evolucija obeh signalov že omogoča nekaj razlag o specifičnem penetracijskem obnašanjem in informira o sposobnosti impregnacije nekaterih tekočin.

Literatura

[1] S. Hochsteiner, S. Lenz, and J. Stultschnik, “Measuring Device and Measuring Method for Measuring the Resin Impregnation of a Substrate,” patent AT517076 (A4), 2016

[2] M. Thébault, A. Kandelbauer, U. Müller, E. Zikulnig-Rusch, and H. Lammer, “Factors influencing the processing and technological properties of laminates based on phenolic resin impregnated papers,” *Eur. J.* Wood Wood Prod., vol. 75, no. 5, pp. 785–806, Sep. 2017

[3] M. Blunt, M. J. King, and H. Scher, “Simulation and theory of two-phase flow in porous media,” *Phys. Rev. a*, vol. 46, no. 12, pp. 7680–7699, 1992

[4] W. M. Mahmud and V. H. Nguyen, “Effects of snap-off in imbibition in porous media with different spatial correlations,” *Transp. Porous Media*, vol. 64, no. 3, pp. 279–300, 2006

[5] A. Clarke, T. D. Blake, K. Carruthers, and A. Woodward, “Spreading and imbibition of liquid droplets on porous surfaces,” *Langmuir*, vol. 18, no. 8, pp. 2980–2984, 2002

[6] R. Lenormand, E. Touboul, and C. Zarcone, “Numerical models and experiments on immiscible displacements in porous media,” *J. Fluid Mech.*, vol. 189, pp. 165–187, 1988

[7] J. F. Oliver, L. Agbezuge, and K. Woodcock, “A diffusion approach for modelling penetration of aqueous liquids into paper,” *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 89, no. 2–3. pp. 213–226, 1994

[8] R. Bollström *et al.*, “Printability of functional inks on multilayer curtain coated paper,” *Chem. Eng. Process. Process Intensif.*, vol. 68, pp. 13–20, Jun. 2013

[9] W. E. Morris and J. M. Verdi, “Decorative core stock laminating varnishes and phenolic resin core and decorative laminates produced therefrom,” US2801198A, 1957.

[10] S. E. Palazzolo and H. O. Mccaskey, “Modified phenolic resin system and laminates made therefrom,” US3707296A, 1970.

[11] J. Salvatore E. Palazzolo, Harold O. McCaskey, “High yield saturating paper,” US 4060450 A, 1977

[12] J. A. Bristow, “Liquid absorption into paper during short time intervals,” *Sven. Papperstidning-Nordisk Cellul.*, vol. 70, no. 19, pp. 623–629, 1967

[13] R. P. Baldwin and K. N. Thomsen, “Chemically modified electrodes in liquid chromatography detection: A review,” *Talanta*, vol. 38, no. 1, pp. 1–16, Jan. 1991

[14] L. Gorton *et al.*, “Selective detection in flow analysis based on the combination of immobilized enzymes and chemically modified electrodes,” *Anal. Chim. Acta*, vol. 250, pp. 203–248, Oct. 1991

[15] J. Collins and A. P. Lee, “Microfluidic flow transducer based on the measurement of electrical admittance,” *Lab Chip*, vol. 4, no. 1, pp. 7–10, 2004

[16] S. Bohm, J. I. Rodgers, A. McNeilage, J. Moffat, M. Stiene, and T. Richter, “Microfluidic analytical system with position electrodes,” US6990849B2, 2006

[17] E. G. Emmert and S. A. Torstenson, “Apparatus for determining absorption and liquid penetration of paper,” US2931977A, 1960

[18] M. Thébault, A. Kandelbauer, E. Zikulnig-Rusch, R. Putz, S. Jury, and I. Eicher, “Impact of phenolic resin preparation on its properties and its penetration behavior in Kraft paper,” *Eur. Polym. J.*, vol. 104, pp. 90–98, Jul. 2018

[19] R. Lenormand, C. Zarcone, and A. Sarr, “Mechanisms of the displacement of one fluid by another in a network of capillary ducts,” *J. Fluid Mech.*, vol. 135, pp. 337–353, Oct. 1983

Primerjava termovizijskih kamer za merjenje telesne temperature pri ljudeh

Vid Mlačnik [1], Igor Pušnik [1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: vid.mlacnik@gmail.com, igor.pusnik@fe.uni-lj.si

Comparison of thermal imagers for measurement of human body temperature

Abstract
In this paper, application of uncooled thermal imaging for human body temperature screening is evaluated in terms of accuracy and reliability. Accuracy of thermal imaging and associated mathematical model is analyzed and assessed for commercially available products, taking into account both the measurement instrument and measured subject.

Results of accuracy analysis, gathered to the extent of available data, are added to common accuracy of the core body temperature measurement and assessment process. Theoretical combined uncertainty of the measurement is considerably higher than required by medical profession for fever screening, especially before calibrations were carried out.

1 Uvod

Že od izbruha SARSa leta 2002, predvsem pa med aktualno Covid-19 pandemijo se na trgu pojavlja množica novih izdelkov za omejevanje širjenja nalezljivih bolezni z uporabo termovizijskih kamer. Takšne rešitve se predvidoma uporabljajo na mestih množičnih vstopov v npr. tovarne, letališča, trgovine, zdravstvene ustanove, itd, kjer se dobro odražajo prednosti tovrstnih izdelkov, to so hitrost meritve, samodejen zajem in obdelava meritev ter brezkontaktno delovanje termovizije. Problem brezkontaktnega merjenja pa vsekakor predstavlja točnosti in negotovost meritve, ki je teoretično višja kot pri kontaktnih metodah merjenja, kar močno pogojuje uporabnost in zanesljivost tovrstnih inštrumentov.

V praksi med nakupom pri večini izdelkov kupca pričakajo enake, optimistične trditve o točnosti meritev, razlike pa se opazijo šele po natančnem preskusu izdelka z laboratorijsko opremo.

Natančnost in točnost tehnologije brezkontaktnega merjenja telesne temperature je odvisna od mnogih vplivnih veličin, vsaka od njih pa v neki meri prispeva k končnemu rezultatu. Za ovrednotenje točnosti merilnega sistema je pomembno ugotoviti vse vplivne veličine in preveriti njihov vpliv na negotovost končnega rezultata. Pri tem je vredno poudariti, da velik doprinos k negotovosti meritve predstavlja slabo poznavanje lastnosti človeka.

Da bi preverili točnost tehnologij brezkontaktnega merjenja telesne temperature z uporabo termovizije, smo v Laboratoriju za metrologijo in kakovost Fakultete za elektrotehniko UL, kot nosilcu nacionalnega etalona za termodinamično temperaturo izvedli ovrednotenje nekaj izdelkov različnih cenovnih razredov, merilne rezultate pa smo ovrednotili in primerjali.

2 Fizikalni model merjenja

Idealno črno telo s temperaturo nad 0 K v skladu s Stefan-Boltzmannovim zakonom oddaja toplotno sevanje sorazmerno četrti potenci absolutne temperature objekta. Dejanski tok sevanja toplotne energije nekega sivega telesa je pogojen z emisivnostjo površine tega objekta, torej je dejanski sevalni tok za faktor emisivnosti ϵ nižji od sevanja idealnega črnega telesa pri enaki temperaturi in spektralnih ter položajnih pogojih. Na poti proti merilnemu inštrumentu se izsevanemu signalu lahko pridruži tudi odbiti signal (enačba 1), kot posledica vpadnih signalov zaradi odbojnosti objektov v okolici. Odbojnost se običajno preračuna iz izmerjenih in vnesenih parametrov okolice (emisivnost, temperatura ambienta, vlaga, idr.), ki je predpostavljeno homogena.

$$S_{vir} = \epsilon \cdot S_{obj} + (1 - \epsilon) \cdot S_{amb} \qquad (1)$$

Odbiti signal je odboj sevanja objektov iz okolice, ki se pri odboju popolno ali spektralno difuzno odbije od opazovanega objekta z $(1 - \epsilon)$ vpadne jakosti.

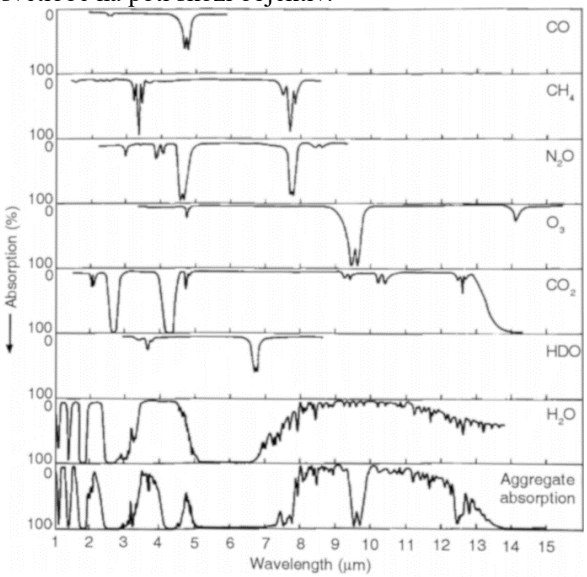
Na prenosni poti nastopa atmosfera, ki opisano sevanje vzdolž prepotovane razdalje običajno filtrira. Zrak zaduši določene valovne dolžine, ki so specifične za nastopajoče pline in vodno paro. Absorpcijo atmosfere vzdolž določene poti imenujemo atmosfersko okno, vidno na sliki 1, lastnost prepuščanja sevanja pa transmisivnost τ, opisano s skalarjem med 0 in 1 (enačba 2).

Kot polprozoren medij lahko obravnavamo prenosno pot od merjenega objekta do senzorja, kar zajema zrak in lečo.

$$S_{kamera} = \tau_{prenosna_pot} \cdot S_{vir} \qquad (2)$$

Signal po prehodu skozi zrak preko leče termovizijske kamere vpade na senzor in mu preda toploto. Na vpadlo sevanje na senzor vpliva še popačitev na leči - tako geometrijska popačenja, kot absorpcija sevanja, pa tudi difuzija in nepravilnosti ostrenja zaradi širokega razpona

valovnih dolžin, ki različno vpliva na lomne kote IR svetlobe na poti skozi objektiv.



Slika 1: Absorpcijski spekter atmosfere

Opisano sevanje v merilnem inštrumentu vpade na senzor toplote. Najpogosteje implementirani nehlajeni senzor toplote je mikrobolometer, čigar slikovni elementi so osnovani na substratu s temperaturno odvisno upornostjo. Temperatura slikovnega elementa se spreminja glede na vpadlo in izsevano toplotno energijo, ta pa je odvisna od sevanja ospredja in ozadja senzorja. Pri znanih parametrih sevanja ozadja je mogoče iz obnašanja slikovnih elementov izračunati vpadlo sevanje ospredja meritev za izračun sevalne temperature merjenca.

3 Merjenec in biologija

Za kvalitetno odločanje v zdravstvu kot enega izmed parametrov potrebujemo zanesljivo meritev telesne temperature. V zdravstvu je zaželena negotovost temperaturnih inštrumentov nižja ali enaka 0,3 °C, kar je z uporabo termovizije praktično nemogoče.

Površinska temperatura kože je pri ljudeh močno nehomogena. S stališča preprostosti uporabe so najprimernejša mesta merjenja odkriti deli kože kot sta obraz in vrat. Temperatura na mestu meritve mora biti čim manj odvisna od temperature ambienta in delovanja človeške termo-regulacije, kar pomeni da je primerno merilno mesto dobro prekrvavljeno in slabo toplotno izolirano. V okolici obrazne in temporalne arterije je najprimernejše merilno mesto v bližini notranjih očesnih kotov (slika 2), saj je tam temperatura najmanj pogojena s temperaturo okolice, poleg tega pa tam skladno z željami izmerimo najvišjo temperaturo. Zanesljivost vseh merilnih mest je poleg tega pogojena tudi z variabilnostjo postavitve žil pri merjenih osebkih [1].

Vsem merilnim mestom je skupno, da ne odražajo prave notranje telesne temperature. K temu problemu pristopimo s preračunom na predvideno telesno temperaturo, za kar uporabimo matematični model, ki so ga Ng, Kawb in Chang [2] pridobili z regresijo linearne funkcije med izmerjeno telesno (izmerjeno z ušesnim termometrom) in površinsko sevalno temperaturo, izmerjeno s termovizijo. Model je prikazan v tabeli 1.

Tabela 1: Korelacija med merjeno površinsko in preračunano notranjo temperaturo pri ljudeh, pridobljena z regresijo linearne funkcije [2].

Območje	Oči (pravokotnik)
Neodvisna spr. (X)	Telesna temperatura
Odvisna spr. (Y)	Maks. meritev območja oči
Velikost vzorca	310 oseb
Korelac. koef. (r^2)	0,5509
St. odklon modela	0,8292 °C
Korelacijska funkcija	$X = 1{,}1752 \cdot Y - 4{,}9327$



Slika 2: Običajna postavitev obraznih žil

4 Določitev merilnih napak in negotovosti

Na prenosni poti nastopa mnogo predpostavljenih vrednosti, ki lahko v primeru odstopanj prispevajo k napaki meritve telesne temperature.

Prispevke k napaki zaradi geometrijskih popačenj slike se praviloma odpravi s kalibracijo kamere po izdelavi inštrumenta. Preostanek negotovosti v samem merilnem inštrumentu obravnavamo kot celoto in jo statistično izmerimo za dve lastnosti – homogenost slike in ponovljivost meritev kamere.

Ključno vlogo za jakost izmerjenega sevanja igrajo emisivnost kože, temperatura okolice v odsevu in temperatura ter sestava zraka na prenosni poti. Ti parametri so delno predvidljivi, odstopanja pa prispevajo

k napaki meritve. Idealno želimo v prostoru suh zrak in čim krajšo prenosno pot med merjencem in merilnim inštrumentom ter ozadje homogene temperature in emisivnosti.

Ne glede na atmosferske vplive pa oddaljenost merjenca vpliva na prostorsko ločljivost meritev oz. velikost ciljnega področja na sliki. Pomembno je, da merjeno področje zajema dovolj slikovnih elementov, da se pravilno izloči vpliv prostorske diskretizacije in prostorskega prehodnega pojava na toplotni sliki (t.i. size of source effect). Idealno merjenje se izvaja v fokusni ravnini termovizijske kamere, zaradi velikosti na sliki pa je zaželeno, da je oddaljenost merjenca čim nižja.

Na izmerjeno sevanje zaradi nehomogene smerne porazdelitve izsevane toplote vpliva tudi kot merjenja – idealno merimo površino pod pravim kotom, odstopanja od tega kota pa dodatno doprinesejo k napaki matematičnega modela pri obdelavi vpadle energije [3].

Celotno negotovost merilnega sistema izračunamo po enačbi (3), pri čemer upoštevamo občutljivostne koeficiente c_i in negotovost vplivnih veličin u_i.

$$u_{celotna} = \sqrt{\sum_i (c_i \cdot u_i)^2} \qquad (3)$$

Pomembno je poudariti, da izračunani podatki veljajo pri pravilni uporabi in popolni izločitvi vplivnih dejavnikov, denimo velikosti merjenca na sliki, vpliva prostorskega prehodnega pojava ali neizostrenosti slike, kota na normalo površine ...

5 Analiza termovizijskih kamer

Preskusili smo štiri termovizijske sisteme za presejalno merjenje ljudi s povišano telesno temperaturo. Vsem izdelkom je bila skupna navedena negotovost ±0,3 °C pri merjenju telesne temperature. Prav tako jih združuje način delovanja s samodejno zaznavo obrazov na vidni sliki s pomočjo umetne inteligence, pri čemer izstopa izdelek proizvajalca Dahua, ki omogoča avtomatsko zaznavo obrazov in proženje meritev tudi ob odsotnosti svetlobe na sliki vidnih spektrov.

Tabela 2: seznam ovrednotenih izdelkov
* vrednost je ocenjena na podlagi prikazane slike, uradni podatek ni znan

Proizvajalec (in model) toplotne kamere	iFOV (mRad)	IR ločljivost
YBFace		8x1 (ocena)
Hengtong	3,8	32x32
Nuctech	1,7	384x288
Dahua TPC-BF5421-T	1,3	400x300
FLUKE TiS45	3,9	160x120
FLIR T650sc	1,3	640x480

Cenovno najugodnejša izdelka proizvajalcev Hengtong optic-electric in YBFace sta osnovana na android sistemu z dodano toplotno kamero zelo nizke ločljivosti. Sistema sta namenjena vhodni identifikaciji zaposlenih s sočasnim brezkontaktnim merjenjem temperature. Ločljivosti termalnih kamer sta pri obeh izdelkih nizki, vidno polje pa zajema le del polja vidne slike. Edina dostopna meritev je meritev temperature obraza, ki je zaokrožena na desetinko stopinje. S statističnim ponavljanjem smo preverili raztros meritev pri merjenju kalibracijskega telesa.

Podjetji Nuctech in Dahua ponujata sisteme za video nadzor infrastrukture. Merilna sistema se dobavita z zunanjim referenčnim virom IR sevanja. Prednost tovrstnih izvedb je izboljšana stabilnost meritev, saj referenčna meritev zniža negotovost meritve temperature zaradi lezenja in nehomogenosti temperature ozadja toplotnega senzorja v merilni napravi. Sistema na osnovi prepoznanih obrazov na vidni sliki izmerita temperaturo pripadajoče regije na toplotni sliki. Zaradi nedostopnosti izmerjenih vrednosti posameznih slikovnih elementov smo lahko ovrednotili le zunanji vir sevanja.

Za primerjavo smo ovrednotili še dve splošno-namenski termovizijski kameri (slika 3).



Slika 3: splošno-namenski kameri proizvajalcev FLUKE (levo) in FLIR (desno)

Kameri nizkega (FLUKE) in visokega (FLIR) cenovnega ranga sta namenjena industrijski rabi, obe pa podpirata izvoz meritev za nadaljnjo obdelavo.

6 Primerjava termovizijskih kamer

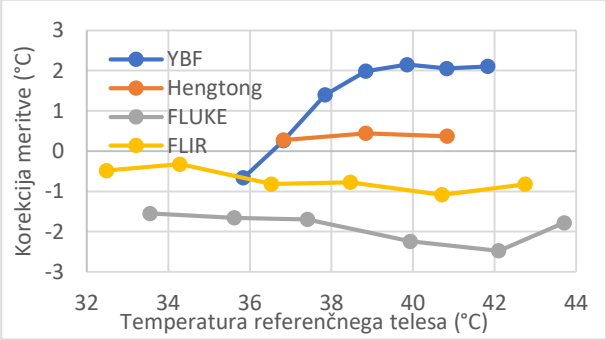
Laboratorijska oprema je temperaturno umerjena in sledljiva po ITS-90 temperaturni lestvici do mednarodnih etalonov. Pri ovrednotenju kamer je bilo kot vir referenčne temperature uporabljeno črno telo v vodni kopeli s izračunano emisivnostjo 0,998 in razširjeno negotovostjo 0,2 °C.

Cenovno najugodnejši napravi proizvajalcev YBFace in Hengtong sta pri ovrednotenju izkazali nenavadno funkcijo. Sistema pri meritvah pod neznano mejno vrednostjo prikažeta na videz naključno vrednost z drugačnim raztrosom in nespremenljivim povprečjem. Vpliv naključnih vrednosti je viden v grafu korekcij naprav (slika 4) pri sistemu proizvajalca YBFace, kjer korekcija nenadno pada sorazmerno referenčni temperaturi kalibracijske naprave. Pri izračunu negotovosti so bile meritve te kamere pri kalibracijskih točkah pod 40 °C zavržene kot neveljavne.

Tabela 3: korekcija kamer v ciljnem območju glede na referenčno temperaturo telesa in razširjena negotovost kamer pri merjenju površinske in telesne temperature ljudi.

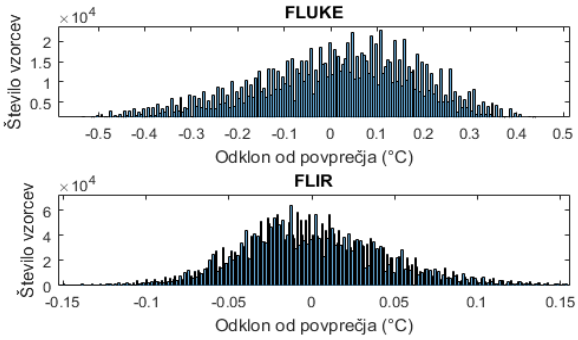
Temp. merjenca (°C)	Korekcija (°C)			
	YBF	Heng-tong	FLUKE	FLIR
33			-1,5	-0,4
34			-1,6	-0,3
35			-1,6	-0,5
36	-0,5		-1,7	-0,7
37	0,5	0,3	-1,7	-0,8
38	1,5	0,4	-1,8	-0,8
39	2,0	0,4	-2,0	-0,8
40	2,1	0,4	-2,2	-1,0
41	2,1	0,4	-2,4	-1,0
42	2,1		-2,5	-0,9
$U_{koža,k=2}$	**0,3**	**0,6**	**0,9**	**0,6**
$U_{k=2}$	**1,9**	**2,0**	**2,1**	**2,0**

Vse merjene kamere smo ovrednotili v laboratorijskih pogojih. Za izmerjene negotovosti se upošteva, da je kamera umerjena in brez sistematskega pogreška. Nobena od kamer ni bila primerno predhodno kalibrirana (slika 4), kar bi pri uporabi po navodilih proizvajalca dodatno povečalo napako meritve.



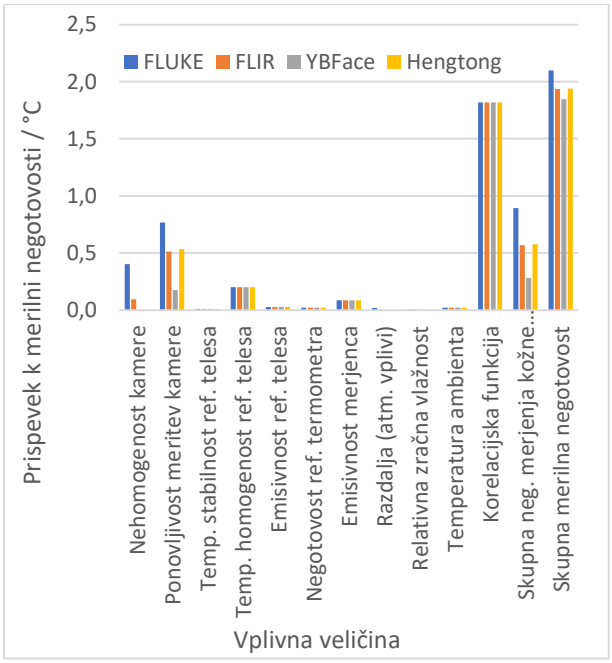
Slika 4: korekcija ovrednotenih naprav glede na referenčno temperaturo

Homogenost celotnega kadra kamere smo ovrednotili z večjim referenčnim črnim telesom v vodni kopeli (slika 5).



Slika 5: primerjava histogramov odstopanj od povprečja slike pri merjenju

Slika 6 prikazuje utežene posamezne doprinose negotovosti k skupni merilni negotovosti. Pri tem je pomembno izpostaviti, da ponovljivost meritev pri izdelku YBF in Hengtong sočasno odraža tudi nehomogenost kamer. Negotovost korelacijske funkcije je predvidena skupna negotovost pri merjenju in oceni notranje telesne temperature merjenih osebkov. Pri interpretaciji je pomembno upoštevati, da so v negotovosti korelacijske funkcije upoštevani tudi podatki o negotovosti uporabljene termovizijske kamere, ne pa tudi negotovost ušesnega termometra.



Slika 6: Posamezni doprinosi ($c_i \cdot u_i$) k merilni negotovosti sistema merjenja in ocene telesne temperature

7 Zaključek

Želena merilna točnost medicinske stroke pri diagnostiki povišane telesne temperature velja za ±0,3 °C. V procesu ovrednotenja smo ugotovili dejanske negotovosti inštrumentov, ki večinoma presegajo želeno negotovost že pri merjenju temperature kože, sploh pa jo presegajo ob upoštevanju negotovosti korelacijske funkcije za izračun telesne temperature iz najvišje temperature očesnega področja.

Literatura

[1] G. Dickson et al., "The variability of the facial artery in its branching pattern and termination point and its relevance in craniofacial surgery," European Journal of Plastic Surgery, vol. 37, str. 1-8, september 2013.

[2] E. Y. K. Ng, G. J. L. Kaw in W. M. Chang "Analysis of IR thermal imager for mass blind fever screening," Microvascular research, vol. 68, št. 2, str. 104-109, September 2004.

[3] R. Vardasca, A. R. Marques, J. Diz, A. Seixas, J. Mendes, E. F. J. Ring, " The influence of angles and distance on assessing inner-canthi of the eye skin temperature," Thermology international, vol. 27, št. 4, str. 130-135, 2017.

LOEE 4.0 – oddaljeno merjenje vektorskega polja z uporabo 3D kartezičnega stroja

Kristjan Vuk Bališ[1], Janez Zaletelj[2], Marjan Dolinšek[1], Iztok Humar[1]

[1]LOEE, [2]LUCAMI, FE, UL
E-pošta: kristjan.vuk@fe.uni-lj.si

LOEE 4.0 – Remote measurements of vector field employing an universal 3D cartesian machine

Abstract. *This paper presents a progress towards an establishment of remote laboratory exercises of the Fundamentals of electrical engineering course for 1st grade undergraduate students. The project has been started to overcome the issues of current situation, preventing students from being a part of supervised laboratory exercises at the faculty. The key components which are being developed for the remote measurements of the electric field are the Teslameter, the Cartesian positioning machine and the web portal. The system presented in the paper will enable students to remotely observe and control position of the measurement device, and to observe, save and process the values of magnetic field of the coil. The system will offer complete features of the corresponding hands-on lab exercise.*

1 Uvod

Pričujoče epidemiološke okoliščine nas učijo, kako zelo je pomembna tudi agilnost pedagoških pristopov v ustreznem spremstvu tehnoloških rešitev in orodij. Neposredno delo študentov na laboratorijski opremi, ki pomeni pridobivanje nepogrešljivega paketa veščin v smislu zlivanja praktičnega in teoretskega znanja, je bilo v minulem semestru v veliki meri hendikepirano. Glede na aktualne epidemiološke trende smo se v Laboratoriju za osnove elektrotehnike in elektromagnetiko – LOEE spustili v razvoj univerzalnega sistema za oddaljeno izvajanje laboratorijskih vaj pri poučevanih predmetih.

Glede na valovni značaj elektromagnetnega pojava moramo v elektromagnetiki v splošnem v obravnavo spustiti vse štiri dimenzije prostor-časa. Ta kriterij pade tudi na konstelacije laboratorijskih eksperimentov, kjer imamo v najbolj »neugodnem« primeru opravka z merjenjem časovno-spremenljivega prostorskega vektorskega polja, kakršno je npr. magnetno polje ter pripadajoča vektorja; vektor gostote magnetnega pretoka ter vektor magnetne poljske jakosti.

Stroj, ki bi zadoščal takšni štiri dimenzionalni nalogi, je tri dimenzijski kartezični stroj in pa temperatura nad absolutno ničlo, da steče četrta, časovna dimenzija; pa tudi polprevodniki morajo »oživeti«.

2 Zahteve

2.1 Vektorska 3D meritev

Ker so merjene veličine vektorske, moramo poleg obvladovanja štiridimenzionalnega prostor-časa zagotoviti tudi tridimenzionalno, tj. vektorsko meritev (npr. gostote magnetnega pretoka). V ta namen predstavljamo razviti Teslameter 3D in pripadajočo vektorsko Hallovo sondo, ki jo sestavljajo trije enaki enodimenzijski magnetorezistivni senzorji, nameščeni na FR4 origami, oblikovan v strukturo trirobnika.

2.2 Oddaljeni dostop

Izhodišče in pobuda za tovrstno nadgradnjo laboratorijskih eksperimentov je natanko možnost upravljaja na daljavo z ustrezno povratno informacijo (video strujanje merilne konstelacije in pa pridobivanje aktualnih merilnih rezultatov). Potrebna je torej implementacija telemetrije in telekomande, pri čemer naj slednje omogoča karseda dober približek izkušnji, kakršne je študent deležen pri opravljanju vaje v živo. Ciljna tehnologija za omogočenje oddaljenega dostopa je protokolni sklad TCP/IP ter obstoječe in po meri izdelane rešitve na aplikacijskem sloju. Upravljanje z napravo bo mogoče prek grafičnega uporabniškega vmesnika v smislu spletne aplikacije.

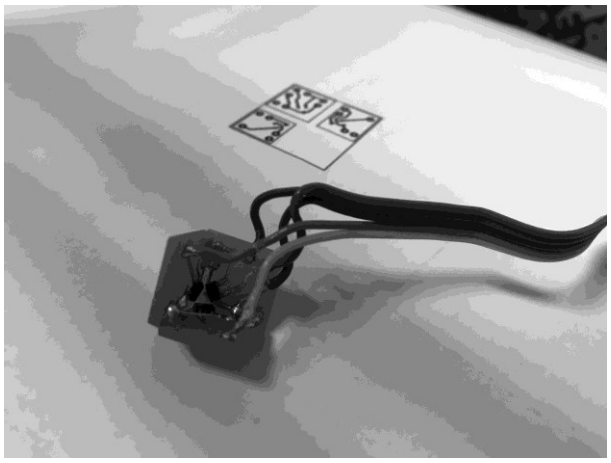
3 Teslameter 3D

V duhu nadaljevanja laboratorijske tradicije, ki je že rodila maloserijski teslameter, je v sklopu tega projekta inicializiran razvoj teslametra za vektorsko merjenje nizkofrekvenčnega magnetnega polja. Prvi izziv je načrtovanje in izgradnja vektorske Hallove sonde.

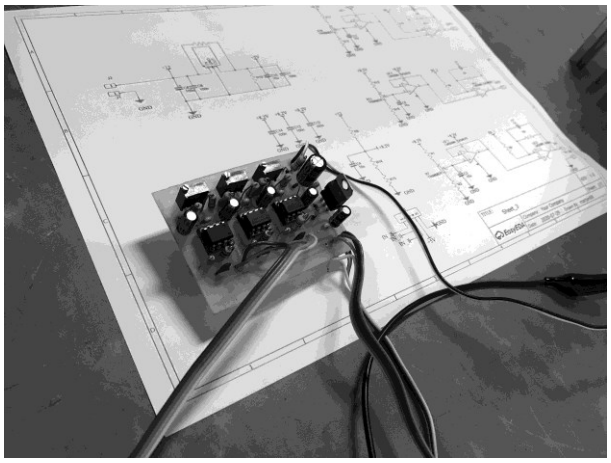
3.1 FR-4 origami

Za izgradnjo vektorske Hallove sonde sta bili na voljo dve možnosti. Uporaba integriranega 3-osnega Hallovega senzorja ali njegova sestava iz treh enoosnih. Z argumenti ekonomičnosti, demonstrativne moči in pa drugih tehničnih karakteristik enih in drugih senzorjev smo se odločili za slednjo možnost. Tri SMD čipe s SOT-23 ohišjem smo postavili v ortogonalno konstelacijo na načrtovani FR4 origami.

Prototip Hallove sonde je nastal v matičnem laboratoriju, kjer smo za potrebe razvoja tega projekta vzpostavili kapacitete za izdelavo prototipnih tiskanin s tehniko prenosa tonerja. Slika 1 prikazuje fotografijo drugega prototipa vektroske Hallove sonde.



Slika 1. Fotografija 2. prototipa vektorske Hallove sonde



Slika 2. Fotografija prototipnega vezja Teslametra 3D

3.2 Elektronsko vezje Teslametra 3D

Uporabljenim čipom je treba zagotoviti stabilno in šuma očiščeno enosmerno napajalno napetost v višini 5 V. Meritev komponente vektorja gostote magnetnega pretoka je dostopna na izhodu čipa kot analogna napetost, superponirana prednapetosti v višini polovice napajalne napetosti – v praksi se izkaže, da ta napetost odstopa do 10 % od pričakovane vrednosti, kot navaja tudi podatkovni list čipa, zato ne gre računati na nek generični postopek nastavljanja inštrumenta.

Ker mora teslameter »otipati« že zelo šibka magnetna polja, po gostoti primerljiva z zemeljskim, bo za dobro digitalno ločljivost (po analogno-digitalni pretvorbi z amplitudnim območjem 0-5 V) glede na specificirano občutljivost uporabljenega Hallovega senzorja v višini 14 mV/mT , potrebno ustrezno visoko nizkofrekvenčno ojačenje. V ta namen uporabimo operacijski ojačevalnik v invertirajoči vezavi brez razklopnega kondenzatorja na vhodu in izhodu. V povratno vezavo poleg upora dodamo vzporedni kondenzator, ki omeji pasovno širino ojačevanika in oslabi šum. S tem v resnici delamo kompromis, kajti četudi je zahteva obvladovanje magnetostatičnega polja, nas pri zmanjševanju pasovne širine omejuje najmanjša časovna konstanta RC, ki določa frekvenco zajema meritev pri krajevno spreminjajočem se polju ob premikanju sonde.

Za analogno-digitalno pretvorbo ojačenega signala na izhodu operacijskega ojačevalnika vprežemo ceneni Microchip-ov mikrokrmilnik serije attiny412. Ta ima integriran ADC, ki lahko po časovnem multipleksu vzorči signale na vseh šestih GPIO pinih čipa. Mi bomo za pretvorbo sicer potrebovali tri pine, dva pa bosta služila za serijsko komunikacijo z mikroračunalnikom Raspberry Pi (RPi), preko katere bomo pošiljali digitalizirano 10-bitno meritev.

Slika 2 prikazuje fotografijo izdelanega prototipnega vezja Teslametra 3D.

3.3 Nastavljanje teslametra

Nastavljanje se začne s čim bolj točno nastavitvijo vseh treh referenčnih prednapetosti na neinvertirajočih vhodih operacijskih ojačevalnikov, ki ojačujejo signal s senzorjev. Medtem ko je sonda izpostavljena ničelnemu magnetnemu polju, nastavimo referenčne prednapetosti tako, da se karseda približamo polovični napetosti ciljnega intervala. Tega lahko pogojuje npr. dovoljeno amplitudno območje na vhodu analogno-digitalnega pretvornika, s katerim bomo meritev digitalizirali.

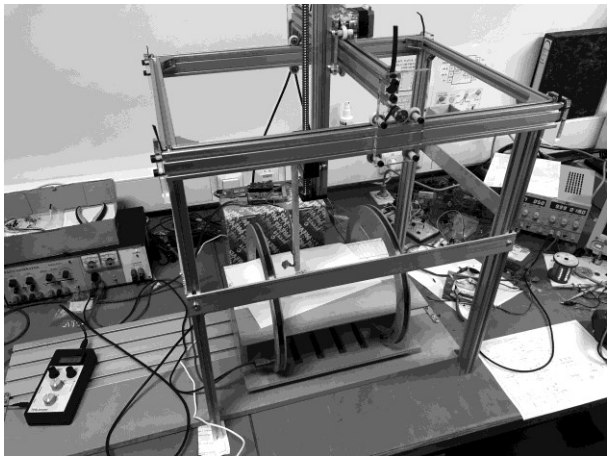
Naslednji korak je odčitek že digitalizirane meritve za posamezen senzor in zapis le-te v pomnilnik kot digitalne vrednosti pri ničelnem polju. Nato posamezen senzor na sondi izpostavimo polju znane gostote, pripadajočo digitalno vrednost pa zopet zapišemo v pomnilnik. Ob, sodeč po podatkovnem listu uporabljenih čipov, upravičeni predpostavki, da je odnos med analognim izhodnim napetostnim signalom in detektirano gostoto magnetnega pretoka linearen, lahko sedaj izračunamo konstanto pretvorbe v obliki tesla na najmanj uteženi bit digitalne pretvorbe [T/LSB].

Kolikšna bosta merilni doseg in ločljivost pri dani bitni globini analogno-digitalne pretvorbe, je stvar načrtovanega/nastavljenega ojačenja operacijskih ojačevalnikov.

4 3D kartezični stroj

Prototipni kartezični stroj je bil izdelan po meri s kriterijem pokrivanja gabaritov od dimenzijsko najmanjših do največjih eksperimentov, s katerimi razpolaga laboratorij. Premike po koordinatnih oseh omogočajo trije koračni motorji, ki jih je za tiho in gladko kinematiko stroja smiselno poganjati v režimu mirkokorakanja. V ta namen se uporabljajo ustrezna gonilna vezja (za vsak koračni motor po eden), ki so nasajena na RPi ščit proizvajalca PROTONEER PiCNC, namenjen krmiljenju CNC strojev. Ščit z

nadrejenim mikroračunalnikom RPi komunicira s pomočjo asinhrone serijske komunikacije UART.



Slika 3. Fotografija prototipnega kartezičnega stroja

Srce PiCNC ščita je Atmelov 8-bitni mikrokrmilnik. Ta služi kot interpreter G-kode, ki je standarna oblika podajanja ukazov najrazličnejšim računalniško vodenim strojem. Tovrstne ukaze mikrokrmilniku posredujemo s pomočjo UART serijskega vmesnika.

Vsi trije uporabljeni koračni motorji imajo specificiran kot koraka na 1,8°. V režimu mikrokorakanja 1/8 na ta način dosežemo 1600 korakov na en polni obrat motorja, kar nam glede na uporabljena vodila postavi teoretsko mejo resolucije premikov na 25 mikrometrov po oseh x in y ter 2,5 mikrometrov po z osi. Na ta način lahko z zastavljeno robotizacijo laboratorijskih vaj dosežemo veliko večjo krajevno natačnost pozicioniranja sonde, kar tudi izkušnjo nekoliko približa današnjim razmeram v industriji.

5 Oddaljen ogled in upravljanje

Kot sledi iz uvodnih zahtev, želimo upravljanje z napravo omogočiti na daljavo preko interneta z uporabo spletnega vmesnika. Pri zasnovi in izvedbi vmesnika je potrebno upoštevati zahteve glede avtentikacije uporabnikov in spletne varnosti.

5.1 Sistemska arhitektura

Slika 5 prikazuje sistemsko arhitekturo oddaljenega laboratorija.

Nadzor nad kartezičnim strojem, to je premikanje sonde in pa zajem meritev teslametra, zaupamo mikroračunalniku RPi, ki je povezan v lokalno fakultetno omrežje. Na RPi teče Node-RED strežnik, s katerim lahko vizualno v obliki pretokov programiramo priključene naprave in na enostaven način dostopamo do nizkonivojskih in visokonivojskih vmesnikov ter komunikacijskih protokolov. Raspberry tako na eni strani komunicira s ščitom in pa dodatno vpletenim mikrokrmilnikom, na drugi strani pa komunicira s spletnim strežnikom. V obliki MQTT sporočil dobiva ukaze za pozicioniranje ter izvedbo meritev, sam pa pošilja MQTT sporočila z rezultati meritev ter statusom merilne naprave.

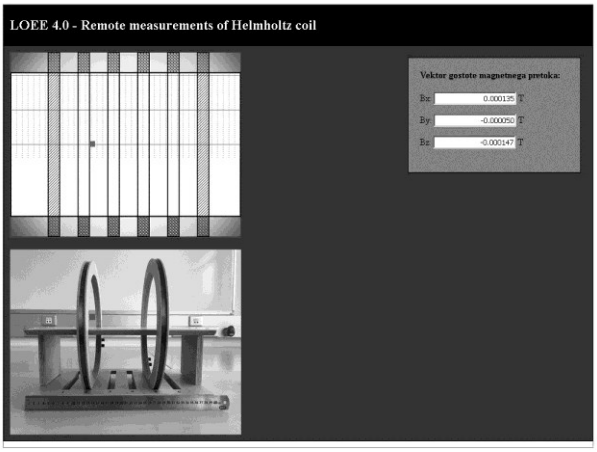
Na RPi je priključena še kamera, ki je lahko bodisi RPi kamera modul, priključena neposredno na namensko vodilo, bodisi običajna spletna kamera. Program »Motion« skrbi za zajem slike ter zagotavlja spletni strežnik, prek katerega dostopamo do slike kamere, ki se samodejno osvežuje. Ker RPi nima javnega IP-ja, se prikaz slike vrši preko proxy strežnika, ki preusmeri HTTP zahtevo na RPi strežnik.

Javni spletni strežnik skrbi za oddaljeni dostop do enega ali več merilnih mest, nadzor dostopa do posameznih mest glede na njihovo razpoložljivost, overjanje uporabnikov s pravicami do dostopa ter prikaz statusa in rezultatov meritev. Večina funkcionalnosti je izvedena v okviru Node-Red pretokov, komunikacija z merilnimi mesti pa poteka po MQTT aplikacijskem protokolu. Node-Red ponuja svoj programirljiv uporabniški vmesnik Dashboard s privzetim naborom modulov za interakcijo s klientom in prikaz podatkov, v katerega pa lahko vgradimo tudi svojo programsko kodo.

5.2 Uporabniški scenarij in spletni vmesnik

Uporabnik si dostop do aplikacije zagotovi z avtentikacijo. Na nivoju prototipa se za ta namen uporablja Google Sign-In API, s pomočjo katerega se avtentikacija najprej izvrši na strani klienta, nakar se pridobljeni žeton (JWT) pošlje aplikacijskemu strežniku, ki preveri identiteto in pooblastila avtenticiranega uporabnika.

Uporabnik je nato preusmerjen na uporabniški vmesnik Node-red Dashboard, kjer lahko spremlja status posameznih naprav oziroma merilnih mest. Ob vsaki uporabnikovi akciji se preverja istovetnost uporabnika preko tehnologije piškotkov.



Slika 4. Posnetek spletnega grafičnega vmesnika za upravljanje z napravo in prikaz meritev

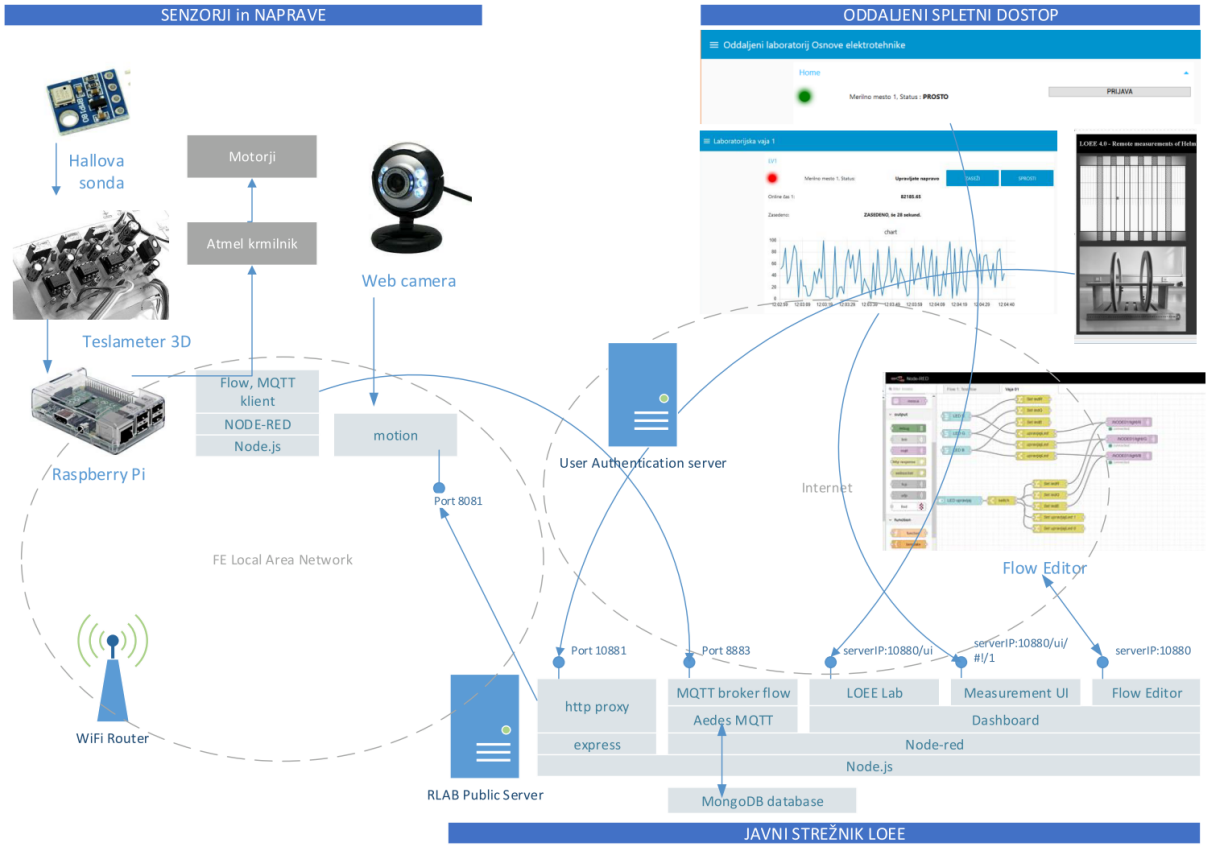
V kolikor je merilno mesto na voljo ter prosto, uporabnik pa je avtoriziran za njegovo uporabo, lahko ta le-tega zaseže za vnaprej določen časovni interval. Na ta način si uporabnik pridobi pravico do nadzora merilne

naprave, medtem ko imajo dostop do telemetrijskih podatkov vsi aktivni uporabniki.

5.3 Uporabniški vmesnik za oddaljeno merjenje

Prilagojeni uporabniški vmesnik sestavlja prikaz interaktivne skice merilne konstelacije v dveh projekcijah, na katerih se s povlekom miške pozicionira sondo v 3D območju. Kot povratna informacija je v uporabniški vmesnik vključen tudi živ video kader kartezičnega stroja z vsebovano merilno konstelacijo, ki daje uporabnik-u/om vpogled v stanje naprave. Na osnovnem vmesniku je prikazan set telemetrijskih podatkov, to je trokomponentna meritev gostote magnetnega pretoka, ki jo zagotavlja razvita vektorska merilna sonda.

Slika 4 prikazuje posnetek osnovnega spletnega grafičnega vmesnika za upravljanje z napravo.



Slika 5. Sistemska arhitektura oddaljenega laboratorija

6 Zaključek

Zastavljeni projekt je izrazito interdisciplinarne narave, njega razvoj pa terja širok nabor elektrotehniških veščin s področij elektromagnetike, elektronike, telekomunikacij, računalništva in nenazadnje programiranja na več nivojih. Zaradi kompleksnosti izvedbe pričakujemo, da bo prva produkcijska različica sistema na voljo šele nekje ob izteku koledarskega leta. Večina predstavljenih prototipnih rešitev je v zgodnji fazi razvoja in nam služijo kot »proof of concept« pri razvoju končnega sistema.

Literatura

[1] I. Humar, D. Tacar, P. Kokelj: Razvoj, izdelava in uporaba ročnega merilnika gostote magnetnega pretoka - teslametra. Zbornik dvajsete mednarodne Elektrotehniške in računalniške konference ERK 2011, 19.-21. september 2011, Portorož, Slovenija. Zv. A, str. 79-82.

[2] P. Kokelj, D. Tacar, I. Humar: Razvoj, izdelava in uporaba merilnika gostote magnetnega pretoka - teslametra. Zbornik devetnajste mednarodne Elektrotehniške in računalniške konference ERK 2010, 20.-22. september 2010. Portorož, Slovenija. Zv. B, str. 369-372.

[3] P. Kokelj, D. Tacar,. Laboratorijske vaje iz OE - magnetno polje. Zbornik osme Elektrotehniške in računalniške konference ERK 1999, 23. - 25. september 1999, Portorož, Slovenija. Zv. B, str. 405-408.

[4] D. Križaj, M. Polenšek, M.Jankovec: Helmholtzov par kot okolje za spoznavanje osnovnih lastnosti magnetnega polja ob uporabi računalniško podprtega merilnega okolja. Zbornik enaindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2012, 17.-19. september 2012, Portorož, Slovenija. Zv. B, str. 269-272.

Akustika in elektroakustika

Acoustics and Electroacoustics

Zahteve za odmevnice in pristop k projektiranju odmevnice akustičnega laboratorija InnoRenew CoE

dr. Rok Prislan

InnoRenew CoE, Livade 6, SI-6310 Izola, Slovenija
E-pošta: rok.prislan@innorenew.eu

Requirements for reverberation chambers and the design approach to the InnoRenew acoustic laboratory chamber

Several important acoustic quantities and parameters are measured in the acoustic environment of a reverberation chamber. In the chamber the presence of the diffuse sound field is assumed, which is difficult to justify due to the absence of well accepted diffuse sound field quantifiers. As a consequence, it is also difficult to define exact design goals for reverberation chambers. In the article some requirements for reverberation chambers are summarized as defined by the standardized measurement procedures, such as for determining the diffuse sound filed absorption coefficient of materials (α), sound power of noise sources (L_W) and random-incidence scattering coefficients of surfaces (s). Furthermore, several aspects of the initial acoustic design of the reverberation chamber emerging as part of the acoustic laboratory of InnoRenew CoE are presented.

1 Uvod

V InnoRenew CoE se ukvarjamo z raziskovanjem obnovljivih materialov in trajnostne gradnje ter vpeljavo znanstvenih dognanj v industrijsko prakso. Pri tem se posebej posvečamo inovacijskemu in interdisciplinarnemu proučevanju lesa in njegove uporabe. Pomembno razvojno-raziskovalno področje v InnoRenew CoE je tudi akustika, zato je vzpostavljen akustični laboratorij, ki poleg strokovnega kadra in laboratorijske opreme zajema tudi laboratorijske prostore. V gradnji sta gluha soba in odmevnica, pri čemer so izzivi ob načrtovanju slednje predstavljeni v prispevku.

Odmevnica ni tipski prostor, za izgradnjo katerega bi bila na voljo konkretna navodila ali načrti. Nekatere zahteve za odmevnice so sicer dane v sklopu opisa standardiziranih merilnih postopkov, a se jih le peščica nanaša na fizične lastnosti prostora. Večina zahtev je namreč danih za zvočno polje, ki se vzpostavi v prostoru, zaradi česar so obvezen del akustičnega projektiranja odmevnice tudi napredne računalniške simulacije akustike prostora.

2 Predpostavka difuznega zvočnega polja

V odmevnicah predpostavimo prisotnost difuznega zvočnega polja, ki je v akustičnem slovarju [1] definiran kot: " *idealizirano zvočno polje sestavljeno iz neskončno mnogo nekoreliranih ravnih valov, ki imajo enakomerno smerno porazdeljeno intenziteto. Skupna intenziteta je torej nič.* "

Takšno polje je homogeno in izotropno, kar je priročna predpostavka v teoretičnih izpeljavah. Ker pa difuznega polja v praksi ni mogoče vzpostaviti, kljub temu, da ga v odmevnicah pričakujemo, gre torej za precej kontroverzen koncept. Pri tem stroka ni poenotena [2, 5] glede definicije difuznega zvočnega polja, njegovega ovrednotenja (karakterizacije) in razumevanja povezave med odbojem zvoka na mejnih površinah in vzpostavljenim zvočnim poljem v prostoru. Kljub prepoznani praktični uporabnosti koncepta difuznega zvočnega polja obstaja torej večja vrzel v njegovem razumevanju, kar povzroča negotovosti tudi pri projektiranju odmevnic [6].

3 Zahteve za odmevnice

Nekatere zahteve za odmevnice in zvočno polje, ki se v njih vzpostavi, so dane v sklopu standardiziranih merilnih postopkov, ki se izvajajo v odmevnicah. Med najbolj relevantnimi področnimi standardi so:

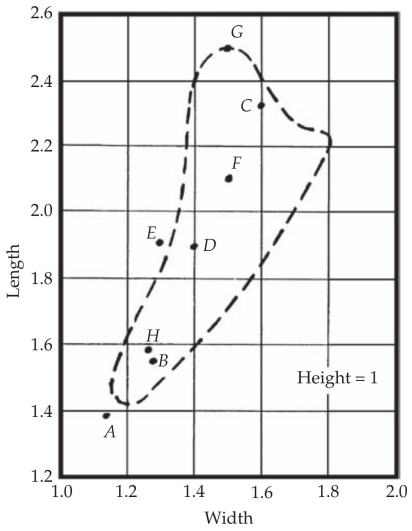
- SIST EN ISO 354:2004 [7], ki obravnava merjenje koeficientov zvočne absorpcije (α). Meritev temelji na vplivu, ki ga ima vnos vzorca materiala tipične velikosti 10-12 m^2 na odmevni čas v prostoru.
- ISO 17497-1:2004 [8], ki obravnava merjenje koeficientov sipanja na površinah, in sicer ob naključnem vpadu zvočnega vala.
- SIST EN ISO 3741:2010 [9], ki obravnava meritev zvočne moči in ravni zvočne energije virov hrupa. Za nekatere specifične vire hrupa poznamo tudi dodatne merilne standarde, kot npr. ISO5135:1999 [10] za dušilne in zaporne elemente prezračevalnih sistemov.

Praviloma uporabljamo isto odmevnico za več vrst meritev, kar pomeni, da ločevanje med zahtevami posameznega standarda ni smiselno, saj mora odmevnica zadoščati zahtevam vseh.

3.1 Geometrijske zahteve

Nekatere zahteve so vezane na geometrijo odmevnice:

- Volumen prostora mora presegati $V = 200\,\mathrm{m}^3$, a se za meritve tudi v frekvenčnih pasovih pod 100 Hz priporoča bistveno večje volumne od $200\,\mathrm{m}^3$. Ker pa večji volumni povečujejo zvočno atenuacijo, ki povzroča pretirano dušenje v območju visokih frekvenc, se le redko gradijo odmevnice volumnov večjih od $300\,\mathrm{m}^3$.
- Največja razdalja med poljubnima točkama v prostoru mora biti manjša od $1.9 \cdot V^{1/3}$, kar pomeni, da zasnova odmevnice kot izrazito podolgovatega ali ploščatega prostora ni dopustna.
- Če je prostor oblike kvadra, dimenzije stranic ne smejo biti v celoštevilskem razmerju, saj je s tem distribucija prostorskih resonanc neugodna. Zato so dana konkretna priporočila za razmerja stranic, lahko pa uporabimo tudi Boltov pristop [11] (glej sliko 1).
- Za doseganje zadostno homogenega zvočnega polja so dana priporočila glede velikosti in oblike visečih difuzorjev, ki jih v prostor umeščamo po izgradnji.



Slika 1: Optimalno razmerja stranic prostora oblike kvadra (Boltovo območje [11])) označeno kot območje omejeno s prekinjeno črto (vir: [12]).

3.2 Akustične zahteve

Preostale zahteve za odmevnice so dane za njene akustične lastnosti in zvočno polje, ki se v prosotu vzpostavi. Med temi gre našteti:

- Pod dano frekvenčno mejo f_0, mora biti povprečen koeficient absorpcije α nižji od 0.16, nad njo pa nižji od 0.06. Frekvenčno mejo se izračuna na osnovi volumna prostora

$$f_0 = 2000 \cdot V^{-1/3} \quad [\mathrm{Hz}]. \tag{1}$$

- Največja dopustna absorpcija v prostoru je omejena z volumnom prostora

$$A \leq 0.3 \cdot V^{2/3}. \tag{2}$$

- Ob delovanju zvočnega vira je predpisana dopustna deviacija zvočnega tlaka med merilnimi mesti.
- Zagotoviti moramo nizko raven hrupa ozadja v času meritve, kar dosežemo z visoko zvočno izolativnostjo ločilne konstrukcije. Sestava te je praviloma masivna, in sicer armirano betonska. Poleg tega odmevnica nima direktnega togega stika z okolico, temveč stoji na ustrezno dimenzioniranem izolacijskem sloju.
- Za čas meritve morata biti temperatura in vlaga stabilni in torej znotraj definiranega dopustnega območja. Ker gre najpogosteje tudi za toplotno izredno izolirane prostore, se težave z ohranjanjem temperature in vlage praviloma pojavijo zgolj ob meritvah zvočnih virov, ki oddajajo veliko toplote ali vplivajo na vlago v prostoru.

4 Pristop k projektiranju

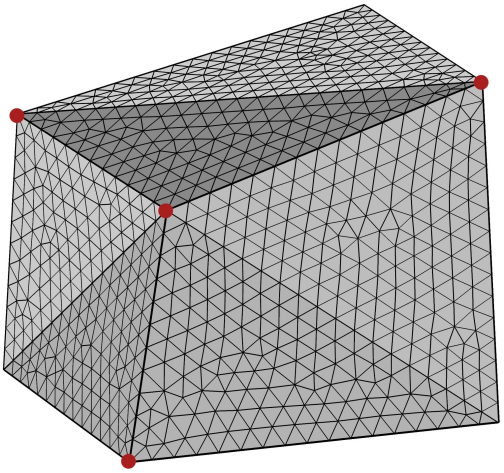
Raziskovalni inštitut InnoRenew CoE gradi nov kompleks inštituta, del katerega je tudi akustični laboratorij. Osnovna omejitev za odmevnico je tako prostor, ki je v objektu na razpolago, in dopustna obremenitev nosilne konstrukcije. S tem je bila že v grobem določena geometrija odmevnice, dodatno pa smo pregledali še priporočila aktualne strokovne literature [13, 14, 6] in primere dobre prakse [15, 16]. Na osnovi tega je bila sprejeta odločitev, da obodne površine odmevnice ne bodo vzporedne, saj lahko to ob vnosu absorpcijkega materiala bistveno vpliva na homogenost zvočnega polja.

Obodne površine bodo grajene iz armiranega betona, opaženje katerih pa prinaša dodatne omejitve. Izredno kompleksne geometrije prostora bi namreč povzročile veliko podražitev izvedbe, zato bo odmevnica grajena iz devetih ravni stranic. Od teh bodo tla in dve steni štirikotne ter med seboj pravokotne, preostanek pa bo iz trikotnih delov, kot je prikazano na sliki 2.

Natančna lega vogalov prostora, ki so na sliki 2 označeni z rdečo, ni določena, temveč je podvržena akustični optimizaciji geometrije. Tako so v času pisanja prispevka v teku računalniške simulacije, na podlagi katerih bo v nadaljevanju definirana končna geometrija odmevnice.

Optimizacija geometrije odmevnice zajema geometrijske metode modeliranja akustike prostora [17, 18], ki jih uporabimo za prepoznavo periodičnih trajektorij. Te so neželene, saj lahko povzročijo persistentno zvočno polje v ravnini, ki znižuje homogenost in izotropnost, ki sta temeljni zahtevi difuznega zvočnega polja.

Optimizacija geometrije prostora zajema tudi uporabo valovnih metod modeliranja akustike prostora [4], in sicer uporabo metode končnih elementov (računalniško orodje COMSOL Multiphysics). Tako se prostor diskretizira na končne volumske elemente, kot je vidno v obliki mreže



Slika 2: 3D model odmevnice: tla in dve steni (v prikazu niso vidne) tvorijo štirikotne stranice, ki so med-seboj pravokotne, preostale stranice pa so trikotne. Lege rdeče označenih vogalov prostora so predmet akustične optimizacije geometrije.

na sliki 2. Simulacije z metodo končnih elementov zajemajo nizko frekvenčno območje, v katerem so relevantni posamezni modalni načini prostora. Pri tem nas zanima distribucija prostorskih resonanc, dodatno pa izvedemo virtualni eksperiment merjenja koeficientov absorpcije skladno z ISO 354 [7].

V sklopu tega v prostor umestimo točkovne zvočne vire in opazovane točke, ki predstavljajo umestitev zvočnika in mikrofonov. Upoštevamo tudi ustrezne robne pogoje, in sicer robni pogoj realne impedance za toge površine in poroznega zvočno vpojnega sloja za vzorčni material. Tako pridobimo frekvenčni odziv prostora in odmevni čas, ki služita kot optimizacijska kriterija.

5 Zaključki

Predstavljeni so bili današnji izzivi akustičnega projektiranja odmevnic, ki izhajajo iz omejenega razumevanja difuznega zvočnega polja in neoprijemljivih zahtev merilnih standardov. Povzete so bile zahteve, ki jih za odmevnice podajajo standardizirani postopki merjenja akustičnih količin in parametrov v odmevnici.

Na primeru odmevnice, ki se projektira v sklopu akustičnega laboratorija InnoRenew CoE, so bile prestavljene računalniške simulacije akustike prostora, na podlagi katerih izvajamo optimizacijo geometrije prostora. Z metodo končnih elementov kot kriterij optimizacije vzamemo distribucijo prostorskih resonanc in vpliv vnosa absorpcijskega materiala na frekvenčni odziv in odmevni čas. Z geometrijskimi metodami preverjamo pojavnost neželenih periodičnih trajektorij. V času pisanja prispevka so v teku računalniške simulacije, na podlagi katerih bo v nadaljevanju definirana končna geometrija odmevnice.

Zahvala

To delo je bilo podrto s sredstvi Okvirnega programa Evropske Unije, Obzorje 2020 (H2020 WIDESPREAD-2-Teaming: #739574, InnoRenew CoE) in Republike Slovenije, Investicijsko financiranje Republike Slovenije in Evropske unije iz Evropskega sklada za regionalni razvoj.

Literatura

[1] C. Morfey: *The Dictionary of Acoustics*, Academic Press, London (2001)

[2] T. J. Schultz: *Diffusion in reverberation rooms*, Journal of Sound and Vibration, 16(1), 17 - 28, 1971

[3] M. Schröder: *Die StatistischenParameter der Frequenzkurven von grossen Räumen*, Acta Acust. United Ac., 4, 594–600, 1954

[4] R. Prislan: *Modalni in statistični pristop pri opisu zvočnega polja v prostoru*, Elektrotehniški vestnik, 80(5), 184–188, 2013

[5] C.-H. Jeong: *Diffuse sound field: challenges and misconceptions*, Proceedings of 45th International Congress and Exposition on Noise Control Engineering, 1015-1021, 2016

[6] M. Vercammen: *On the Revision of ISO 354, Measurement of the Sound Absorption in the Reverberation Room*, 23rd International Congress on Acoustics, 2019

[7] *ISO 354:2003: Acoustics – Measurement of sound absorption in a reverberation room*, Geneva, Switzerland

[8] *ISO 17497-1:2004: Acoustics – Sound-scattering properties of surfaces – Part 1:Measurement of the random-incidence scattering coefficient in a reverberation room*, Geneva, Switzerland

[9] *ISO 3741:1999: Acoustics – Determination of sound power levels and sound energy levels of noise sources using sound pressure - Precision methods for reverberation test rooms,* Geneva, Switzerland

[10] *ISO 5135:1999: Acoustics – Determination of sound power levels of noise from air-terminal devices, air-terminal units, dampers and valves by measurement in a reverberation room,* Geneva, Switzerland

[11] R. H. Bolt: *Note on Normal Frequency Statistics for Rectangular Rooms,* The Journal of the Acoustical Society of America, 18(1), 130 - 133, 1946

[12] F. A. Everest, K. C. Pohlmann: *The Master Handbook of Acoustics, 5th*, The McGraw-Hill Companies, Inc (2009)

[13] M. Nolan, M. Vercammen, C.-H. Jeong in J. Brunskog: *The use of a reference absorber for absorption measurements in a reverberation chamber*, Proceedings of Forum Acusticum, 2014

[14] M. Nolan, M. Vercammen, C.-H. Jeong: *Effects of different diffuser types on the diffusivity in reverberation chambers*, Proceedings of EuroNoise, 2015

[15] P. D'Antonio, M. Nolan, E. Fernandez-Grande in C.-H. Jeong: *Design of a New Testing Chamber to Determine the Absorption, Diffusion and Scattering Coefficients*, Proceedings of the 23rd International Congress on Acoustics, 2259 - 2266, 2019

[16] P. D'Antonio, M. Nolan, E. Fernandez-Grande in C.-H. Jeong: *Design of a New Testing Chamber to Measure the Absorption Coefficient Down to 25 Hz*, 4027 - 4034, 2019

[17] R. Prislan, G. Veble in D. Svenšek: *Ray-trace modeling of acoustic Green's function based on the semiclassical (eikonal) approximation*, The Journal of the Acoustical Society of America, 140, 2695 - 2702, 2016

[18] R. Prislan: *Osnove geometrijskega modeliranja akustike prostora*, Zbornik šestindvajsete mednarodne elektrotehniške in računalniške konference, 285–288, 2014

Acoustical project of the new Banca Popolare di Sondrio in Berbenno Service Center - I

Marcello Brugola [1], Franz Policardi-Antoncich [2, 3],

[1] *Studio di Acustica Brugola*
[2] *LUCAMI, Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1000 Ljubljana, Slovenija*
[3] *Raziskovalni Inštitut za nove tehnologije in energetiko (R.I.N.T.E.) d.o.o., Stritarjeva 6/a, 4000 Kranj, Slovenia*

marcello.brugola@brugola.eu

Abstract

North Italian Banca Popolare di Sondrio (NIBS) recently acquired an old industrial warehouse with the intention of obtaining its new Service Center; a multimedia space to host specific meetings, the board seat and a series of learning rooms for the various employees and office workers refreshing courses had to be carefully projected and realized in accordance with architect's requirements. Our engineering studio was entrusted with the architectural and PA acoustics projects of the shoe-box shaped room and the complete Audio-Video system. Iterative procesess through specific softwares leaded to the final design and realization. In this paper we will present the basic project acoustic requirements and the obtained results for RT_{60}, SPL, RASTI, D_{50} and C_{80} values. This paper does not focus on possible external disturbing noises and on the Audio-Video system.

1 Introduction

Post construction acoustic refurbishing is a known and complex subject. Many different constraints have to be taken into account, as for example external noise impact on internal acoustics, possible building modifications, architect's requirements and last but not least cost. Designing and building a structure from scratch is often simpler and mentally stimulating but the job in acoustic buros requires a lot of flexibility. Many constructions not originally thought and designed so that acoustics play a primary role, are subsequently adapted for specific uses where acoustics, audio and video become predominant. Even some famous original and very expensive realizations examples led to bad results, often due to the designer's excessive confidence. Let us solely cite here two of the most important: both the New York Philharmonic Hall realized in 1960 by Bolt Beranek and Newmann and the Sydney Opera House realized in 1973 by Jørn Utzon with more than $100.000.000 both proved to be acoustically very problematic if not disastrous. In the first case more than $2.000.000 were not enough to improve acoustics and in 1976 it was even necessary to completely demolish the interior of the room and redesign it from scratch [1], while for the second it has been necessary to resort to subsequent very substantial and expensive modifications for more tha $100.000.000, which will end this year [2]. So good acoustics is a challenging goal.

2 The room

The building NIBS acquired some years ago was an industrial warehouse. A complete refurbishing process had to be projected and managed regarding acoustics and audio-video multimedia installation spaces, to be able to host specific meetings, the board seat and a series of learning rooms for refreshing courses. The main shoe-box shaped hall is 31 meters long, 16m wide and from 7,5 up to 9 m high with a slightly wood pitched ceiling; the floor is characterized by 5 horizontal planes: the lowest for the speakers, a first slightly rising (3,5%) area for the audience, a horizontal plane for the main aisle, a second slightly rising (3,5%) area for the audience and a rear

last horizontal plane taking to four exit doors, as showed in figure 1.



Figure 1. Side view of the main hall

The space is characterized by 540 seats. Audience is divided in 3 x 10 rows upwards and intersecated by a horizontal 3 m large aisle, which takes to the lateral double doors exits, so that 6 seat areas can be easisy defined, grouped by 3 in the direction of width. Every row has 9 seats and every area 90. The rear space is 2,5 m and the front 3,7 m up to the speaker's position.
Lateral distance from seats to walls is around 2 m and the inter-area distance is 2 to 3 seats, ie. 1,2 to 1,8 m. Seat dimension is standard 0,60 m and inter-row distance is 0,40 m; rows are slightly offset to allow better view, all showed in figure 2.

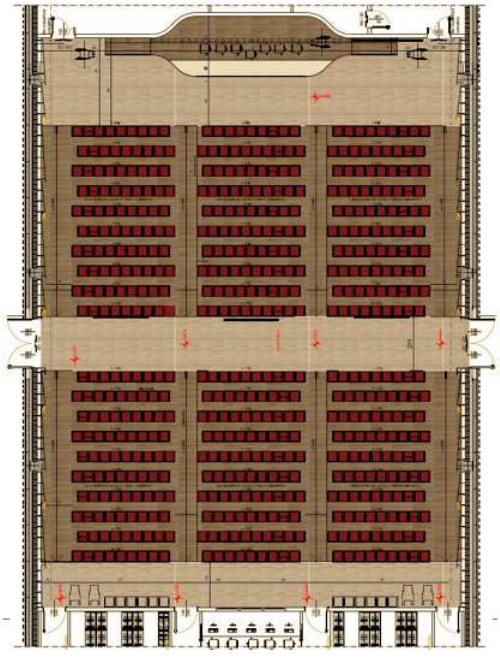


Figure 2. Floor plan of the main hall

3 Acoustical parameters

A mathematical software (Catt Acoustics) [3] was used to develop the most suitable solutions for the enclosed environment. This methodology has also been extended to the two main learning rooms, located on the building ground floor, while for the small rooms we relied on our simplified calculation software. The parameters we used to project and characterize the hall are the most commonly used in architectural acoustics, as shown in table 1, to which we added the Bass Ratio (LF).

Table 1. Acoustical parameters used in this project, with explanation and measure units

T10	Reverberation Time -10 dB (= EDT, Early Decay Time)	s
T15	Reverberation Time - 15 dB	s
T30	Reverberation Time - 30 dB	s
T60	Reverberation Time - 60 dB	s
SPL	Sound Pressure Level	dB
D 50	Deutlichkeit Faktor	%
C 80	Clarity Factor	dB
LF	Lateral Energy Fraction	%
STI	Speech Transmission Index	%

As regards acoustics in confined spaces, there is no specific legislation setting particular maximum or minimum limits for the main acoustic parameters. Each of them, despite having a scientific theoretical definition, is linked to subjective preferences, needs and objective evaluation and it becomes thus difficult to establish an univocal value. Our experience says that it is usually preferable to set an acceptability band for minimum and maximum values and the obtained results iteration process have to be evaluated on the basis of comparison ranges obtained from scientific literature and in the light of the experience and sensitivity of the acoustician. We based our choices on a series of values proposed by some main authors for a conference room with a 5,100 m^3 volume: auditoria Reverbetration Times up to -60 dB (RT_{60}): 0,9 s [4] 1,1 s [5] and 1,3 s [6].
The mentioned values are worldwide considered optimal for the intended use, from which it can be deduced that the RT_{60} in the central frequency bands 500-1000 Hz should ideally be around 1.0 to 1.2 s. Shorter RT_{60} mainly induce some improvements in speech

intelligibility; longer RT_{60} are mainly considered good for musical performances.
A further need was a well controlled Bass Ratio or Warmth, i.e. the absorption curve (practically the inverse of RT_{60}) had to be almost constant over the entire audio band, except for the lower part of the spectrum, with a slight low frequency increase up to a maximum of 130% from the reference 1.000 Hz central band [7]. Too high deviations from these tolerances indicate too high absorption (or reflection) phenomena at certain octaves, creating reflected signal distortions compared to the original one.
The D_{50} values were evaluated in % according to the following criterion:

Table 2. D_{50} evaluative criterion in %

Optimal	D50 > 60
Acceptable	30 < D50 < 60
Not acceptable	D50 < 30

The reference values for C_{80} are the following: speech $\geq$ 3 dB and music $-4 \leq C_{80} \leq 2$ dB.
The LF obtained values were evaluated according to the following criterion:

Table 3. LF evaluative criterion in %

Good	15 < LF < 35
Acceptable	10 < LF < 15
Not acceptable	LF < 10; LF > 35

And last but not least, the STI values have been evaluated according to the following table:

Table 4. STI evaluative criterion in %

Insufficient	< 30%
Poor	da 30% a 45%
Normal	da 45% a 60%
Optimal	da 60% a 75%
Excellent	> 75%

4 The 4 phases

Phase I: model preparation. The first phase focused on the environment geometrical analysis and on the 3D model. The pre-existing structure is characterized by a hard wooden pitched ceiling, which has been considered untouchable by the architect, so our only possiblilty was to focus on the side and back walls. The side walls were subdivided in smaller surfaces, which were slightly tilted to achieve an absorbing-diffusing capacity at the various frequencies and the ceiling has been considered as a multiple scattering surface.
Phase II: surfaces acoustic properties definition. We defined the acoustic characteristics of the materials for each surface element, as well as the absorption coefficients α and the scattering coefficients σ for frequencies ranging from 125 Hz to 4,000 Hz. An iterative process provided a first generical response through CATT-Acoustic simulation software. Figure 3 shows the main lateral walls subdivision in smaller surfaces and their tilting final result process; each surface is visually characterized by a different colour showing the 2 surface materials: perforated wooden panels or sound-absorbing material covered with stretched fabric panels.
Phase III: sources and receivers definition. After another iterative process, the potentially most critical source-receiver combinations were identified. With this important information it was then possible to re-calculate the acoustic parameters of the room.
Phase IV: simulation. The software calculated the sound decay curves over time (echogram) and the acoustic parameters for each specified source-receiver combinations. In accordance with the initial results, it was then possible to adjust the model in order to correct any performance deficit and increase the acoustic quality. Some changes were necessary as from the client requests, in order to obtain good results without compromising aesthetics and architectural choices.

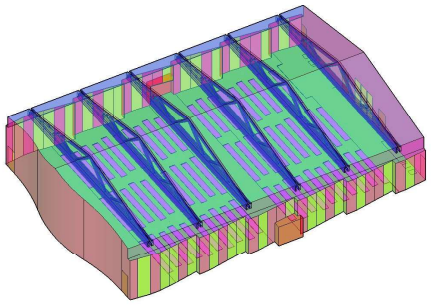


Figure 3. CATT-Acoustic simulation software 3D image of the hall: different colours define different absorbing-reflecting material

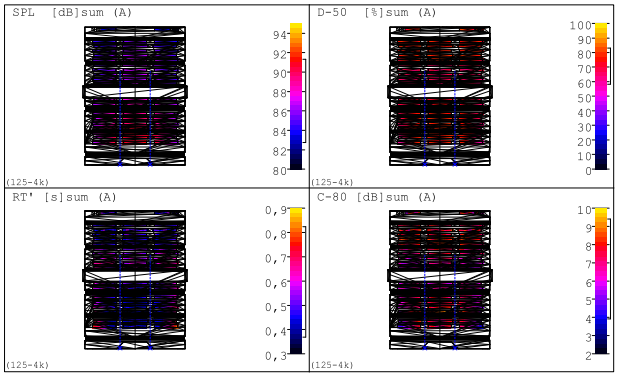
5 The sound source

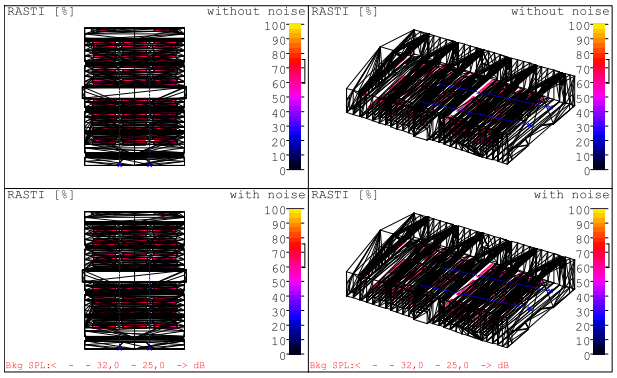
We considered a generic sound source with predefined directivity to take into account the audio system characteristics at an advanced design stage; we also considered two PA arrays installed on a truss above the speaker/ stage area. The client's request foresaw that the room would be mainly used for conferences and not for music, so we choose a voice signal as our main test sound signal. The average male-female spectral composition for a normal vocal effort is defined by the American standard ANSI S3.79 in terms of sound pressure level at a 1 meter distance along the source axis (Lp1m_a, expressed in dB), as shown in table 5:

Table 5. SPL average male-female spectral composition for normal vocal effort defined in ANSI S3.79

¤	125Hz¤	250Hz¤	500Hz¤	1kHz¤	2kHz¤	4kHz¤	8kHz¤	16kHz
Lp1m_a¤	51.2¤	57.2¤	59.8¤	53.5¤	48.8¤	43.8¤	38.8¤	33.8¤

These levels have been prudently raised to approximately 85 dB at 500 Hz, while maintaining the energy ratios between the other octave bands. We also performed Full Detailed Calculation and Early Part Detailed ISM in order to discover the potentially most critical source-receiver combinations. Based on these data it was possible to identifie and calculate the final room acoustic parameters. Fortunately the possible critical reflected sound signals with a delay time higher than the "*echo threshold*" of 50 ms did not arise. A subsequent analysis was also carried out through Audience Area Mapping software. Figures 4a and 4b visually illustrate the acoustic parameters spread final results for the most significant source-receiver combinations.





Figures: 4a SPL D50, RT and C80 Acoustic parameters final results for the most significant source - receiver combination; 4b. RASTI Acoustic parameters final results for the most significant source - receiver combination

6 Final remarks

At the end of the iterative process all the D_{50} values resulted excellent (> 60%) and evenly spread over the listening area as well as the C_{80}, which were suitable for speech listening (> 3 dB). The Bass Ratio or LF values were only between 6 and 30%, very variable due to the too smooth untreatable wodden ceiling. This was compensated by very good and uniform STI and RASTI parameters across the whole room. The SPL value also appears within the recommended limits (-4/5 dB maximum) throughout the whole room.
In this paper we demonstrate that an attentive software implementation procedure and a good acoustician's evaluationis able to spring to good room acoustics for the needed purpouse.

Literatura

[1] F. Nucibella, Studio e analisi acustiche di uno spazio polivalente, PhD. thesis, Parma 2012

[2] https://www.dw.com/en/german-sound-engineers-to-improve-sydney-opera-house-acoustics/a-19470173

[3] https://www.catt.se/

[4] M. Long, Architectural Acoustics, Academic Press Inc, 2nd edition 2013

[5] V.O. Knudsen, Cyril M. Harris, Acoustical Designing in Architecture, John Wiley & Sons Inc, First edition, 1950

[6] Makrinenko, Leonid I., Acoustics of Auditoriums in Public Buildings, Acoustical Society of America, 1994

[7] http://www.concerthalls.org/?page_id=90

Uporaba prostorskega zvoka v avtomatiziranih vozilih

Harsh Sanghavi [1], Kristina Stojmenova[2], Chihab Nadri [1], Sangjin Ko[1], Myounghoon Jeon[1], Jaka Sodnik[2]

[1]Mind Music Machine Lab, Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia
[2] Fakulteta za elektrotehniko, Univerza v Ljubljani, Slovenija

E-pošta: harshks@vt.edu, kristina.stojmenova@fe.uni-lj.si, cnadri@vt.edu, sangjinko@vt.edu, myounghoonjeon@vt.edu, jaka.sodnik@fe.uni-lj.si

Abstract. *This study explores the use of spatial sound displays for communicating takeover requests in automated vehicles with lower levels of autonomy. It compares two directional sounds (from a free lane and from a lane with a hazard) with a traditional non-directional auditory display, which plays the sound from the center of the vehicle. The study was conducted in a motion-based driving simulator in two locations – Slovenia and Virginia. All participants were engaged in four 9-minute driving tasks in level-3 automated vehicle. Each driving session contained three hazardous events with takeover request (in total 12 requests per user). The results showed shorter reaction times for the directional sound displays compared to the non-directional sound display. Regardless of the sound display, the participants in Slovenia had statistically significantly shorter reaction times compared to participants in Virginia.*

1 Uvod

Tehnologija avtonomnih vozil je v zadnjem desetletju prevzela osrednjo vlogo v avtomobilskih raziskavah. Razlog za veliko zanimanje za to področje so nedvomno vse predvidene prednosti in varnostne lastnosti, ki naj bi jih avtonomna vozila imela v primerjavi z današnjimi vozili z ročnim upravljanjem. Številne študije so navedle človeški dejavnik kot glavni vzrok prometnih nesreč in sicer v skoraj 95% raziskovanih primerov [1-3], kar upravičuje potrebo po zmanjšanju in sčasoma odpravi človeškega nadzora nad vozilom. Medtem ko je treba popolnoma avtonomno vozilo (stopnja SAE 5 [4]) še razviti, so polavtomatska vozila že na voljo, njihove stopnje avtonomnosti pa vztrajno naraščajo. Takšna vozila pa kljub neprimerno večji varnosti od vozil z ročnim upravljanjem še vedno zahtevajo prisotnost in pozornost voznika v posebnih voznih situacijah. Zelo pomembno je, da vozniku v takšnih primerih zagotovijo varen in hiter prehod iz avtonomnega v ročni način, kar je tudi osrednja problematika in motivacija te raziskave. Na tem področju je bilo izvedenih že veliko raziskav, ki so se osredotočile predvsem na sposobnost voznika za prevzem kontrole ob sproženi zahtevi [5-9]. V tej študiji preučujemo možnosti uporabe prostorskega zvoka, kot opozorila vozniku za prevzem kontrole polavtomatskega vozila in ga primerjamo z najpogosteje uporabljenimi neprostorskimi zvočnimi obvestili.

2 Uporaba zvočnih obvestil za prevzem kontrole nad vozilom

Raziskave so pokazale, da lahko uporaba ustreznega prostorskega (lokaliziranega) zvoka izboljša odzivanje na ciljne dogodke voznika ali pešcev na cesti [10], [11]. Uporaba prostorskih opozorilnih zvočnih signalov za nevarnost trka je pokazala, da je prepoznavanje zvočnih opozoril pred nevarnostjo s sprednje in zadnje strani vozila učinkovitejše od taktilnih opozoril in omogoča signifikantno krajše odzivne čase na dogodke [10]. Podrobnejši pregled uporabe prostorskih zvočnih opozoril je pokazal tudi, kako pomembna so taka opozorila pri informiranju o lokaciji določene nevarnosti in preusmerjanju pozornosti na določeno smer [12]. Raziskave zvočne percepcije pešcev so prav tako pokazale pozitivne učinke uporaba prostorskega zvoka, saj so le-ti zaznali smer prometa s hitrostjo večjo od 19 km/h (12 milj/h) s približno 90-odstotno natančnostjo, kar kaže na pomembnost lokalizirane zvočne informacije [11].

Uporaba usmerjenih zvočnih opozoril pri voznikih v križišču je prav tako pokazala pozitivne učinke, saj se je znatno zmanjšal čas do začetka zaviranja voznikov in tako tudi čas za upočasnitev vozila [12]. Poleg tega so raziskave z uporabo smernosti in modalnosti zvoka pokazale, da so prostorsko usmerjeni zvočni signali, ki nakazujejo smer nevarnosti bistveno izboljšali odzivanje voznikov v primerjavi z neusmerjenimi prikazi, pri čemer je najboljše rezultate dosegel hibridni vizualno-zvočni opozorilni sistem [13]. Marshall in drugi [14] so opozorili tudi na to, da lastnosti zvoka, kot sta glasnost in uporaba visokih tonov s svojo nadležnostjo lahko pomembno vplivajo na zaznano nujnost neke kritične situacije. Navajajo tudi, da ljudje zelo glasne in visoke tone ocenjujejo kot ustreznejše, hkrati pa tudi zelo moteče. Zvočna opozorila z visoko stopnjo nujnosti lahko voznikom pomagajo, da se učinkoviteje odzovejo [15]. Študije so tudi pokazale, da se ljudje hitreje odzovejo na zvočna opozorila, če zvenijo nujno [16], [17]. Pomembna lastnost zvočnega opozorila je tudi njegova intuitivnost, kar pomeni zmožnost prepoznavanja pomena v skladu z uporabnikovimi pričakovanji, predznanjem in splošnim mentalnim modelom [18].

Pozitivni rezultati uporabe prostorskih zvokov v informacijskih sistemih v vozilih so motivirali zasnovo naše študije, ki se je osredotočila na raziskovanje glavnih prednosti in pomanjkljivosti usmerjenih informacij v zvočnih zaslonih za sporočanje zahtev po prevzemu kontrole nad vozilom. Nadalje smo raziskali tudi kulturni vpliv na prevzem kontrole nad vozilom in način uporabe polavtomatskih vozil. Ker je trenutno v ZDA splošno višja stopnja sprejemanja takšnih polavtomatskih vozil v primerjavi z Evropo, smo se odločili za izvedbo enake študije v Virginiji (ZDA) in Sloveniji ter tako primerjali rezultate o vozniških lastnostih na obeh lokacijah.

3 Metodologija

Poskus je bil izveden na simulatorju vožnje podjetja Nervtech [19]. Sestavljen je iz dirkalnega avtomobilskega sedeža, volanskega sistema s pravim volanom in športnih stopalk. Prometni simulacijski scenarij je bil prikazan na treh 48-palnih zaslonih, ki so omogočali 120° vodoravnega vidnega polja.
Vozni scenariji so vključevali vožnjo po tripasovni avtocesti, kjer je bila omejitev hitrosti 110 km/h (oziroma 70 milj/h v Virginiji). Vidnost na cesti se je s pomočjo megle znižala na približno 100 metrov (slika 1).



Slika 1. Zmanjšana vidljivost z uporabo megle.

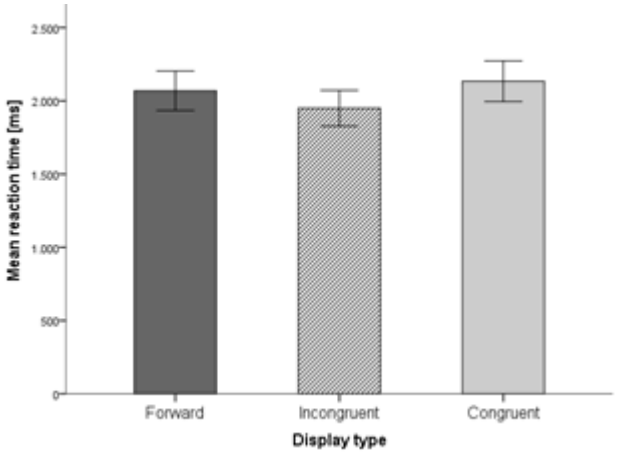
Za študijo je bila uporabljena simulacija avtomatiziranega vozila s 3. stopnjo avtonomnosti po SAE [4]. Avtomatski sistem vožnje (ADS) je bil vključen ves čas, razen v nepričakovanih situacijah (na primer prometna nesreča, divjad na cesti, gradbena dela itd.). Ko je bil ADS vklopljen, je vozilo vedno vozilo po srednjem pasu s konstantno hitrostjo 110 km/h. Testiranci so lahko sistem ADS kadar koli izklopili in ponovno vklopili s preprostim potegom leve ročice na volanu. Nepričakovana situacija je vedno povzročila zaporo dveh od treh voznih pasov (srednjega in levega ali desnega pasu), tako da je moral voznik zamenjati vozni pasu ali zavirati in tako ustaviti vozilo, da bi se izognil nesreči.
V študiji je sodelovalo 72 udeležencev - 36 v Sloveniji in 36 v Virginiji. Udeleženci so bili razdeljeni v 3 skupine, od česar je 24 udeležencev študijo zaključilo z neusmerjenim zvočnim zaslonom, 24 z zaslonom, ki je oddajal zvok iz smeri voznega pasu, kjer se je nahajala ovira in 24 z zaslonom, ki je oddajal zvok iz smeri prostega voznega pasu.
V študiji so sodelovali izključno udeleženci z veljavnimi vozniškimi dovoljenji. Nihče od udeležencev ni imel težav s sluhom.

Poskus je vključeval ogrevalno vožnjo in štiri 9-minutne preizkuse. Ogrevalna vožnja je trajala približno 3 minute in je bila namenjena udeležencem, da se seznanijo s simulatorjem vožnje, sistemom ADS, zvočnim zaslonom in igrico 2048. Igrica 2048 je bila namenjena odvračanju pozornosti s ceste v času, ko je bil vklopljen ADS sistem. Udeleženci so igrali 2048 na tabličnem računalniku, ki so ga v času ročnega upravljanja vozila enostavno odložili v naročje ali na naslonjalo za roke v vozilu. V vsakem od štirih preizkusov so bili udeleženci izpostavljeni trem kritičnim situacijam, v katerih so morali prevzeti nadzor nad vozilom. Čas dogodka in lokacija prostega pasu sta bila naključno izbrana, da bi se poskušali izogniti učinkom učenja (ang. learning effect) in predvidevanja (ang. anticipation effect). Ker je imelo vozilo visoko stopnjo avtomatizacije, so morali udeleženci prevzeti kontrolo nad vozilom le v primeru onemogočene avtomatizacije vozila zaradi nepričakovane situacije ali ovire. Ko se je pojavila ovira, je voznik prejel zahtevo s pomočjo zvočnega opozorila (prikaza). V raziskavi so bili uporabljeni trije tipi zvočnih opozoril za prevzem kontrole: en neusmerjen in dva usmerjena zvoka. Neprostorski oz. neusmerjen zaslon je predvajal zvočno obvestilo z sredine vozila. Za preskuse z usmerjenim zvokom smo uporabili kongruentna in inkongruentna obvestila - zvoke iz smeri voznega pasu z oviro ali zvoke iz smeri prostega pasu. Zvočno opozorilo je bilo sestavljeno iz dveh prevladujočih frekvenc (880, 1760 Hz), ki sta se štirikrat ponovili. Opozorilni zvok se je predvajal z intenziteto 65 dB, zvok okolice pa s 60 dB. Po predvajanju zahteve za prevzem je avtonomno vozilo prešlo v neavtonomni način. To je vozniku omogočilo nadzor nad vozilom. Pri ročnem upravljanju je vozilo imelo sistem avtomatskega menjalnika.

4 Rezultati študije

Leveneov test je bil uporabljen za oceno enakosti odstopanj, test Shapiro-Wilk pa za oceno normalne porazdelitve vsake skupine podatkov. Na podlagi teh rezultatov so bili za analizo podatkov uporabljeni ustrezni parametrični in neparametrični testi.
Analiza je pokazala ugodnejše rezultate ob uporabi usmerjenih zvočnih obvestil v primerjavi z neusmerjenim zvokom. Reakcijski časi so bili najkrajši za prostorski zvok, ki se je predvajal s smeri voznega pasu z oviro (M = 1950,1 ms, SD = 1004,66 ms), sledil mu je neprostorski zvok (M = 2068,98 ms, SD = 1101,52 ms), medtem ko so bili za usmerjen zvok iz prostega voznega pasu najdaljši reakcijski časi (M =

2133,90 ms, SD = 1095,87 ms). Razlike v primerjanih povprečnih reakcijskih časih za različne vrste zaslonov (prostorski v primerjavo z neprostorskim) niso bile statistično pomembne F (2, 819) = 2,002, p = 0,136 (slika 2).



Slika 2. Povprečni reakcijski časi na različne opozorilne zvoke ob zahtevi po prevzemu kontrole nad vozilom

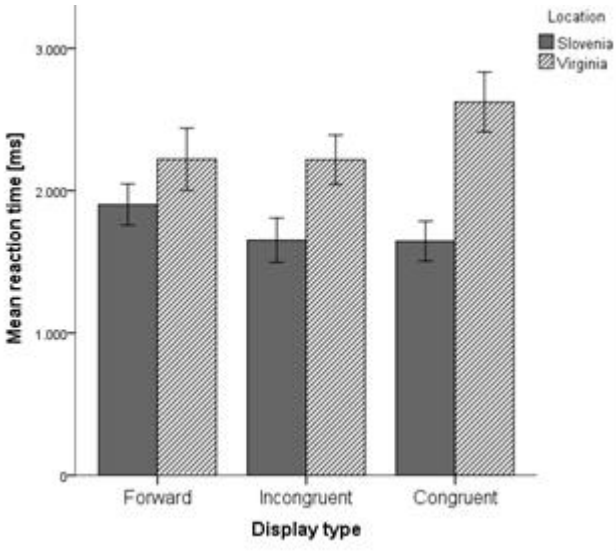
Analiza za Slovenijo je pokazala, da je za usmerjene zvoke statistično pomembno zmanjšanje povprečnih reakcijskih časov (povprečje vseh reakcij iz obeh usmerjenih zaslonov) (M = 1648,42 ms, SD = 853,63 ms) v primerjavi z neusmerjenimi zvoki (M = 1902,18 ms, SD = 815,69 ms), F (1, 395) = 7,851, p = 0,005. Kljub temu, da je bil za Virginio trend podoben slovenskemu, so rezultati iz Virginije pokazali, da ni bilo statistično pomembnih razlik v povprečnem reakcijskem času med neusmerjenim zvokom (M = 222,29 ms, SD = 1292,42 ms) in usmerjenimi zvoki (M = 2419,05 ms, SD = 1163,95 ms), F (1, 423) = 2.543, p = 0,112.

Pri 88,3 % udeležencev je bila prva reakcija zaviranje v primerjavi z le 11,7 % udeležencev, ki so odreagirali z vrtenjem volana. Podrobnejši pregled podatkov za različne lokacije je pokazal, da je v Sloveniji 76,6 % udeležencev ob zvočnem opozorilu najprej zaviralo, 23,4 % pa jih je zavrtelo volan. Zanimivo so v Virginiji prav vsi udeleženci najprej zavirali. Poleg tega so imeli udeleženci v Sloveniji neglede na tip zvočnega opozorila statistično bistveno krajše reakcijske čase v primerjavi z udeleženci v Virginiji, $F(2, 816) = 7{,}053$, $p = 0{,}001$. (slika 3). Post-hoc rezultati so pokazali, da so:

- udeleženci v Sloveniji imeli krajše reakcijske čase (M = 1902,18 ms, SD = 92,42 ms) na neusmerjene v primerjavi z udeleženci v Virginiji (M = 2220,29 ms, SD = 88,02 ms), p = 0,011;

- udeleženci v Sloveniji imeli krajši reakcijski čas (M = 1651,69 ms, SD = 92,05 ms) tudi na usmerjen zvok iz smeri voznega pasu z oviro v primerjavi z udeleženci v Virginiji (M = 2217,2 ms, SD = 87,09 ms), p < 0,001;

- udeleženci v Sloveniji imeli krajši reakcijski čas (M = 1645,47 ms, SD = 87,4 ms) za usmerjen zvočni prikaz iz smeri prostega pasu v primerjavi z Virginijo (M = 2622,32 ms, SD = 87,4 ms), p = 0,001.



Slika 3. Povprečni reakcijski časi različnih zvokov v Sloveniji in Virginiji.

5 Zaključki

V študiji smo za potrebe načrtovanja boljših zaslonov in zvočnih obvestil za prevzem kontrole nad vozilom v avtomatiziranih vozilih ocenjevali tipe prostorskih zvočnih opozoril. Rezultati študije so pokazali, da so usmerjeni zvoki izzvali hitrejše reakcijske čase v primerjavi z neusmerjenim zvokom. Ko smo opazovali razlike v reakcijskih časih med obema usmerjenima prostorskima zvokoma, razlik nismo našli. Ne glede na zvočni signal so imeli udeleženci v Sloveniji statistično bistveno krajše reakcijske čase v primerjavi z udeleženci v Virginiji. Opazna razlika je bila tudi pri načinu prevzema kontrole nad vozilom. Med tem je ko vsak četrti voznik v Sloveniji najprej prijel za volan in zamenjal pas, so se vsi testirani vozniki v Virginiji vsakič odzvali na opozorilo le z zaviranjem.
Te razlike med obema opazovanima mestoma kažejo, da kulturni dejavnik pomembno prispeva k voznikovemu vedenju v polavtomatskih vozilih. To kaže, da bi moralo prepoznavanje takšnih kulturno odvisnih razlik v vozniškem vedenju igrati pomembno vlogo v procesu načrtovanja polavtomatskih vozil za njihovo varno uporabo v različnih mestih po svetu.

Zahvala

To raziskavo je finančno podprla Slovenska raziskovalna agencija v okviru raziskovalnega programa »ICT4QoL - informacijske in komunikacijske tehnologije za kakovost življenja«, št. P2-0246 in bilateralnega raziskovalnega projekta »Vpliv kulturnih razlik v modeliranju vedenja voznikov: Vloga voznika

pri različnih stopnjah avtomatizacije vozil v ZDA in Sloveniji.«, št. BI-US/19-21-008.

Literatura

[1] Boyce, T.E. and E.S. Geller. An instrumented vehicle assessment of problem behavior and driving style: Do younger males really take more risks? Accident Analysis & Prevention, 2002. 34(1): str. 51-64.

[2] Smith, B.W. Human Error as a Cause of Vehicle Crashes. 2013; Dostopno na: http://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes.

[3] Sabey, B.E. and H. Taylor, The Known Risks We Run: The Highway, in Societal Risk Assessment, R.C. Schwing and W.A. Albers, Editors. 1980, Springer: Boston, MA.

[4] SAE International. SAE J3016: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. 2014. Dostopno na: http://standards.sae.org/j3016_201401/.

[5] N. Merat, A. H. Jamson, F. C. H. Lai, and O. Carsten, "Highly automated driving, secondary task performance, and driver state," Hum. Factors, vol. 54, št.. 5, str. 762–771, Oct. 2012.

[6] C. Gold, D. Damböck, L. Lorenz, and K. Bengler, "'Take over!' How long does it take to get the driver back into the loop?," Proc. Hum. Factors Ergon. Soc. Annu. Meet., vol. 57, št. 1, str. 1938–1942, Sep. 2013.

[7] K. Zeeb, A. Buchner, and M. Schrauf, "What determines the take-over time? An integrated model approach of driver take-over after automated driving," Accid. Anal. Prev., vol. 78, str. 212–221, 2015.

[8] V. Melcher, S. Rauh, F. Diederichs, H. Widlroither, and W. Bauer, "Take-Over Requests for Automated Driving," Procedia Manuf., vol. 3, str. 2867–2873, 2015.

[9] D.-I. D. Damböck, D.-I. M. Farid, D.-S. L. Tönert, and K. Beng-Ler, "Übernahmezeiten beim hochautomatisierten Fahren."

[10] C. Ho, H. Z. Tan, and C. Spence, "The differential effect of vibrotactile and auditory cues on visual spatial attention," Ergonomics, vol. 49, št. 7, str. 724–738, Jun. 2006.

[11] B. K. Barton, T. A. Ulrich, and R. Lew, "Auditory detection and localization of approaching vehicles," Accid. Anal. Prev., vol. 49, str. 347–353, Nov. 2012.

[12] C. Ho and C. Spence, "Assessing the effectiveness of various auditory cues in capturing a driver's visual attention," J. Exp. Psychol. Appl., vol. 11, št. 3, str. 157–174, Sep. 2005.

[13] Y. Zhang, X. Yan, and Z. Yang, "Discrimination of Effects between Directional and Nondirectional Information of Auditory Warning on Driving Behavior," Discret. Dyn. Nat. Soc., vol. 2015, str. 1–7, Mar. 2015.

[14] Y. C. Liu and J. W. Jhuang, "Effects of in-vehicle warning information displays with or without spatial compatibility on driving behaviors and response performance," Appl. Ergon., vol. 43, št. 4, str. 679–686, 2012.

[15] D. C. Marshall, J. D. Lee, and P. A. Austria, "Alerts for In-Vehicle Information Systems: Annoyance, Urgency, and Appropriateness," Hum. Factors J. Hum. Factors Ergon. Soc., vol. 49, št. 1, str. 145–157, 2007.

[16] J. L. Burt, D. S. Bartolome, D. W. Burdette, and J. R. Comstock, "A psychophysiological evaluation of the perceived urgency of auditory warning signals," Ergonomics, vol. 38, št. 11, str. 2327–2340, 1995.

[17] E. C. Haas and J. G. Casali, "Perceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise," Ergonomics, vol. 38, št. 11, str. 2313–2326, 1995.

[18] S. Garzonis, S. Jones, T. Jay, and E. O'Neill, "Auditory icon and earcon mobile service notifications," 2009, str. 1513.

[19] Vengust, M., Kaluža, B., Stojmenova, K. and Sodnik, J. NERVteh Compact Motion Based Driving Simulator. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct, 2017, str. 242-243. ACM.

Detekcija infrazvoka z zvočnikom v funkciji mikrofona

Samo Beguš, Gregor Geršak

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: samo.begus@fe.uni-lj.si

Infrasound detection with loudspeaker operating as microphone

In this paper we present a novel method for detection of infrasound using an ordinary loudspeaker. We tested 6 different loudspeakers in three configurations: loudspeaker alone with no enclosure, loudspeaker system with a compression type enclosure and loudspeaker system with a bass reflex enclosure, as the most common loudspeaker design. The frequency response of the loudspeakers have been measured and analysed in view of a possible use as an infrasound sensor - microphone. Loudspeakers were shown that could be used, especially when expected infrasound levels are high. Sensitivities and frequency responses are provided.

1 Vpliv infrazvoka na človeka

V zadnjih letih je opažen in dokumentiran negativni vpliv zvokov infrazvočnih in ultrazvočnih frekvenc na zdravje ljudi, ki so mu (dolgotrajno) izpostavljeni [1-4]. V vsakdanjem življenjskem okolju je vedno več generatorjev infrazvoka (npr. promet, vetrne elektrarne [5]), hkrati je v veljavi le malo priporočil in varnostnih standardov, ki urejujejo to področje [6, 7]. Zato so raziskave o vplivu ultrazvoka in infrazvoka na človeka zelo pomembne in potrebne, prav tako merjenje izpostavljenosti infrazvoku.

Aplikacije na pametnih telefonih omogočajo zaznavanje in merjenje nivoja zvočnega tlaka. Točnost, ponovljivost in zanesljivost delovanja aplikacij je odvisna od več parametrov – strojne opreme, programske opreme, validacije strojne in programske opreme ter okoliških parametrov. To velja še posebno za aplikacije, namenjene zaznavanju infrazvoka [8].

2 Zaznavanje in merjenje infrazvoka

Za zaznavanje in merjenje infrazvoka lahko uporabimo fonometre z ustreznimi mikrofoni, po možnosti z razširjenim frekvenčnim območjem v področju infrazvoka.

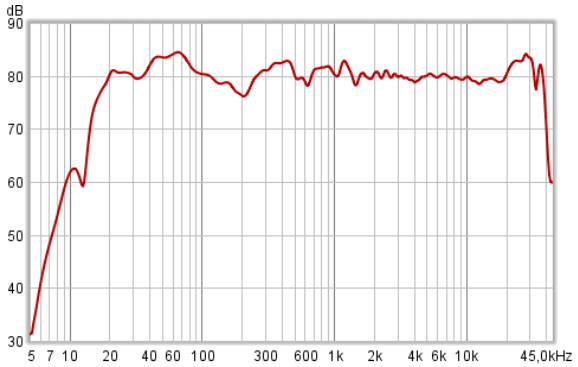
Merilni mikrofon proizvajalca Brüel & Kjær, tip 4191 s spodnjo frekvenčno mejo -3 dB od 1 Hz do 2 Hz [9] je široko uporabljen in vsestranski mikrofon. Podobno velja tudi za dva namenska mikrofona za merjenje infrazvoka, predpolarizirani tip 4964 istega proizvajalca, s spodnjo frekvenčno mejo -3 dB od 0,01 do 0,05 Hz [10] in nepolarizirani kondenzatorski mikrofon tip 4193 [11]. Uporabijo se lahko keramični mikrofoni [12] in predelani splošno uporabni elektret-kondenzatorski mikrofoni [13]. Takšne in podobne konfiguracije zahtevajo namenski mikrofon, s katerim ob ustrezni strojni in programski opremi lahko zagotovimo sledljive meritve.

Za enostavno detekcijo, kjer želimo le zaznavanje infrazvoka, bi lahko uporabili nenamenske, enostavne in cenovno dostopne mikrofone. Cilj prispevka je preveriti, če je možno zaznavati infrazvok z običajnim zvočnikom, ki je sestavni del vsakega avdio sistema.

Izmerili smo šest različnih zvočnikov: brez ohišja, s kompresijskim ohišjem in z bas-refleks ohišjem. Izmerili smo njihovo frekvenčno odvisnost in občutljivost.

3 Metode

Meritve občutljivosti zvočnikov v funkciji mikrofona smo izvedli v gluhi sobi na Fakulteti za elektrotehniko (FE). Kot vir zvoka smo uporabili kombinacijo nizkotonskega zvočnika Genelec 7050C [14] in monitor zvočnika Genelec 8040A [15]. Njun skupni frekvenčni odziv je prikaza na sliki 1.



Slika 1. Frekvenčni odziv sistema za generiranje zvoka v gluhi sobi na FE.

Referenčne meritve uporabljenih zvočnih virov smo izvedli z merilnim kondenzatorskim mikrofonom Brüel & Kjær tip 4191 s predojačevalnikom Brüel & Kjær tip 2669 in Nexus ojačevalnikom. Za zajemanje in generiranje testnih signalov smo uporabili USB zvočno kartico Creative E-MU 0404 pri frekvenci vzorčenja 96 kHz in 24 bitni globini ter 23 dB ojačanja. Zvočnik smo priključili na mikrofonski, XLR vhod. Spodnja frekvenčna meja zvočne kartice -3 dB je pri frekvenci 2 Hz. Za izvajanje meritev in obdelavo smo uporabili program Room Equalization Wizard (REW) [16], Microsoft Excel [17] in Audacity [18].

Glede na omejitve sistema za generiranje zvoka, zvočne kartice, mikrofona in mikrofonskega predojačevalnika smo vse analize in prikaze omejili pri spodnji frekvenčni meji 5 Hz. Tu bi opozorili še na pojav presluha električnega signala med vhodoma zvočne

kartice pri nizkih frekvencah, pod 5 Hz. Posledično smo uporabljali vedno le en vhod.

3.1 Potek meritve frekvenčnega odziva in občutljivosti

Nivo jakosti hrupa v gluhi sobi je do 12 dBA in okrog 45 dB, zato je pri meritvah pri nizkih frekvencah in nivojih jakosti <40 dB potrebno uporabiti metode za merjenje s pomočjo ozkopasovne frekvenčne analize, da dobimo primerno razmerje signal/šum, da lahko ugotovimo občutljivost zvočnika in izločimo vpliv hrupa v gluhi sobi.



Slika 2. Fotografija postavitve v gluhi sobi na FE. Levo: merjeni zvočnik – mikrofon; desno: Genelec zvočnika.

Meritve frekvenčnega odziva smo izvedli v gluhi sobi, kjer smo namestili zvočnika za generiranje zvoka in izmerili frekvenčni odziv z mikrofonom B&K 4191. Fotografija postavitve v gluhi sobi je prikazana na sliki 2. Merili smo na razdalji 0,5 m, ker smo želeli doseči čim višji nivo zvočnega tlaka tudi pri nizkih frekvencah. Genelec zvočnika sta specificirana za frekvenčno območje (-3 dB) tip 8040A: <45 Hz ter >20 kHz in (-6 dB) za tip 7050C: 24 Hz do 85 Hz [14, 15]. Na takšni razdalji in pri nivoju zvočnega tlaka 60 dB smo lahko merili odziv do 8,8 Hz in pri 50 dB do 5,1 Hz.

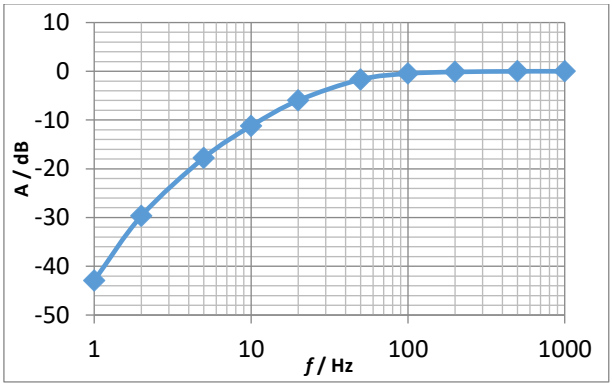
Tabela 1: Uporabljeni zvočniki v testu.

Oznaka zvočnika	Vrsta ohišja	Št. zv.	Velikost nizkotonskega zvočnika	Širina / cm	Višina / cm	Globina / cm
U	Brez	1	8 cm X 12 cm	12	8	5
K1	Kompresijsko	2	Φ 20 cm	28,5	38	20
K2	Kompresijsko	1	Φ 8,5 cm	15	24	15,5
B1	Bas refleks	2	Φ 12 cm	16	24	24
B2	Bas refleks	2	Φ 12 cm	14,5	23,8	23,4
B3	Bas refleks	2	Φ 16,6 cm	20,5	32,5	29

Pri merjenju odziva zvočnika kot mikrofona smo zamenjali merilni mikrofon 4191 z izbranim zvočnikom in izvedli meritev, občutljivost smo določili z upoštevanjem občutljivosti zvočne kartice in nastavljenega ojačanja. V tabeli 1 so zbrani podatki in uporabljene oznake zvočnikov.

Pri ročni meritvi občutljivosti izmerimo napetost na sponkah zvočnika z voltmetrom pri znanem nivoju zvočnega tlaka.

3.2 Zajem signalov z diktafonom

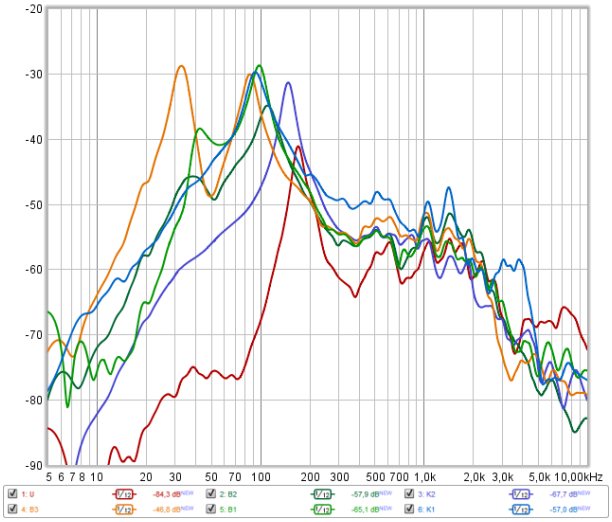


Slika 3. Frekvenčni odziv mikrofonskega vhoda diktafona Olympus LS11.

Namesto zvočne kartice kot sistema za zajem signalov smo preverili možnost uporabe diktafona z mikrofonskim vhodom. Na sliki 3 je prikazan frekvenčni odziv diktafona Olympus LS11 [19]. Iz grafa občutljivosti vidimo, da se zmanjša za 20 dB pri frekvenci 4 Hz. Sledi, da je sicer možno zaznavati infrazvok z zvočnikom in diktafonom s frekvenčno korekcijo, vendar z omejenim dinamičnim območjem zaradi slabše občutljivosti pri nizkih frekvencah.

4 Rezultati

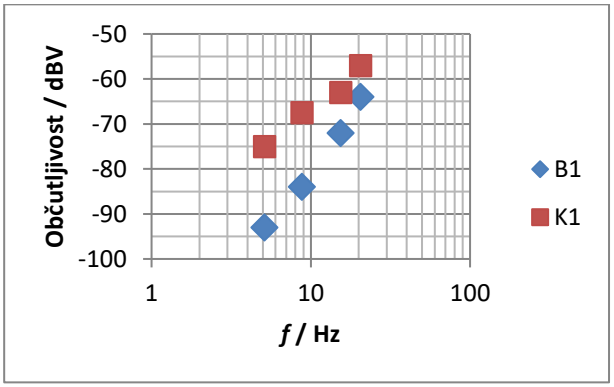
V nadaljevanju so predstavljeni rezultati meritev frekvenčnega odziva in občutljivosti.



Slika 4. Občutljivost zvočnikov v funkciji mikrofona izražena v dBV preračunano na nivo zvočnega tlaka 94 dB.

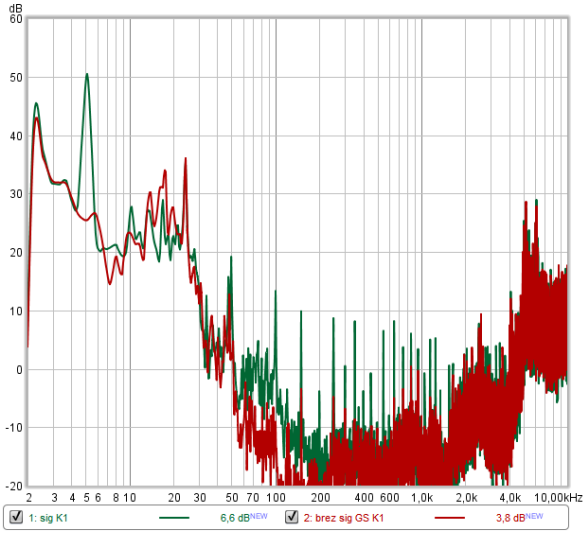
Na sliki 4 so prikazane občutljivosti in frekvenčne odvisnosti uporabljenih zvočnikov z oznakami posameznih zvočnikov, kot v tabeli 1 z 1/12 oktavnim glajenjem.

Občutljivost v dBV je bila izmerjena s programom REW, vrednosti pa korigirane na nivo zvočnega tlaka 94 dB. Pri nizkih frekvencah (pod 10 Hz; za zvočnik U pa že pod 90 Hz) je razmerje signal/šum prenizko za pravilen prikaz občutljivosti in frekvenčne odvisnosti na grafu, zato smo meritve v frekvenčnem območju od 5 Hz do 20 Hz izvedli ročno, amplitude signalov smo merili s pomočjo frekvenčno analize pri pasovni širini posamezne spektralne komponente 0,732 Hz. Rezultati za zvočnika K1 in B1 so prikazani na sliki 5.



Slika 5: Občutljivost zvočnikov K1 in B1 pri nizkih frekvencah.

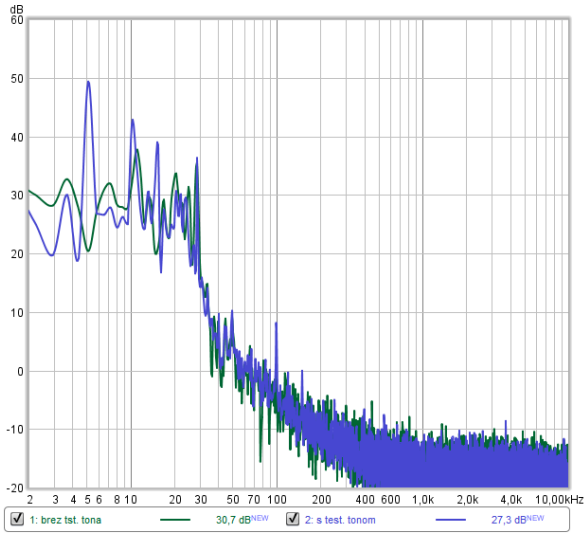
Primera spektra signala brez in s prisotnim testnim tonom pri frekvenci 5,13 Hz je prikazan na sliki 6, s frekvenčno korekcijo občutljivosti. Spekter hrupa izmerjen z merilnim mikrofonom B&K 4191 v gluhi sobi pa na sliki 7.



Slika 6: Spektra signala zajetega z zvočno kartico brez in s testnim tonom pri frekvenci 5,13 Hz z zvočnikom K1 v funkciji mikrofona. Uporabljena je frekvenčna korekcija občutljivosti.

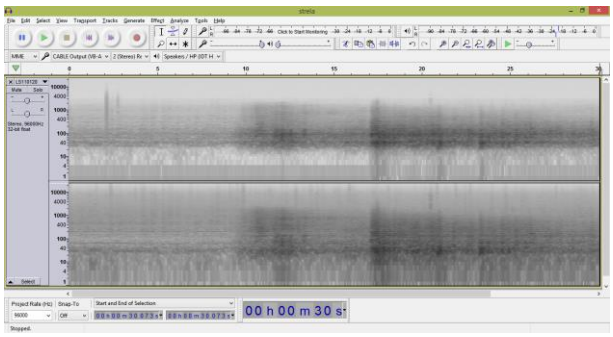
Merjenje dodatno otežuje hrup pri nizkih frekvencah v gluhi sobi, amplituda testnega signala pa je omejena z amplitudo generiranega signala z Genelec zvočniki pri nizkih frekvencah. Zvočnika delujeta izven specifikacij, pri frekvencah nižjih od 24 Hz.

Zvočniki z bas refleks izvedbo ohišja imajo pričakovano dve izraziti resonanci v nizkofrekvenčnem področju in upadanjem občutljivosti proti nizkim frekvencam, ostale konfiguracije pa izrazito resonanco in upad občutljivosti pri nizkih frekvencah in visokih frekvencah.



Slika 7: Spekter zvok z in brez testnega tona pri frekvenci 5,13 Hz izmerjen z merilnim mikrofonom B&K 4191 v gluhi sobi.

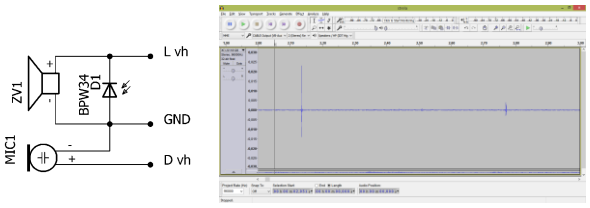
Povišan nivo spektra signala na sliki 7 pri prisotnem testnem tonu je posledica lastnega šuma Genelec zvočnikov. Prav tako se poviša nivo motenj omrežne frekvence (50 Hz in harmonske komponente) pri vključenem napajanju za zvočnike do 20 dB. Motnje se lahko zaduši z ustreznimi pasovno-zapornimi siti. Amplitudo motenj bi lahko zmanjšali tudi z uporabo koaksialnih kablov med zvočnikom in zvočno kartico (uporabljeni so bili neoklopljeni dvožilni kabli za zvočnike).



Slika 8: Spektrogram zvoka po udaru strele, prikazan s programom Audacity. Zgoraj: signal zvočnika v funkciji mikrofona; spodaj: mikrofon EM172. Prvi udar strele je ob času t = 2,137 s.

Kot primer uporabe zvočnika kot mikrofona za zaznavanje infrazovka je na sliki 8 prikazan spektrogram zvoka groma po udaru strele. Izračunani razdalji udara ob

upoštevanju temperature v okolici 15 °C sta 3,4 km in 5,2 km. Uporabljan je bil zvočnik B1 in mikrofon Primo EM172 s snemalnikom Olympus LS11. Zvočnik je bil priključena na L mikrofonski vhod, z vzporedno vezano fotodiodo, za hkratno indikacijo udara strele, slika 9; EM172 pa je bil priključen na D vhod snemalnika.



Slika 9: Levo: Priključitev zvočnika, mikrofona in fotodiode na snemalnik; desno: prikaz posnetega signala ob udaru strele.

S pomočjo fotodiode lahko enostavno določimo čas udara strele, slika 9, kjer se lahko vidi posamezne razelektritve.

Kljub temu, da na sliki 8 ni uporabljana kompenzacija frekvenčne odvisnosti zvočnika kot mikrofona je jasno opazen prisoten infrazvok v frekvenčnem območju do 20 Hz.

5 Zaključek

V prispevku smo pokazali, da lahko zaznavamo infrazvok z običajnim zvočnikom ter zvočno kartico kot sistemom za zajemanje in vključenim predojačevalnikom (ojačanje 23 dB). Izmerili smo šest različnih zvočnikov: brez ohišja, s kompresijskim ohišjem in bas-refleks ohišjem, njihovo frekvenčno odvisnost in občutljivost.

V prihodnjih študijah bi bilo dobro izmeriti ekvivalentni nivo šuma takšnega sistema, za kar pa bi potrebovali sistem za generiranje zvoka in infrazvoka z nižjo spodnjo frekvenčno mejo. Iz slike 6 lahko sicer ocenimo ekvivalentni nivo šuma cca 30 dB pri 5 Hz, a je ocena nezanesljiva zaradi hrupa v gluhi sobi.

Literatura

[1] European Environment Agency, Noise in Europe 2014, Copenhagen, 2014, ISBN 978-92-9213-505-8.

[2] D. Krahé, D. Schreckenberg, F. Ebner, C. Eulitz, U. Möhler, Machbarkeitsstudie zur Wirkung von Infraschall, Umweltbundesamt, Dessau, 2014, http://www.umweltbundesamt.de/publikationen/machbarkeitsstudie-zu-wirkungen-von-infraschall.

[3] Salt, A. N. & Lichtenhan, J. T., How Does Wind Turbine Noise Affect People?, Acoustics Today, 2014, 10, 20 – 27.

[4] Health Protection Agency (UK) 2010: Health effects of exposure to ultrasound and infrasound (RCE-14) – Report of the independent Advisory Group on Non-ionising Radiation. ISBN 978-0-85951-662-4. Online available via the website of the Health Protection Agency http://www.hpa.org.uk/

[5] Branko Zajamšek, Kristy Hansen; Dobre prakse merjenja ter modeliranja hrupa vetrnih elektrarn, objavljeno v: Dovjak, Mateja (glavni in odgovorni urednik, član uredniškega odbora, drugo). Hrup vetrnih elektrarn in mogoči vplivi na življenjsko okolje = Wind turbine noise and possible impacts on living environment : mednarodni znanstveni posvet = international scientific consultation, Ljubljana, 15. 1. 2020 : zbornik prispevkov. Ljubljana: Slovensko društvo za akustiko: Fakulteta za gradbeništvo in geodezijo, 2020. 1 spletni vir (1 datoteka PDF (80 str.)), ilustr.

[6] ISO 7196:1995-3: Acoustics - Frequency-weighting characteristic for infrasound measurements, ISO 1995

[7] DIN 45680:2013-09: Messung und Beurteilung tieffrequenter Geräuschimmissionen (Norm-Entwurf), Deutsches Institut für Normung, 2013. <Measurement and assessment of low-frequency noise immissions (Draft Standard)>

[8] InfraSound Detector https://play.google.com/store/apps/details?id=com.microcadsystems.serge.infrasounddetector&hl=sl

[9] https://www.bksv.com/-/media/literature/Product-Data/bp2212.ashx 4191

[10] https://www.bksv.com/-/media/literature/Product-Data/bp2478.ashx 4964

[11] https://www.bksv.com/-/media/literature/Product-Data/bp2214.ashx 4193

[12] M. Watson, B. McIntosh and D. ReVelle, "A meteor infrasound recording system," *ICASSP '76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, 1976, pp. 786-789, doi: 10.1109/ICASSP.1976.1170020.

[13] Takanao Sugimoto, Kenji Koyama, Yosuke Kurihara and Kajiro Watanabe, "Measurement of infrasound generated by wind turbine generator," 2008 SICE Annual Conference, Tokyo, 2008, pp. 5-8, doi: 10.1109/SICE.2008.4654612.

[14] https://www.genelec.com/7050c

[15] https://www.genelec.com/previous-models/8040a

[16] https://www.roomeqwizard.com/

[17] https://www.microsoft.com

[18] https://www.audacityteam.org/

[19] https://www.olympus.co.uk/site/en/archived products/audio/audio_recording_1/ls_11/index.pdf

Vodno žigosanje audio vsebin z uporabo diskretne valčne transformacije (DWT)

Nejc Bevk, Urban Burnik, Marko Meža

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: ne.bevk@gmail.com

Audio watermarking based on Discrete Wavelet Transform (DWT)

Abstract. *In the article, we first introduce the motivation for watermarking of digital content. We highlight the key features of watermarking systems and some of the most popular techniques and algorithms in use according to the literature. We focus on blind watermarking based on discrete wave transformation. We reviewed the watermark insertion and extraction algorithm in detail and performed validation on a speech recording and on a rock music record. Our findings demonstrate that the watermarking of speech recording is slightly more robust and resistant to signal manipulations.*

1 Uvod

Sodobna digitalna tehnologija široki populaciji omogoča preprosto izmenjavo zvočnih posnetkov v odprtih podatkovnih omrežjih. Ilegalen prenos avtorsko zaščitenih del postaja velik problem tudi v glasbeni industriji, kjer predstavlja dejansko izgubo finančnih sredstev. Potrebujemo postopke, s katerimi lahko spremljamo prenose, identificiramo lastnika, dokažemo lastništvo, overimo vsebino in nadziramo kopiranje zvočnih datotek. Ena izmed rešitev je ta, da v signal zvočnega posnetka na čim manj zaznaven način vgradimo meta podatke, ki jih je kasneje mogoče prebrati oziroma dekodirati. Tak postopek imenujemo vodno žigosanje vsebin, vgrajeni vodni žig pa nam služi kot dokazilo o izvoru označenega posnetka. Informacije morajo biti vdelane tako, da jih kasneje lahko odkrijemo in na njihovi podlagi dokažemo pristnost ali avtorstvo posnetka. V primerjavi z steganografijo ali kriptografijo nam vodno žigosanje omogoča, da digitalni podatki kljub žigosanju ohranijo normalno uporabnost oziroma, da dostop do njih ni tehnološko omejen [1]. Glavne lastnosti sistemov za vodno žigosanje so slišnost, robustnost, varnost, podatkovni tovor in računska zahtevnost [2]. Pomembnost kriterijev se spreminja z namenom uporabe vodnega žiga.

1.1 Lastnosti sistema za vodno žigosanje

Slišnost ali zaznavnost – je lastnost, s katero ocenjujemo, kako zaznaven je vodni žig v označeni vsebini. Vodno žigosanje naj bi bilo čim bolj transparentno in naj ne bi kvarilo zaznavanja originalnega signala. V ta namen uporabljamo psiho-akustične modele, ki izkoriščajo lastnosti človeškega zaznavanja zvoka [2].

Robustnost – je lastnost s katero opisujemo, če vodni žig po različnih obdelavah signala še vedno pravilno zaznavamo. Gre za mero zanesljivosti vodnega žiga ob različnih namernih ali nenamernih napadih, ki vplivajo na signal. Klasični primeri tovrstnih procesov so dodajanje šuma, prevzorčevanje, razne kompresije ali druge modifikacije avdio signala. Pomembno je omeniti, da je stopnja robustnosti močno pogojena z namenom aplikacije vodnega žiga [2] [1].

Varnost – sistema za vodno žigosanje dosežemo tako, da podatke, ki jih bomo žigosali, že pred ali med postopkom ustrezno šifriramo. V ta namen lahko ustvarimo tajne ključe ali uporabimo različne šifrirne tehnike, s katerimi v proces vdelave in detekcije vodnega žiga dodamo naključnost. Tako morebitnim napadalcem preprečimo, da bi zaznali ali uničili vodni žig s statistično detekcijo ali obratno uporabo algoritma [2] [3].

Podatkovni tovor – se običajno nanaša na količino bitov, prenešenih v časovni enoti. V okviru vodnega žigosanja zvočnih posnetkov nam določa količino vgrajenih podatkov in se meri v bitih na sekundo (bit/s). Podatkovni tovor se med posameznimi tehnikami in algoritmi vodnega žigosanja močno razlikuje [2].

Računska zahtevnost – je lastnost, s katero merimo obseg izvajanja operacij, ki jih potrebujemo za vgradnjo ali zaznavo vodnega žiga. Zaradi različnih potreb pri uporabi vodnih žigov je zahtevnost med posameznimi postopki lahko zelo različna. Za doseganje boljše računske zahtevnosti moramo skleniti kompromis med robustnostjo in slišnostjo [2].

1.2 Aplikacije vodnega žiga

Dokazilo lastništva, oziroma zaščita avtorskega dela je eno izmed glavnih gonil razvoja tehnologij za vodno žigosanje zvočnih posnetkov. Z vodnim žigom lahko v zvočni posnetek vstavimo podatke, v katerih je naveden lastnik medija. Ta lahko kasneje z detekcijo žiga uveljavlja lastništvo. Vodni žig za ta namen mora biti izredno robusten in imeti visoko stopnjo varnosti, hkrati pa je lahko kratek, zato ne doprinese velikega podatkovnega tovora [2] [1].

Vodni žigi za *spremljanje prenosov* nam omogočajo zbiranje informacij o prenosih vsebine. Tovrstne informacije se nato uporablja za statistične analize in trženje. Robustnost pri tovrstnih vodnih žigih ni tako pomembna, saj je tveganje, da bo prišlo do popačenja, manjše. Veliko pomembneje je, da je vodni žig nezaznaven.

Vodne žige lahko uporabljamo tudi za *overitev vsebine*. Tako lahko preverimo, če je bila kakorkoli spremenjena ali obdelana. Tovrstni vodni žigi so zato zelo krhki in propadejo že ob manjših spremembah vsebine. Ker je vodni žig vgrajen v vsebino, se z njo tudi

spreminja. To nam omogoča, da do neke mere izvemo, kako s kakšnim napadom je bila vsebina napadena [2].

Vodno žigosanje se aplicira tudi za *nadzor kopiranja* vsebin, ki preprečuje izdelavo nezakonitih kopij avtorsko zaščitenih del [2].

2 Orodja in postopki

2.1 Ogrodje sistema in postopek digitalnega žigosanja zvočnega posnetka

V literaturi zasledimo veliko število različnih sistemov za vodno žigosanje zvočnih posnetkov. V grobem je vsak sistem sestavljen iz treh glavnih delov:

- generator vodnega žiga,
- vgrajevalnik vodnega žiga in
- detektor vodnega žiga.

Na začetku postopka vzamemo vsebino, ki jo želimo zavarovati z vodnim žigom. To bo originalen zvočni signal. Z generatorjem vodnega žiga izdelamo vodni žig, ki je lahko različnih oblik, na primer slika, črke ali zaporedje besed. Vgrajevalnik nato vgradi signal vodnega žiga v originalen zvočni signal, tako da dobimo žigosan signal. Takšen signal je pripravljen na prenos, med katerim bo naletel na običajne spremembe med prenosom ali namerne napade, ki želijo uničiti vodni žig. Na koncu detektor izvleče vodni žig iz signala, ki je zaradi poti nekoliko drugačen od žigosanega signala. Detektor nato primerja izvlečeni vodni žig z originalnim vodnim žigom in potrdi, če je vodni žig pristen [2].

2.2 Tehnike in algoritmi

V literaturi zasledimo veliko različnih tehnik, ki koristijo številne algoritme za dodajanje vodnega žiga v zvočne posnetke. Poznamo tehnike, ki vgrajujejo vodne žige v časovni domeni. Ena izmed takšnih metod je tehnika s kodiranjem najmanj pomembnega bita. V postopku vdelave žiga zamenjamo najmanj pomemben bit vsakega vzorca z vodnim žigom, ki ga sestavlja niz binarnih števil [2]. Tehnika sama po sebi ni zelo robustna, zato se zanjo pogosto uporabljajo izboljšave. Prednost tehnike je v tem da je računsko nezahtevna, zato se uporablja pri realno-časovnih aplikacijah [1].

Sledijo tehnike, ki za vgradnjo vodnega žiga ne uporabljajo časovne ampak transformirane domene. Takšne metode navadno vsebujejo dva koraka pred vgradnjo vodnega žiga in sicer domensko transformacijo in inverzno transformacijo. Takšna tehnika je na primer tehnika žigosanja z razširjenim spektrom. Tehnika zaradi njene robustnosti velja za eno najpopularnejših in najvarnejših tehnik [1]. Tehnika razširi vodni žig skozi spekter originalnega signala na način, da je energija, prisotna v vsakem frekvenčnem koraku, zelo majhna in težko zaznavna. Slabost omenjene tehnike je, da proces žigosanja doda slišne popačitve zvočnim posnetkom. Zato se ob tehniki pogosto uporablja tudi amplitudno oblikovanje s pomočjo psiho-akustičnega modela [2] [3].

V nekoliko novejši literaturi zasledimo tudi uspešno uporabo diskretne valčne transformacije kot način za uspešno vgradnjo vodnega žiga v zvočne posnetke [4] [5]. Pri tovrstni tehniki uporabljamo multiresolucijsko analizo in sintezo. Med dekompozicijo, je signal razstavljen na del z visokimi frekvencami in del z nizkimi frekvencami. Del nizkih frekvenc je nato ponovno razstavljen na isti način nizkih in visokih frekvenc. Takšno razstavljanje se ponavlja dokler ne dosežemo željene stopnje. Z vsakim nivojem se časovna resolucija razpolovi, frekvenčna pa podvoji, zato s to tehniko dosežemo zelo dobro frekvenčno resolucijo pri nizkih frekvencah in dobro časovno resolucijo pri visokih frekvencah [2] [4].

3 Preizkus slepega vodnega žigosanja z uporabo diskretne valčne transformacije (DWT)

V literaturi najdemo podrobne opise vodnega žigosanja zvočnih posnetkov z uporabo diskretne valčne transformacije [4] [5]. V projektu želim predstaviti in preizkusiti algoritem za vodno žigosanje zvočnih posnetkov, ki ga najdemo v [5]. Cilj je razložiti algoritem in testirati njegovo delovanje v orodju Matlab s pomočjo že izdelane skripte [6].

3.1 Opis postopka

Omenjeni postopek za vodno žigosanje zvočnih posnetkov uporablja diskretno valčno transformacijo in diskretno kosinusno transformacijo. Za izboljšanje rezultatov se uporabi tudi šifriranje z Arnoldovo transformacijo, ki dobro premeša podatke. V algoritmu se uporabi diskretna valčna transformacija tretje stopnje. Za vodni žig se uporabi dvobarvno sliko, ki je šifrirana z cikličnimi vrednostmi in dodatno premešana z uporabo Arnoldovega algoritma. V naslednjem delu sledijo kratki opisi uporabljenih tehnik.

3.2 Opis uporabljenih tehnik

Diskretna valčna transformacija (DWT) – Kot že prej omenjeno, je valčna transformacija uporabna za analizo tako splošnih, kot bolj specifičnih lastnosti signala. Tovrstna transformacija razstavi osnoven avdio signal v seznam komponent z omejeno velikostjo oken. S ponovnim sestavljanjem komponent lahko rekonstruiramo vhodni signal. Z uporabo DWT preoblikujemo signal z 2-D filtri, ki razdelijo signal v štiri nize multiresolucijskih koeficientov. To so koeficienti približkov nizkih frekvenc ter komponente vertikalnih, horizontalnih in diagonalnih podrobnosti signala. Dekompozicijo lahko izvedemo v več nivojih, s čimer povečamo robustnost in s tem odpornost na resnejše napade [5] [1].

Diskretna kosinusna transformacija (DCT) – je razširjena vrsta transformacije, ki se uporablja tudi v slikovni in video kompresiji. Upošteva vzorce zvočnega

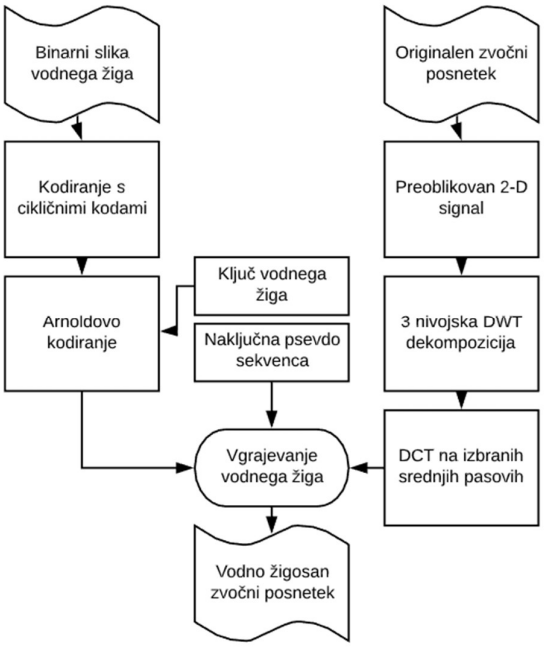
signala kot vsoto kosinusne funkcije, ki oscilirajo pri različnih frekvencah in amplitudah. Transformacija je podobna diskretni Fourierjevi transformaciji ampak uporablja samo realna števila. V primeru kompresije je DCT bolj smiselna kot diskretna sinusna transformacija, ker nam vrne manj koeficientov, s katerimi se učinkovito lahko približamo originalnemu signalu. V omenjenem algoritmu za vodno žigosanje se DCT uporablja za zajem koeficientov pasu srednjih frekvenc, s katerimi lociramo ključno informacijo za vodno žigosanje [5]. Po končni transformaciji lahko ponovno sestavimo signal z inverzno DCT.

Arnoldov kodirni algoritem – se v omenjenem algoritmu uporablja, da doda dodaten nivo zaščite vodnemu žigu. Z algoritmom namreč dodatno kodiramo sliko ki služi kot vodni žig. Uporablja se samo na signalih velikosti $N \times N$, kot definirano v (1),

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} mod N \quad (1)$$

kjer sta (x, y) koordinati originalnega vodnega žiga, (x', y') pa sta koordinati kodiranega žiga. N predstavlja višino in širino signala, ki ga želimo obdelati. Dekodiranje signala je odvisno od kodirnega ključa, ki definira tudi, kolikokrat je bilo kodiranje izvedeno [5].

Vgrajevanje vodnega žiga

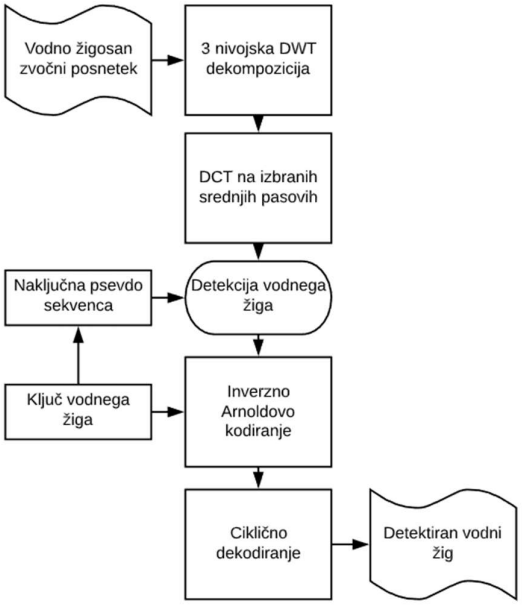


Slika 1 Diagram poteka algoritma za vgradnjo vodnega žiga. Prirejeno po [5].

Postopek vgrajevanja vodnega žiga poteka v naslednjih korakih, kot je prikazano na Slika 1. Zvočni signal je najprej preoblikovan v 2-D matriko. Za vodni žig se izbere binarna slika dimenzij $N \times N$. Z uporabo DWT nato razstavimo 2-D matriko na pod pasove [L,H,V,D]. Izberemo srednja pasova H in V ter ponovimo DWT. Za boljšo robustnost dodamo še drugi in tretji nivo DWT, kjer vedno izberemo srednje pasove, A2, B2, C2 in D2. Vsak pas se razdeli na $\frac{N}{4} \times \frac{N}{4}$ bloke. V vsakem bloku opravimo DCT transformacijo. S ključem vodnega žiga nato ustvarimo 2 naključni psevdosekvenci. Dolžina obeh sekvenc mora biti enaka številu koeficientov srednjih pasov. Binarna slika vodnega žiga je preoblikovana v matriko, ki jo je mogoče zakodirati s periodičnimi kodami. Ciklično kodiranje se ponovi še enkrat za dodaten nivo zaščite. Novo dobljeno matriko dodatno kodiramo s ključem vodnega žiga. Biti kodiranega vodnega žiga so nato vdelani v zvočni signal, nad katerim izvedemo še inverzni DCT in DWT, tako da dobimo vodno žigosan zvočni posnetek [5].

3.3 Detekcija vodnega žiga

Detekcija vodnega žiga poteka v korakih, ki so prikazani na Slika 2 in sicer tako, da vodno žigosan zvočni posnetek ponovno razstavimo s tremi nivoji DWT. Na novo izdelamo dve naključni psevdosekvenci. Tako kot v postopku vgrajevanja izvedemo DCT na izbranih koeficientih srednjih pasov. Vgrajen vodni žig izvlečemo glede na korelacijo med prej izbranimi koeficienti srednjih pasov in psevdosekvenc. Pridobimo novo matriko, ki jo s pomočjo ključa vodnega žiga pošljemo skozi obraten Arnoldov kodirni algoritem. S podobnimi transformacijami kot pri vgradnji žiga in s cikličnim dekodiranjem dokončno izvlečemo binarno sliko vodnega žiga [5].



Slika 2 Diagram poteka algoritma za detekcijo vodnega žiga.

4 Rezultati

Algoritem smo testirali na dveh različnih zvočnih posnetkih. Prvi vsebuje samo človeški govor brez glasbe, drugi posnetek pa je znana rock skladba. Pripravili smo vodni žig, binarno sliko (Slika 3) v velikosti 16x16 slikovnih elementov.



Slika 3 Originalen vodni žig

Preverjali smo napako vodnega žigosanja ob nekaj različnih manipulacijah[1] zvočnega posnetka. To so:

- dodajanje šuma,
- filtriranje s premikajočim povprečenjem,
- spreminjanje vzorčne frekvence,
- obrezovanje.

Prav tako smo s primerjavo originalnega in žigosanega posnetka preverili, kakšen vpliv ima vodni žig na zvočni posnetek. Rezultati so zapisani v Tabela 1. V tabeli so prikazani tudi detektirani vodni žigi.

Tabela 1 Rezultati testiranja

	SSIM	BER	Detektiran vodni žig
Brez vplivov			
Govor	1	0	
Skladba	0.9995	6	
Dodajanje šuma			
Govor	1	0	
Skladba	0.9985	18	
Filtriranje			
Govor	0.9943	64	
Skladba	0.9912	78	
Prevzorčevanje			
Govor	0.9899	80	
Skladba	0.9856	102	
Obrezovanje			
Govor	0.9964	31	
Skladba	0.9964	37	

Bit error rate (BER) – nam prikaže napako med začetnimi in končnimi biti. V primeru algoritma je to število napačno zapisanih slikovnih elementov v sliki zaznanega vodnega žiga.

Structural similarity index (SSIM) – je mera, ki se uporablja za merjenje podobnosti med dvema digitalnima vsebinama. V danem algoritmu na tak način testiramo vpliv vodnega žiga na zaznavne spremembe zvočnega posnetka. Opazuje se originalen zvočni posnetek, ki se primerja z končnim posnetkom, ki vsebuje vodni žig.

5 Zaključek

Skozi članek smo predelali področje vodnega žigosanja digitalnih vsebin, specifično, avdio posnetkov. Posamezne tehnike, ki so temu namenjene, se po uspešnosti razlikujejo od namena uporabe. Velja, da moramo skleniti kompromis med robustnostjo, podatkovnim tovorom in zaznavnostjo. Pod drobnogled smo vzeli algoritem za vodno žigosanje, ki temelji na DWT in DCT ter Arnoldovem kodirnem algoritmu. Postopek ponuja zanimivo rešitev, kako v zvočni posnetek zakodiramo binarno sliko, ki služi kot vodni žig. Za testiranje smo izdelali vodni žig v obliki takšne slike. S pomočjo že pripravljenega algoritma smo preizkusili žigosanje na dveh vrstah zvočnih posnetkov in sicer na posnetku govora in rock skladbi. Preverili smo tudi odpornost vodnega žiga s pomočjo že pripravljenih manipulacij signala. Iz rezultatov je razvidno, da se vodni žig bolje obdrži na zvočnem posnetku govora. Razvidno je tudi, da je vodni žig uporabljenega algoritma zelo nezaznaven, kar pomeni, da se žigosan signal zelo malo razlikuje od originalnega.

Literatura

[1] D. Isovski, „Odpornost zvočnega vodnega žiga na napade,“ 2017.

[2] Y. Lin, W. H. Abdulla in others, Audio watermark, Izv. 146, Springer, 2015.

[3] Y. Xiang, G. Hua in B. Yan, Digital audio watermarking: fundamentals, techniques and challenges, Springer, 2017.

[4] A. M. Joshi, „Hardware Implementation of Audio Watermarking Based on DWT Transform,“ v Ethics, Laws, and Policies for Privacy, Security, and Liability, IntechOpen, 2019.

[5] A. M. Joshi in others, „DWT-DCT based blind audio watermarking using Arnold scrambling and Cyclic codes,“ v 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), 2016.

[6] Subir, „GitHub,“ 99sbr, 2016. [Elektronski]. Available: https://github.com/99sbr/Audio-Watermarking. [Poskus dostopa 17 4 2020].

[1] Manipulacije se bile že vdelane v algoritem [6] in sicer po postopku opisanem v [5]

Suspended False Ceiling acoustic elements theoretical and experimental investigation

[1] Giovanni Amadasi, [2, 3] Franc Policardi-Antoncich
[1] *Vibro-Acoustic SCS & Partners, Via Antoniana 278, 35011 Campodarsego PD, Italy*
[2] *LUCAMI, Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1000 Ljubljana, Slovenija*
[3] *Raziskovalni Inštitut za nove tehnologije in energetiko (R.I.N.T.E.) d.o.o., Stritarjeva 6/a, 4000 Kranj, Slovenia*
g.m.amadasi@scs-controlsys.com

Abstract

This research work investigates the all-together sound package alphacell® software used for Suspended False Ceiling acoustic elements (SFCae), in order to understand the influence of Panel Perforation (PP) and the porous characteristics on Sound Absorption (SA) values. The present study investigates the poro-acoustic materials physical models of each layer and the theoretically achieved results for the multi-layers, by varying and optimizing the physical model parameters of Porosity and Flow-Resistivity. More than 20 standard perforation patterns were also experimentally investigated, confirming the theoretically obtained results. The present paper deals with 14 standard perforation patterns.

1 Introduction

Suspended False Ceiling acoustic elements (SFCae) are widely used in closed spaces to correct room acoustics. They present different layers and surface finishings; their acoustic performance is normally expressed with only an overall absoption coefficient α_W, which does not always represent the real acoustic characteristics of the ceiling. The hypothesis of the present study is that most of the published data are just consequence of specific installation results mixed with ISO 354:2003 and/or ISO 10848-2:2006 standard test procedures. The mentioned standards' results are affected by test methods boundary conditions and these results can not be used "as they are" for room acoustic design and simulation.
Many existing standard products and solutions on the market were investigated and tested using: A) a SCS-9020B© Impedance tube compliant with ISO 10534-2 for Normal incidence (plane waves) α_N coefficient, B) a SCS-9031© small reverberation chamber tuned to match ISO 354 for diffuse field giving the α_{Stat} coefficient and C) a Flow-Resistivity testing device SCS-9023© complying with ISO 29053-2.
The experimental results have been compared to the physical models simulated by the *alphacell®* software (TIMM method), in order to reach a suitable agreement between experimental results and physical model choices. Analysis and optimization processes were conducted varying the characteristic parameters of every single layer using the *alphacell®* [1] software, investigating the presented results.

2 Parameters Definition

Most of the results were theoretically achieved for the multi-layers combination by varying and optimizing the poro-acoustic materials physical model parameters Porosity and Flow-Resistivity on the one side, and by changing the holes geometrical characteristics of the Perforated Panels "PP" on the other side.

2.1 Open Porosity

Open Porosity (ϕ in %) is the ratio between the interconnected void volume and the total volume of a material. The theory states that it is enough to measure the volume vs. mass ratio of an absorber sample, but in real conditions this parameter is difficult to measure because: a) in case of open cell foams it is hard to determine which cells are really open and interconnected, b) in case of fibrous absorbers it is difficult to measure the exact volume of a compressible sample. Just as a general example, poro-acoustic materials achieve up to 95-98% of ϕ for fibrous absorbers. One simple method to measure ϕ consists of filling the sample pores with a liquid and measure its volume, but the liquid can contaminate the sample and preclude other parameter measurements. Another method uses thermodynamics: the sample is placed in a chamber which is compressed, leading to a volume differential ΔV, so the internal pressure will increase by ΔP according to Boyles law:

$$V_0 + V_t = -\frac{P_0 + \Delta P}{\Delta P}\Delta V \tag{1}$$

2.2 Static air flow resistivity

Static air flow resistivity (σ) (N.s.m−4), also called Resistivity, it is defined as: $\sigma\phi\vec{v} = -\vec{\nabla} p$, an expression representing the generalized Darcy's law in which ϕ is the open porosity of the material, $\vec{v}$ the velocity of the fluid particles subjected to the pressure gradient $\vec{\nabla}p$, and $\phi\vec{v}$ is the fluid flow inside the porous material.
The definition of σ is based on the assumption that $\vec{v}$ and $\vec{\nabla} p$ are constants (Static Resistivity). Some authors prefer to use the static viscous permeability k_0 instead of the static air-flow resistivity. The static permeability, expressed in m^2, is defined as $k_0=\eta\sigma$, where η is the dynamic viscosity of air (~1.84 × 10−5 N.s.m−2 at normal temperature and pressure conditions). It should be pointed out that k_0 does not depend on the fluid property, whereas σ does.

σ represents the visco-inertial effects at low frequencies (viscous boundary layer of the same order of magnitude of the characteristic pore size). The Delany-Bazley and Delany-Bazley-Miki physical models [2] use only this parameter to describe the behavior of fibrous acoustical materials. σ values in poro-acoustical materials spans between 10^3 and 10^6 N.s.m−4.

2.3 Transparency Index

Transparency Index (TI) considers a perforated plate as a porous material using macroscopic parameters. The two extremes 10% open area surfaces are either making a single large hole in the center or making very fine perforations overall the surface. In the first case, instead of a transparent facing material, there is a small completely open area at the surface center and the rest is completely opaque to sound reflecting all of it. In the second case, the entire surface is almost completely transparent to sound, because the tiny solid areas between the holes are too small to interact with the sound waves. TI index was introduced by Schultz [3] and is defined by the following formula:

$$TI = nd^2/ta^2 = 0.04\ P/\pi ta^2 \quad (2)$$

Where:
n = number of perforations per square inch (in^{-2});
d = perforation diameter (in);
t = sheet thickness (in);
a = shortest distance between holes (in);
a = b – d, where
b = on-center hole spacing (in);
P = percent (not fractional) surface open area.
Acording to the Delany-Bazley-Miki physical model [2], the open porosity ϕ determines the sound attenuation i.e. the perforated plate screen effect, becoming more important at high frequencies. Figure 1 shows the attenuation at 10kHz as a function of TI.

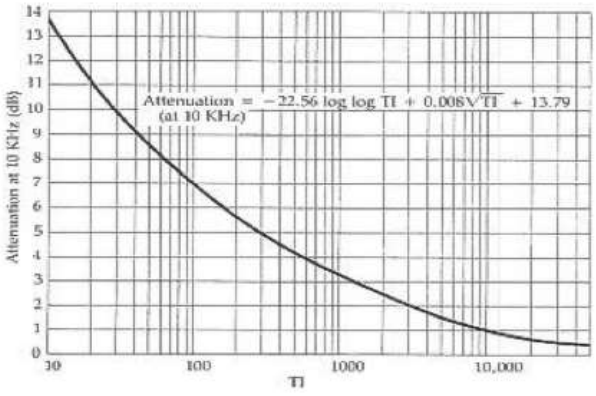


Figure 1. TI index: Schultz graph attenuation at 10kHz

Table 1 shows a comparison between two different PP with an 8x TI index difference (PP1 11.562, PP2 1.439), in which the sound attenuation at 10kHz drops from 2.83 down to 0.91 dB.

Table 1. Strong TI index difference in two perforated plates

Parameter	sample 1	sample 2
	mm	mm
n = number of perforations per sq in; n = 1/A	396,9100	73,6700
d = perforation diameter	0,5800	1,6700
t = sheet thickness	0,4700	1,5900
a = shortest distance between holes; a = b – d	0,0311	0,0594
b = on-center hole spacing	1,3700	3,1800
A = sheet area per hole = $b^2\cos 30^0$	0,0025	0,0136
angle degree °	30,0000	30,0000
P % open area - Porosity = $[(pd^2/4)/A] \times 100$	16,2400	24,9900
TI = nd2/ta2	11.562,00	1.439,000
Attenuation at 10 kHz = -22.56 log log TI + 0.008vTI + 13.79dB	0,9100	2,8300

3 Simulation models validation

The *alphacell*® software can compute the acoustic behavior of very different systems, as for example those composed by air-gaps, isotropic & orthotropic solids and isotropic porous media. It uses Delany-Bazley-Miki, Hamet, Attenborough, and Johnson-Champoux-Allard-Pride-Lafarge models. The software takes into account the material's Pore Size Distribution (PSD), Micro-Perforated Panels (MPP) with circular, rectangular or slit-like perforations also, and it is suitable for Woven and Non-Woven textiles too.
alphacell® is capable to apply the model developped by Biot in 1956 [4], [5] to all previous acoustic models, in order to include the elastic effects of the porous frame as: visco-elastic isotropic, limp and rigid body behaviour, heavy-layer like, impervious screen and the double porosity by Olny & Boutin [6]. The software can compute multi-layered structure acoustic responses with arbitrary incidence plane waves, modal sound field, diffuse sound field, mechanical force, turbulent boundary layers and even rain on the roof and many more.
Experimental data from different PP (PP1 to PP7) were extracted using the Kundt impedance tube as in A) and the Flow-Resistance device as in C). The validation procedure compared all the experimental data as in B) for every PP sample with software simulation results. The layers sequence model used in the the Kundt tube consists of a PP, a TNT thin tissue and an air gap. This last one was generally chosen 200 mm as it is the distance between the inner face of the TNT tissue and the Kundt tube end. The TNT tissue thickness was kept constant between 0.2/0.25mm and the variables were 1) the PP perforation characteristics and 2) the PP thickness.

Different PP were chosen from the normal industrial production in order to represent a broadband spectrum of the available SFCae.

3.1 Normal incidence sound wave: real data vs. simulation data

Sound waves expansion in a free or semi-free field is practically spherical. Interactions between sound waves in free air and a ceiling are strongly influenced by their incidence planes. These are are theoretically infinite, but in order to achieve a suitable comprehension, sound wave incidence planes are studied for normal (90°) incidence and for approximized random incidence planes named Diffuse Field.

Table 2 shows the holes geometric data for all the PP samples; circular perforation is the common one, representing 6 of the 7 samples, while only 1 sample is made with square holes. The number of perforations spans from a minimum of 14,26 up to 101,38 holes /sq. inch. Hole dimensions are also very different, spanning from 0,75 up to 4 mm, with a variability more than 5 x, so that TI values span between a minimum of 34 up to 824.

Parameter	Comments		PP 1	PP 2	PP 3	PP 4	PP 5	PP 6	PP 7
Perforation	C=circular, Sq=squared		C	C	Sq	C	C	C	C
n = number of perforations per sq in; n = 1/A			25.34	36.50	14.26	25.81	101.38	36.50	74.48
d = perforation diameter		mm	2.50	2.00	4.00	0.75	1.00	2.00	2.00
t = sheet thickness		mm	0.70	0.60	0.60	0.60	0.50	0.50	5.00
a = shortest distance between holes: a = b – d		mm	0.1378	0.1181	0.1575	0.1673	0.0787	0.1181	0.0591
b = on-center hole spacing		mm	6.00	5.00	8.00	5.00	3.00	5.00	3.50
A = sheet area par hole = $b^2 \cos 30^0$		mm2	0.0395	0.0274	0.0701	0.0388	0.0099	0.0274	0.0134
angle deg		deg	45.00	45.0	45.0	-	45.0	45.0	45.0
P % open area - Porosity= $[(pd^2/4)/A] \times 100$		%	19.2736	17.7625	27.7539	1.7663	12.3351	17.7625	36.2500
TI = nd2/ta2			469.20	686.66	603.51	34.02	1 287.49	823.99	672.65
Attenuation at 10 kHz = -22.56 log log TI + 0.008vTI + 13.79		dB	4.34	3.78	3.97	9.66	2.96	3.53	3.81

Table 2. 7 PP samples hole geometry data

Table 3 shows Flow Resistivity "r" data for all PP samples; sample thickness is almost the same, the measured r values range from 4.962 up to 8.329 and the RS Specific r ranges from 38,95 up to 65,38. r ranges from a minimum of 7.790 up to a huge 128.395, with a 16,5x difference factor.

Parameter	Comments		PP 1	PP 2	PP 3	PP 4	PP 5	PP 6	PP 7
d sample	thikness	mm	0.70	0.60	0.60	0.60	0.50	0.50	0.50
R (measured)	Dp/Qv	Pa s/m^3	6 022	6579	8 190	8 329	8 178	6 414	4 962
Dp	diff. Pa (mic.)	Pa	23.64	25.82	32.15	32.69	32.10	25.17	19.48
Qv	Volume flow	m^3/s	3.93E-03						
Rs Specific R	R A = Dp/v	Pa s/m	47.27	51.65	64.29	65.38	64.20	50.35	38.95
v	air speed	m/s	0.5						
A	Sample Area	m^2	0.00785						
r Flow Resistivity	Rs/d = Dp/(v d) (Ns/m4)	Pa s /m^2	67 532	86 075	107153	108971	128395	100700	7790

Table 3. 7 PP samples Flow Resistivity "r" data

Table 4 shows r data for all PP samples but with an added TNT layer; sample thickness remains the same. The measured r values range from 38.140 up to 63.070, and the RS Specific r ranges from 299,40 up to 495,1 both with an enormous increase from Table 3. The difference with/without TNT spans from 65.089 up to 737.994 i.e. more than 11 times.

Parameter	Comments		PP1 + TNT	PP2+ TNT	PP3 + TNT	PP4 + TNT	PP5 + TNT	PP6 + TNT	PP7 + TNT
d sample	thikness	mm	0.70	0.60	0.60	0.60	0.50	0.50	0.50
R (measured)	Dp/Qv	Pa s/m^3	38 140	49190	63070	44 520	52 020	53 420	46 420
Dp	diff. Pa (mic.)	Pa	149.70	193.1	247.6	174.7	204.2	209.7	182.2
Qv	Volume flow	m^3/s	3.93E-03						
Rs Specific R	R A = Dp/v	Pa s/m	299.40	386.1	495.1	349.5	408.4	419.4	364.4
v	air speed	m/s	0.5						
r F. Resistivity	Rs/d = Dp/(v d) (Ns/m4)	Pa s /m^2	427713	643569	825166	582470	816714	838694	72879
Difference w/out TNT			360180	557494	718013	473499	688319	737994	65 089

Table 4. 7 PP samples + TNT Flow Resistivity "r" data

After the PP complete objective analysis and measurement, a model validation was performed, based on Impedance tube results for α_N coefficient and *alphacell*® software on PP1 and on PP4 for the normal incidence plane. The comparison was performed in 1/3 octave bands from 50 Hz up to 1.5 kHz and the showed tolerances are quite acceptable for acoustic modeling.

Figure 2 verifies PP1 + TNT ϕ = 19% with σ 360 kPa s/m^2 and the compliance is very good or at least good in all frequency bands, with slight differences in the upper frequency bands.

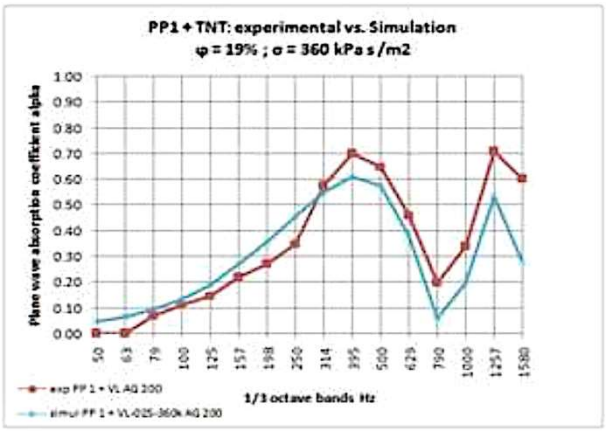


Figure 2. PP1 + TNT experimental vs. simulation

Figure 3 verifies PP4 + TNT ϕ = 1,8% with σ 580 kPa s/m^2 and the compliance is very good or at least good in all frequency bands, with slight differences in the lowest frequency band.

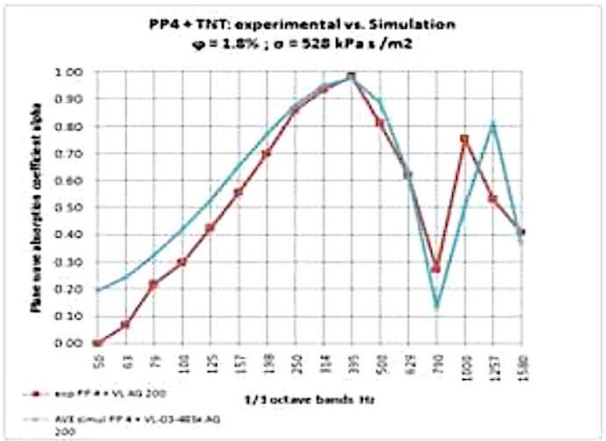


Figure 3. PP4 + TNT experimental vs. simulation

3.2 Diffuse field incidence sound wave: real data vs. simulation

As already mentioned, a complete sound wave incidence study must take into account the almost infinite random non α_N normal incidence planes named α_{Stat} Diffuse Field. In this situation sound waves bounce the surface from multiple statistically random non defined planes, interacting with the surface and in this case with its holes in a different way.
Figure 4 shows how the random α_{Stat} Diffuse Field incidence planes differs from the α_N normal incidence plane, with a significant sound absorption drop in a PPx sample (any sample with ϕ >12%) + TNT and PP4 + TNT.

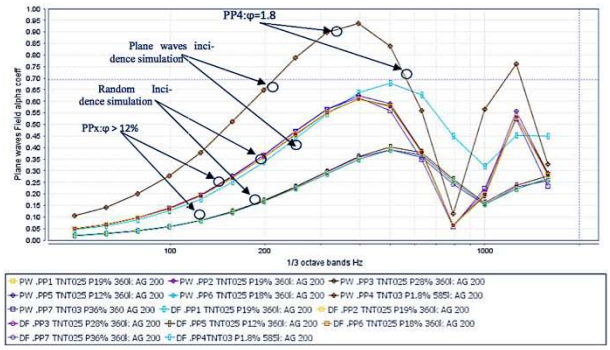


Figure 4. α_N normal incidence plane and α_{Stat} Diffuse Field simulation for a PPx sample (any sample with ϕ >12%) + TNT and PP4 + TNT of 1,8 %

The simulation was then extended up to 5 kHz (Figure 5) for a PPx sample (any sample with ϕ >12%) and PP4 (ϕ = 1.8%), both for Plane waves "PW" and Random incidence α_{Stat} Diffuse Field conditions.

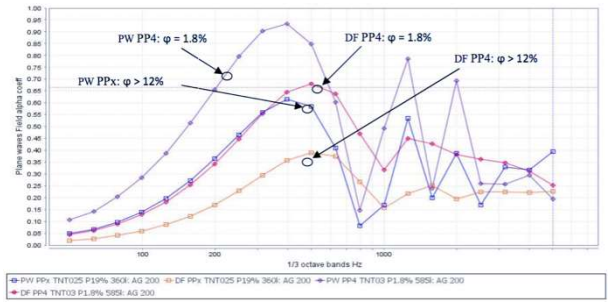


Figure 5. α_N normal incidence plane and α_{Stat} Diffuse Field simulation for a PPx sample (any sample with ϕ >12%) + TNT and PP4 + TNT with ϕ 1,8 % extendet do 5 kHz

The simulation substantially confirmed the PP + TNT behaviour.

4 Conclusion

After the experimental validation of the model simulation, it has been shown how it is possible to study multi-layers in order to correctly design a Suspended False Ceiling acoustic element SFCae through a specific software.
The simulation studied the holes diameter optimisation by applying *alphacell*® features to PP at first instance by considering geometry parameters (D and φ) and estimating the α_{Stat} up to 5 kHz.
The simulation took into account the Plate tickness, the holes diameters at constant porosity, the Flow Resistivity and the Open Porosity in order to maximize the sound absorption whith a the specific air flow resistance adding TNT tissue obtaining a significant sound absorption as well as a full spectrum good balance absorption.
This paper also demonstrate that to some extent the single-number rating α_W EN ISO 11654, based on the limited frequency interval 250 - 4.000 Hz, does not provide sufficient information to acousticians and it does not take into account the balance of α_{Stat} Diffuse Field in the full spectrum range.

Aknowledgements

The authors wish to acknowledge the precious help of Atena Spa for the PP samples and the TNT tissues, Matelys research raboratory for the scientific support and bright ideas in using *alphacell*® software, SCS-Euroacoustic laboratory for experimental test on porous materials, R.I.N.T.E. and LUCAMI for DOE and present paper writing.

References

[1] alphacell.matelys.com, 10. 7. 2020

[2] Delany M. E. and Bazley E. N., Acoustical properties of fibrous absorbent materials, Applied Acoustics 3, 1970, pp. 105-116, DOI: 10.1016/0003-682X(70)90031-9

[3] Theodore J. Schultz, Acoustical uses for perforated metals, Ph.D. dissertation, Houston - Texas, 1986

[4] M. A. Biot, *Theory of propagation of elastic waves in a fluid saturated porous solid. I Low frequency range*, J. Acoust. Soc. Am. 28, pp. 168–178, 1956

[5] M. A. Biot, *Theory of propagation of elastic waves in a fluid saturated porous solid. II Higher frequency range*, J. Acoust. Soc. Am. 28, pp. 179–191, 1956

[6] X. Olny, C. Boutin, Acoustic wave propagation in double porosity media, The Journal of the Acoustical Society of America 114, 73 (2003)

Računalništvo in informatika

Computer and Information Science

Ekstrakcija ključnih besed filmov iz podnapisov

Jan Popič, Timi Ornik, Nejc Planer, Borko Bošković, Janez Brest

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,
Koroška cesta 46, 2000 Maribor
E-pošta: jan.popic1@um.si

Keyword extraction from movie subtitles

With large amounts of video content it is important to have as much metadata as possible. Important metadata in video content includes keywords which summarize the essential elements in a given video. In this paper, we present a framework for an automated approach to keyword extraction from video content using their associated subtitles. Framework is composed of three primary building blocks: preprocessing, named-entity recognition and keyword extraction. The goal of the first two blocks is the removal of information irrelevant to the keywords. In the final stage, the actual keywords are extracted from the preprocessed subtitles. In the experiment, we compare the keywords extracted with our framework to a set of predefined keywords obtained from the Internet Movie Database website. We test our method on Slovene language and compare it to a similar English model.

1 Uvod

V današnjih dneh je na voljo ogromna količina video vsebin, le-te pa postajajo del našega vsakdana. Ker pa je v tej poplavi vsebin pogosto težko najti takšno, ki bi nas zanimala, je posebej pomembno, da so dobro označene z metapodatki. Eden izmed ključnih metapodatkov o video vsebini so ključne besede. Z avtomatsko ekstrakcijo ključnih besed lahko neizmerno pohitrimo sicer ročno označevanje. Ključne besede so uporabne predvsem za priporočilne sisteme, pripise in oznake vsebin in povezovanje s socialnimi omrežji.

Ekstrakcija ključnih besed je definirana kot postopek, ki vhodnemu dokumentu dodeli nekaj besed ali besednih zvez, s katerimi najbolje opišemo njegovo vsebino [1]. Ker ima večina video vsebin v današnjih dneh vsaj neko obliko pripadajočih podnapisov, smo se v našem delu osredotočili samo na ekstrakcijo ključnih besed iz slovenskih podnapisov. Zaradi najlažjega dostopa do podnapisov in obstoječih metapodatkov o vsebini, se bomo omejili samo na filme. Enak pristop bi lahko preslikali na vsako video vsebino, ki ima podnapise.

2 Sorodna dela

V [2] so se avtorji ukvarjali z ekstrakcijo ključnih besed iz prepisov (transkripcije) sestankov. Avtorji so izpostavili nekaj pomembnih razlik med prepisi govora in navadnim pisanim besedilom (knjige, članki, prispevki, itd.), ki veljajo tudi za našo domeno analize podnapisov video vsebin:

1. V nenačrtovanem pogovoru sta v stavku povprečno dve pomenski besedi, za razliko od pisanega besedila, kjer je v posameznem stavku od 4 do 6 pomenskih besed. To lahko predstavlja problem osnovnim algoritmom ekstrakcije ključnih besed, ki temeljijo na frekvenci pojavitev.
2. V govorjenem besedilu je pogovor manj strukturiran. Manjkajo tudi dodatne informacije, ki jih lahko zasledimo v drugih pisanih besedilih (naslovi, odstavki, poglavja, itd.).
3. Za razliko od pisanih besedil, kjer je navadno en sam avtor, imajo v prepisih različni govorci različne sloge govorjenja in različno rabo besed.
4. Ker je struktura govora ohlapna in ni zagotovila, da je pravilna, se lahko pojavijo težave z označevanjem besednih vrst (angl. *POS tagging* oz. *part-of-speech tagging*).

Avtorji prispevka [3] so po našem znanju edini, ki so uporabili dodatne informacije na voljo v podnapisih. Njihov pristop je deljen na 5 osnovnih gradnikov: predobdelava, prepoznava imenskih entitet, ločevanje pogovorov, ekstrakcija ključnih besed in prilagoditev uteži ključnih besed. Poročali so o 11,2 % izboljšavi metrike *F1* z algoritmom TextRank [4] in uporabo prej navedenih korakov v primerjavi z uporabo algoritma TextRank nad celotno vsebino podnapisov.

3 Zasnova algoritma

Naš primarni cilj je bil izdelava sistema za ekstrakcijo ključnih besed filmov iz podnapisov, ki so v slovenskem jeziku. Za enostavnejše nadaljnje delo in preprosto ter intuitivno uporabo smo želeli izdelati ogrodje, ki omogoča enostavno konfiguracijo komponent (odstranjevanje in spremembo posameznih delov ogrodja).

Pri snovanju našega pristopa smo se zgledovali po podobnem pristopu za angleški jezik, ki so ga predstavili avtorji v [3]. Vendar naš pristop ne vsebuje ločevanja pogovorov, prav tako se v koraku prepoznave imenskih entitet nismo omejili samo na unikatne besede, ki se nikoli ne pojavijo z malo začetnico, temveč smo dovolili malo

število ponovitev zaradi morebitnih slovničnih in besednih napak v samih podnapisih. Naš pristop je razdeljen na tri osnovne gradnike:

a) Predobdelava

V tej fazi smo iz podnapisov izločili nepotrebne informacije: vse značke HTML (barvne, glasbene, način govora, označbe dogajanja), znake, ki nakazujejo spremembo govorca ('-') in dvojne presledke. Prav tako smo iz vsake datoteke s podnapisa odstranili zadnja dva podnapisa, ker le-ta ponavadi vsebujeta podatke o avtorju prevoda, kar ni pomembno za ekstrakcijo ključnih besed. Nato smo odstranili nepomembne in nepotrebne besede s pomočjo v naprej definiranega NLTK (angl. *Natural Language Toolkit*) seznama nepomembnih besed [5].

b) Prepoznava imenskih entitet (PIE)

V tej fazi smo filtrirali besedne vrste ter ločili imenske entitete (predmete iz resničnega sveta, na primer osebe ali lokacije, ki jih lahko označimo z lastnimi imeni), ki so primerne za ključne besede od tistih, ki to niso.

Najprej smo nad vsebino posameznega podnapisa zagnali algoritem za označevanje besednih vrst, da smo pridobili informacije o strukturi stavka. Iz posameznega stavka se nato izbrišejo vse besede, ki ne predstavljajo želenih besednih vrst (želene besedne vrste so uporabniško nastavljiv parameter).

Izmed besed, ki so ostale v stavku, se kot kandidatke označijo vse besede, ki predstavljajo lastna imena in so se z malo začetnico ponovile največ dvakrat (na ta način omilimo morebitne sintaktične napake v podnapisih).

Za vse kandidatke smo s pomočjo semantičnega leksikona slovenščine sloWNet [6] preverili, ali je njihova leksikalna domena na seznamu dovoljenih (dovoljene domene so uporabniško nastavljiv parameter). Vse, ki niso ustrezale seznamu dovoljenih leksikalnih domen, smo odstranili iz stavkov.

c) Ekstrakcija ključnih besed

Za samo ekstrakcijo ključnih besed smo uporabili dva različna algoritma (TD-IDF [7] in TextRank), ki smo ju nato med seboj primerjali po uspešnosti.

3.1 Zbiranje podatkov

Korak prepoznave imenskih entitet potrebuje nekaj modelov in korpusov (nabor besedil) za pravilno deljenje besed (angl. *tokenizer*), označevanje besednih vrst in leksikalno analizo. Uporabili smo sledeče modele in korpuse:

- seznam nepomembnih slovenskih besed iz NLTK,
- jezikovni model Punkt za deljenje povedi iz NLTK,
- korpus "MULTEXT-East 1984" za označevanje besednih vrst iz NLTK,
- semantični leksikon Open Multilingual Wordnet za dodatne informacije in povezavo s sloWNet iz NLTK,
- model Word2vec "Slovenian CoNLL17 corpus" iz NLPL [8] za mero podobnosti in
- semantični leksikon slovenščine za pridobitev domen in sinonimov sloWNet.

Za uspešno analizo rezultatov potrebujemo dovolj veliko testno množico podnapisov in ključnih besed, ki predstavljajo nabor ključnih besed, ki so jih določili uporabniki in ocenjevali za najbolj primerne. Nabora ključnih besed za filme nismo našli v slovenskem jeziku, zato smo ga pridobili v angleškem jeziku iz spletne strani IMDb [9] (urejene po ocenah uporabnikov). Ključne besede smo nato prevedli v slovenski jezik s pomočjo prevajalnika Microsoft Translator [10]. Zaradi prevajanja smo morali vse ključne besede spremeniti v male črke. Podnapise smo pridobili iz portala Podnapisi.NET [11], kjer pa ni nujno, da obstaja podnapis za vsak izbran film, ta pa je lahko tudi napačen.

Zaradi samega prevajanja in načina pridobivanja podnapisov smo posledično uvedli nekaj napak v naš korpus. Dodatne napake se lahko pojavijo tudi v samih podnapisih, navadno tistih, ki ne izhajajo iz uradnih studiov. Izbrali smo 100 najbolje ocenjenih filmov iz IMDb in pridobili en podnapis ter 60 prevedenih ključnih besed za vsak film. Primer nekaj ključnih besed je viden v tabeli 1.

Tabela 1: Primeri ključnih besed iz IMDb

Film	Prevedene ključne besede
The Matrix (1999)	simulirano resničnost, umetna realnost, prerokba, programer, po apokalipsi, borilnih veščin, hacker, distopija, človeštvo v nevarnosti, tabletke
Fight Club (1999)	presenečenje konča, boj, proti obratu, multiple osebnostne motnje, nespečnost, proti skladnosti, na podlagi novih, skupinsko zdravljenje, proti kapitalizmu, pretep

4 Eksperiment

Naš eksperiment smo zasnovali tako, da smo za vsak film iz podnapisov izluščili ključne besede z našim algoritmom, te pa smo nato primerjali s tistimi pridobljenimi iz IMDb. Za pravilno pridobljeno ključno besedo smo upoštevali dobesedno ujemanje, sinonime pridobljene iz sloWNet in mero podobnosti. Pri tem nismo preverjali ustreznih velikih začetnic, saj ima naša zbirka pridobljenih ključnih besed izključno male črke. Za ovrednotenje smo naš pristop smatrali kot klasifikator v en razred. Iz matrike klasifikacije (angl. *confusion matrix*) smo nato izračunali metriko $F1$, natančnost P (angl. *precision*) in priklic R (angl. *recall*).

Analizirali smo kombinacije algoritmov za ekstrakcijo besed (TF-IDF in TextRank) z in brez modula prepoznave imenskih entitet PIE (glej poglavje 3), saj nas je zanimal vpliv tega modula na kvaliteto ključnih besed.

V modulu PIE smo v tem primeru obdržali samo lastna imena, samostalnike, pridevnike in števila. Iz lastnih imen smo v nadaljevanju izluščili samo tiste, ki imajo veliko začetnico in se ne pojavijo več kot dvakrat z malo začetnico. Izluščene besede smo preverili s sloWNet in odstranili vse, ki ne spadajo v izbrano domeno (obdržane domene *geography, chemistry, telecommunication, astronomy, politics, animals* so bile izbrane s subjektivno oceno primernosti ključnih besed). Pri odstranjevanju neželenih

besednih vrst se v povprečju odstrani 69 % unikatnih besed, pri odstranjevanju neželenih leksikalnih domen - z našimi obdržanimi domenami - pa 1,5 %. Leksikalne domene imajo vpliv predvsem na tip pridobljenih ključnih besed (subjektivna definicija ključne besede), ne toliko na kvaliteto. Skupno se v povprečju odstrani 71 % vseh besed (vključno s ponovitvami).

Tabela 2: Primerjava različnih kombinacij algoritmov v %

Metoda	*F1*	*P*	*R*
TF-IDF	1,468	1,468	1,469
TF-IDF + PIE	**2,320**	**2,319**	**2,321**
TF-IDF + PIE + S	8,061	8,061	8,061
TextRank	1,456	1,461	1,452
TextRank + PIE	1,979	1,988	1,969
TextRank + PIE + S	**8,770**	**8,760**	**8,780**

PIE - Prepoznava imenskih entitet, S - Sinonimi

Vse kombinacije algoritmov in njihovi rezultati so vidni v tabeli 2.
Največjo natančnost brez upoštevanja sinonimov je dosegel algoritem TF-IDF. Brez modula PIE ima algoritem TF-IDF le majhno prednost pred algoritmom TextRank, če vključimo modul PIE se oba rezultata izboljšata, razlika med rezultatoma obeh algoritmov pa postane bolj očitna. Z upoštevanjem sinonimov pa ima TextRank še večjo natančnost.

Nadalje nas je zanimal vpliv nabora besednih vrst, ki se v modulu PIE izbrišejo. Za analizo tega smo izbrali algoritem TextRank, saj se je izkazalo, da dosega najboljše rezultate za naš problem, filtriranje domen pa smo izključili. Kot je razvidno iz tabele 3, imajo največji vpliv na rezultat glagoli, saj se ti v podnapisih velikokrat pojavijo, so pa le redko ključne besede.

Tabela 3: Vpliv nabora besednih vrst v %

Obdržani členi	*F1*	*P*	*R*
Vsi + S	5,435	5,429	5,441
NN, A, NUM, V + S	7,066	6,959	7,176
NN, A, NUM + S	8,660	8,647	8,674
NN, A + S	**8,770**	**8,760**	**8,780**

NN - Samostalniki, A - Pridevniki, NUM - Števila, V - Glagoli, S - Sinonimi

Na koncu nas je zanimala še primerjava z upoštevanjem podobnosti izluščenih ključnih besed s pridobljenimi na najboljših kombinacijah iz prejšnjih eksperimentov (algoritem TextRank, modul PIE s samostalniki in pridevniki ter upoštevanje sinonimov).

Podobnosti besed v vmesnih eksperimentih nismo upoštevali s ciljem večje natančnosti izbire in v izogib popačenja vmesnih rezultatov. Prav tako, bi v eksperimentu, kjer smo primerjali besedne vrste, vplival na izbiro besednih vrst, v primeru glagola bi bili rezultati nepravilno višji, saj si je veliko glagolov in samostalnikov zelo podobnih, vendar se glagoli običajno ne uporabljajo kot ključne besede.

Podobnost besed predstavimo s pragom podobnosti pridobljenim s pomočjo modela Word2vec. S tem želimo zgladiti napake pri prevajanju in poskušamo upoštevati subjektivnost in abstraktnost ključnih besed. Ta korak je uporabljen izključno za grobo analizo rezultatov in ne vpliva na delovanje algoritma. Prag z vrednostjo 0,7 v tabeli 4 predstavlja vrednost mere podobnosti, nad katero se ključna beseda šteje kot pravilni zadetek. Prag 0,7 je bil določen s preliminarni poizkusi in predstavlja smiselno vrednost pri kateri se pomensko podobne besede označijo kot sprejete, pomensko različne besede pa se še vedno zavrnejo. Nižji prag bi sicer prinesel boljše kvantitativne rezultate, a same ključne besede ne bi bile smiselne.

Tabela 4: Rezultati najboljše kombinacije v %

Podobnost	*F1*	*P*	*R*
1,0	8,770	8,760	8,780
0,7	**13,869**	**13,870**	**13,869**

4.1 Opisna analiza

Tabela 5: Primeri izluščenih ključnih besed

Film	Izluščene besede
The Matrix (1999)	vrata, **prerokba**, morfej, konec, časa, poveljnik, **programer**, **program**, **človek**, matrico, sion
Fight Club (1999)	gospod, tyler, življenje, stanovanje, **rak**, tylerja, daj, singer, ime, **pretep**

Primere izluščenih ključnih besed vidimo v tabeli 5. S krepko so označene besede, ki se v pomenu ujemajo s ključnimi besedami iz IMDb. Pojavljajo se imena ključnih oseb v filmu, tudi mesta in prostori ter ostale pogoste besede. Vidimo, da so v prevedenih ključnih besedah iz zbirke IMDb v tabeli 1, veliko bolj opisne besede, ki se v samih podnapisih nikoli ne pojavijo in jih je z našim pristopom nemogoče pridobiti. Te opisujejo dele filma in prizore, npr. presenetljiv konec in grafično nasilje, ali pa opisujejo druge podatke o filmu kot npr. “trilogija” ali “ki temelji na romanu”.

Kljub temu, je naše subjektivno mnenje, da z našim pristopom pridobljene besede relativno dobro opisujejo filme, čeprav je v primerjavi z IMDb majhno število zadetkov.

4.2 Primerjava

Avtorji prispevka [3] so dosegli najboljše rezultate z algoritmom TextRank, prepoznavo imenskih entitet in analizo pogovorov.

Zaradi različnega jezika in korpusov ne moremo neposredno primerjati rezultatov. Vsaka metoda je prilagojena svojemu jeziku a sledi istim načelom ekstrakcije ključnih besed. Primerjavo zglajenih rezultatov lahko vidimo v tabeli 6.

Tabela 6: Primerjava rezultatov v %

Metoda	*F1*	*P*	*R*
Prispevek [3] *(angleščina)*	16,99	30,30	11,80
Naš pristop *(slovenščina)*	13,869	13,870	13,869

4.3 Diskusija

Predlagan pristop k ekstrakciji ključnih besed ima dve veliki omejitvi:

a) Subjektivnost

Ključne besede za določen film so zelo subjektivne narave. Ni standardne definicije kaj je in kaj ni ključna beseda, saj je to odvisno od vsakega posameznika. To onemogoči kvantitativno ocenitev kvalitete algoritmov. Iz tega razloga sta naša eksperimenta uporabna samo za ocenitev vpliva posameznih delov in ne za ocenitev kvalitete pridobljenih ključnih besed.

b) Abstraktnost

Ključne besede pogosto vsebujejo abstraktne pojme, ki opisujejo film (npr. "razmerje oče-sin"). Te se skoraj nikoli ne pojavijo znotraj govora v filmu, kar onemogoči ekstrakcijo takšnih besed z našim pristopom.

5 Zaključek

V prispevku smo predstavili ogrodje za avtomatizirano ekstrakcijo ključnih besed filmov iz njihovih podnapisov. V našem pristopu v prvi fazi odstranimo nepotrebne značke HTML, ki jih lahko zasledimo v podnapisih, in besede brez pomena. Nato sledi faza filtriranja, kjer analiziramo in označimo besedne vrste. Posamezne stavke filtriramo tako, da odstranimo besedne vrste za katere ocenimo, da ne predstavljajo dobrih ključnih besed. Sledi analiza imenskih entitet, pri kateri odstranimo vsa lastna imena, ki ne ustrezajo domenam za katere ocenimo, da so nepomembne za ključne besede. V zadnji fazi ekstrakcije ključnih besed z uveljavljenimi algoritmi iz filtriranih podnapisov izluščimo ključne besede filma.

Za analizo kvalitete in konfiguracije (izbira besednih vrst za odstranitev, domen imenskih entitet za odstranitev in izbira algoritma) našega ogrodja smo pridobili slovenske podnapise filmov iz Podnapisi.NET in pred-definirane angleške ključne besede posameznih filmov iz baze IMDb. Ključne besede smo prevedli v slovenski jezik. Problem ekstrakcije ključnih besed smo, za potrebe vrednotenja, obravnavali kot klasifikacijski problem. Ključne besede iz zbirke IMDb in njihove sinonime smo obravnavali kot pravilne.

Analizirali smo kvaliteto izluščenih ključnih besed z algoritmoma TF-IDF in TextRank, pri čemer smo dodatno analizirali vpliv faze filtriranja (PIE). Rezultati nakazujejo, da dosega najboljše rezultate kombinacija TextRank + PIE z upoštevanjem sinonimov.

Dodatno smo analizirali tudi vpliv nabora besednih vrst, ki se odstranjujejo v fazi PIE. Iz rezultatov je razvidno, da daje naš pristop najboljše rezultate, če se odstranijo vse besedne vrste razen lastnih imen (filtriranje domen), samostalnikov in pridevnikov.

V povprečju smo dosegli ujemanje 13,96 ključnih besed na film, v najslabšem primeru 3 besede, v najboljšem primeru 33 besed.

5.1 Nadaljnje delo

Za ovrednotenje kvalitete pridobljenih ključnih besed s predlaganim pristopom bi potrebovali subjektivne ocene. Te bi lahko pridobili z izvedbo vprašalnika nad določeno populacijo.

Predvidevamo, da bi dosegli bolj smiselno razvrščene ključne besede z vpeljavo dodatne faze analize individualnih pogovorov. V tej fazi bi razdelili podnapise na individualne pogovore, te pa ovrednotili glede na pomembnost. Ključne besede, pridobljene iz pomembnih pogovorov, bi imele večjo težo kot tiste pridobljene iz nepomembnih pogovorov.

Morebitno izboljšanje bi lahko dosegli tudi z izbiro drugih algoritmov za ekstrakcijo ključnih besed in podrobnejšo analizo vplivov naših parametrov (množica dovoljenih leksikalnih domen, največ dovoljenih ponovitev, ...).

Zahvala

J. Brest in B. Bošković priznavata financiranje prispevka s strani Javne agencije za raziskovalno dejavnost Republike Slovenije, raziskovalni program P2-0041 – Računalniški sistemi, metodologije in inteligentne storitve.

Literatura

[1] Slobodan Beliga. Keyword extraction: a review of methods and approaches. *University of Rijeka, Department of Informatics, Rijeka*, pages 1–9, 2014.

[2] Fei Liu, Feifan Liu, and Yang Liu. A supervised framework for keyword extraction from meeting transcripts. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:538 – 548, 04 2011.

[3] Matúš Košút and Marián Šimko. Improving keyword extraction from movie subtitles by utilizing temporal properties. In Rūsiņš Mārtiņš Freivalds, Gregor Engels, and Barbara Catania, editors, *SOFSEM 2016: Theory and Practice of Computer Science*, pages 544–555, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[4] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[5] NLTK. NLTK Data. Dostopno na `https://github.com/nltk/nltk_data`, 2019. [Dostopano 5. maja 2020].

[6] Darja Fišer. Semantic lexicon of slovene sloWNet 3.1, 2015. Slovenian language resource repository CLARIN.SI.

[7] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[8] Nordic Language Processing Laboratory. NLPL word embeddings. Dostopno na `http://vectors.nlpl.eu/repository/`. [Dostopano 23. junija 2020].

[9] IMDb. `https://www.imdb.com/`. [Dostopano 18. junija 2020].

[10] Microsoft Translator. `https://www.microsoft.com/en-us/translator/`. [Dostopano 18. junija 2020].

[11] Podnapisi.NET. `https://www.podnapisi.net/`. [Dostopano 8. maja 2020].

Diferencialna evolucija za učenje agenta umetne inteligence pri igranju splošnih videoiger

Aleš Zamuda, Matjaž Vöröš

Univerza v Mariboru, FERI, Koroška cesta 46, 2000 Maribor
E-pošta: ales.zamuda@um.si, matjaz.voeroes@student.um.si

Differential Evolution for Artificial Inteligence Agent Learning in General Video Game Playing

Abstract. *In recent years, a new chapter for video gaming has opened in the form of General Video Game Artificial Intelligence (GVG-AI) Competition. The Competition challenges the contestant to implement an agent that can maximize the score of played video games with usage of modern optimization algorithms. In this paper, a GVG-AI agent built upon recent optimization algorithms is hence presented. This agent uses Differential Evolution (DE), which was not yet used in a GVG-AI Competition. The performance of the agent is assessed and compared using the GVG-AI benchmark and the results demonstrate that the agent is statistically significantly better than most of the existing ones, but there is still room for improvement.*

1 Uvod

V pametni proizvodnji postajajo z razvojem virtualizacijske tehnologije in zajemanja podatkov vse bolj ključne digitalne sence (digitalni dvojčki) [1]. Tako je tudi pri proizvodnji igralcev računalniških iger, ki skozi digitalno virtualnost dobivajo obliko računalniškega agenta napram poprej le človeških igralcev teh iger. Tak agent vsebuje umetno inteligenco (UI), angl. *Artificial Intelligence* (AI), ki je računska, angl. *Computational Intelligence* (CI), in opravlja operacijo v danem okolju raziskave (angl. *Operational Research* [2]), v tem primeru, igre. V tem prispevku je tako predstavljen pristop k modeliranju in proizvodnji takega agenta UI (AI/CI), kjer proizvodnjo vodimo s pomočjo optimizacijskega algoritma. S tem je tako izhodni rezultat kot učenje proizvedeno v UI, zato tak razvoj igralnega agenta lahko umestimo tudi v področje umetnega življenja (angl. *Artificial Life,* AL) [2].

Rešitve, ki s pomočjo UI igrajo določeno igro ali pa so narejene z določenim namenom, redko služijo čemu drugemu kot lastni domeni [3]. Pravo testno okolje za realno-časovne agente sta tako ogrodji ALE (angl. *Association for the Advancement of Artificial Intelligence*) [4] kot GVG-AI (angl. *General Video Game AI Competition*) [5]. Razlike med ogrodjema ALE in GVG-AI so predvsem, da ogrodje GVG-AI agenta neposredno informira o stanju okolja s pomočjo javanskih razredov in mu omogoča, da stanje razvija ali kopira, s tem pa dobiva informacije o možnih prihodnjih stanjih, v katere lahko napreduje iz trenutnega [3].

Pod pojmom agent sicer razumemo napravo ali program, ki sprejema podatke iz okolice s pomočjo svojih senzorjev in nanjo deluje z aktuatorji [6][7]. Pri inteligentnih agentih je naloga UI, da predstavlja takega agenta, ki signale pretvori v akcije [7]. Pri GVG-AI potrebujemo agenta**,** temelječega na cilju (angl. *Goal-Based Agent*) [7]. Poznati nekaj o trenutnem stanju okolja ni vedno dovolj za določitev naslednje akcije; torej na pravilno odločitev vpliva cilj, do katerega želimo priti [7]. Pot do cilja se tako določa s pomočjo iskalnih strategij in planiranja [6][7]. Različna uporaba UI v videoigrah je tako koristna za oblikovanje boljših in zanimivejših iger. UI ima pri igranju dve vidnejši vlogi [8]: ali videoigro igra kot igralec, s tem izboljša avtomatiziranje igre in omogoča lažji pregled ter ocenitev videoigre ali pa v videoigri sodeluje kot nasprotnik, kar lahko privede do lažje dinamične nastavitve težavnosti videoigre.

Za iskanje optimalne rešitve je potrebno definirati problem in izbrati ustrezno metodo [9]. Pri optimizacijskih problemih imamo podane množice možnih rešitev, kriterijsko funkcijo za vrednotenje in določene omejitve, iščemo pa najboljšo možno rešitev problema oziroma rešitev z najboljšo vrednostjo ocenitvene funkcije [10]. Pojem optimizacije je sicer precej širok, v splošnem pa se algoritmi za optimizacijo uporabljajo tako za UI, kot pri splošnem računanju [8]. Evolucijski algoritmi (EA) so globalni optimizacijski algoritmi [2], saj za razliko od lokalnih optimizacijskih algoritmov [3] [4] točke na začetku razporedimo po večjem prostoru, z njimi pa iščemo najboljšo možno globalno rešitev v nelinearnem odločitvenem modelu, ki pogosto vsebujejo nepopolne lokalne rešitve [5]. EA uporabljamo predvsem za reševanje težkih optimizacijskih problemov, npr. na področju strojnega učenja [2] [11]. V nadaljevanju predstavljen zasnovan agent iz [12] razvija načrt v navideznem modelu (digitalni senci) nekaj milisekund, nato pa deluje v videoigri skozi izvedbo prve akcije v snovanem načrtu ter se ponovno vrne v sprotno razvijanje tega načrta za igro [11]. Agent se od predhodnih agentov razlikuje po tem, da v učenju z EA vključuje nekatere izbrane operatorje modernih optimizacijskih algoritmov [2]. Za tega agenta poročamo tako teste in rezultate, ki jih primerjamo z drugimi pristopi nad operacijo igranja videoiger.

V naslednjem poglavju so predstavljena sorodna dela, v poglavju 3 sledi predstavitev snovanja algoritma GVG-AI in implementacija, nato so v poglavju 4 podani rezultati in v poglavju 5 zaključek.

2 Sorodna dela

V tem poglavju je najprej opisana diferencialna evolucija za optimizacijo, nato sledi še opis ogrodja za tekmovanje inteligentnih agentov za igranje videoiger.

2.1 Diferencialna evolucija za optimizacijo

Osnovno predlogo diferencialne evolucije (lahko tudi diferenčne evolucije, v nadaljevanju DE) [13] lahko opišemo v korakih EA iz [8]:

a) *Inicializacija*: začetno populacijo napolnimo z rešitvami, ki jih generiramo naključno; na primer naključne točke razporejene po iskalnem prostoru. Če že poznamo dobre rešitve, jih lahko vključimo v začetno populacijo. Naj bo $\boldsymbol{I}$ poljubna množica (prostor posameznikov) rešitev $\boldsymbol{x}_{i,G}$, ki tvorijo populacijo posamezne generacije G z velikostjo populacije μ^G, kjer je i v obsegu od $1,2,\dots,\mu^G;\ \mu^G, G \in \mathbb{N}$ [14].

Začetno populacijo v času $G = 0$ zapišemo kot nam kaže enačba (1). Pri DE se morajo komponente rešitve nahajati v mejah komponent $\boldsymbol{B}_{\min} = [B_{\min_1}, B_{\min_2}, \dots, B_{\min_D}]$ in $\boldsymbol{B}_{\max} = [B_{\max_1}, B_{\max_2}, \dots, B_{\max_D}]$ [2] (D nam predstavlja dimenzijo); izračun naključne komponente posameznika v mejah $[\boldsymbol{B}_{\min}, \boldsymbol{B}_{\max}]$ vidimo v enačbi (2). Tukaj $rand(0,1)$ označuje naključno spremenljivko v prostoru $\mathbb{R}$ med 0 in 1.

$$\boldsymbol{P}_0 := \{\boldsymbol{x}_{1,0}, \dots, \boldsymbol{x}_{\mu^0,0}\} \in \boldsymbol{I}^{\mu^0} \tag{1}$$

$$\boldsymbol{x}_{i,j,G} := B_{\min_j} + \left(B_{\max_j} - B_{\max_j}\right) \times rand(0,1) \tag{2}$$

b) *Ocenitev*: za ocenitev vseh rešitev v populaciji uporabimo funkcijo uspešnosti in vsem rešitvam dodelimo oceno. Tako s preslikavo $f : \boldsymbol{x}_{i,G} \to \mathbb{R}$ ocenimo uspešnost rešitve $\boldsymbol{x}_{i,G}$, npr. imenovano *evaluateGVGAI* za primer ocenitvene funkcije GVG-AI.

$$f(\boldsymbol{x}_{i,G}) := evaluateGVGAI(\boldsymbol{x}_{i,G}) \tag{3}$$

c) *Selekcija staršev*: pri DE za selekcijo uporabimo vse posameznike v populaciji in izmed posameznikov $\boldsymbol{x}_{i,G}, \forall i \in \{1,2,\dots,NP\}$, kjer je NP velikost populacije v trenutni generaciji G [15], $NP = \mu^G$, izberemo tri naključne posameznike $\boldsymbol{x}_{r_1,G}, \boldsymbol{x}_{r_2,G}, \boldsymbol{x}_{r_3,G}$, za katere velja $r_1 != r_2 != r_3 != i$, ter po enačbi (4) izračunamo novega posameznika $\boldsymbol{v}_{i,G+1}$ [2], F predstavlja skalirni faktor [15].

$$\boldsymbol{v}_{i,G+1} := \boldsymbol{x}_{r_1,G} + F \times (\boldsymbol{x}_{r_2,G} - \boldsymbol{x}_{r_3,G}) \tag{4}$$

d) *Reprodukcija*: potomce v generaciji ustvarimo s pomočjo križanja staršev. Pri DE najprej izračunamo spremenljivko $j_{rand} = randint(1, D)$, nato v zanki od $1, 2, \dots, D$ po enačbi (2.5), kjer CR predstavlja stopnjo križanja [15], $randint(1, D)$ pa predstavlja $\mathbb{N}$ vrednost v obsegu od 1 do D, izračunamo novega posameznika $u_{i,j,G+1}$ [16]. Tako vektor $\boldsymbol{v}_{i,G}$ kot $\boldsymbol{x}_{i,G}$ sta sestavljena iz komponent $\boldsymbol{v}_{i,G} = [v_{i,1,G}, v_{i,2,G}, \dots, v_{i,D,G}]$ oziroma $\boldsymbol{x}_{i,G} = [x_{i,1,G}, x_{i,2,G}, \dots, x_{i,D,G}]$. Pri križanju lahko uporabimo dvomestno križanje, večmestno križanje ali uniformno križanje [17]; slednje uporabimo pri DE kot:

$$u_{i,j,G+1} := \begin{cases} v_{i,j,G+1} & \text{če } rand(0,1) \le CR \text{ ali } j = j_{rand} \\ x_{i,j,G+1} & \text{sicer} \end{cases} \tag{5}$$

e) *Variacija*: mutacija se izvede nad nekaterimi starši in/ali potomci.

f) *Zamenjava*: v tem koraku izberemo, kateri izmed posameznikov (izbiramo tako med starši kot med potomci) je dovolj dober, da nadaljuje svoje poslanstvo v prihodnji generaciji. Priljubljene metode zamenjave so generacijska (angl. *generational*), kjer starši umrejo in jih nasledijo potomci, stabilna (angl. *steady state*), to uporablja tudi DE, kjer potomci nadomestijo starše le, če je njihova ocena boljša kot ocena staršev, in elitizem, ki je sicer podoben generacijski zamenjavi, a najboljših *x* % staršev preživi.

g) *Prekinitev*: vprašamo se, če smo končali. Odločitev temelji na tem, koliko generacij oziroma ocenitev je preteklo, najvišji oceni pridobljeni s strani rešitev in/ali kakšnem drugem zaključnem pogoju.

h) *Če še nismo končali, pojdimo na korak (b)*.

Vsaka iteracija glavne zanke (vsakič, ko pridemo do koraka *b*) se imenuje generacija, saj je terminologija povzeta iz teorije evolucije [2]. Končno število ocenitev je tako sorazmerna velikosti populacije krat številu generacij [8].

DE je novejši evolucijski algoritem, ki se v globalni optimizaciji uporablja za realno kodiranje numeričnih funkcij [15]. DE je zaradi svoje preprostosti, zmogljivosti in stabilnosti postal zelo priljubljen [18].

DE ima majhno število parametrov, vendar zaradi svoje stabilnosti in prilagajanja problemu, daje boljše rezultate od ostalih evolucijskih algoritmov [15].

Pri DE po začetni inicializaciji populacije, ki nam jo prikazuje enačba (1) sledita mutacija in križanje po enačbah (4) in (5). Po križanju s pomočjo enačbe (6) preverimo, ali se posameznik nahaja v območju $[\boldsymbol{B}_{\min}, \boldsymbol{B}_{\max}]$.

$$\boldsymbol{u}_{i,G+1} := \begin{cases} \boldsymbol{B}_{\min} + (\boldsymbol{B}_{\min} - \boldsymbol{u}_{i,G+1}) & \text{če } \boldsymbol{u}_{i,G+1} < \boldsymbol{B}_{\min} \\ \boldsymbol{B}_{\max} - (\boldsymbol{u}_{i,G+1} - \boldsymbol{B}_{\max}) & \text{če } \boldsymbol{u}_{i,G+1} > \boldsymbol{B}_{\max} \\ \boldsymbol{u}_{i,G+1} & \text{sicer} \end{cases} \tag{6}$$

Nazadnje sledi še izbira med staršem $\boldsymbol{x}_{i,G}$ in dobljenim vektorjem $\boldsymbol{u}_{i,G+1}$. Novega posameznika (za maksimizacijski problem) $\boldsymbol{x}_{i,G+1}$ tako dobimo po enačbi (7) [16].

$$\boldsymbol{x}_{i,G+1} := \begin{cases} \boldsymbol{u}_{i,G+1} & \text{če } f(\boldsymbol{u}_{i,G+1}) > f(\boldsymbol{x}_{i,G}) \\ \boldsymbol{x}_{i,G} & \text{sicer} \end{cases} \tag{7}$$

Med optimizacijskim procesom obdrži algoritem DE točno določene krmilne parametre. Težava se pojavi,

ker najbolj učinkovitih vrednosti pri algoritmu ne poznamo in so še vedno obstoječa vprašanja mnogih raziskovalcev.

Potrjena je bila tudi potreba po spreminjanju parametrov med procesom optimiziranja [16]. Tako je pri algoritmu jDE, ki razširja algoritem DE [18], vsak i-ti posameznik razširjen s samoprilagodljivima parametroma F in CR, ki ju naključimo s stopnjo verjetnosti τ_F oziroma τ_{CR}[16]. Nova parametra $F := F_{i,G+1}$ in $CR := CR_{i,G+1}$ se izračunata po enačbah (8) in (9) [16], kjer sta F_l in F_u spodnja in zgornja meja za F.

$$F_{i,G+1} := \begin{cases} F_l + rand_1 \times F_u & \text{če } rand_2 < \tau_F \\ F_{i,G} & \text{sicer} \end{cases} \quad (8)$$

$$CR_{i,G+1} := \begin{cases} rand_3 & \text{če } rand_4 < \tau_{CR} \\ CR_{i,G} & \text{sicer} \end{cases} \quad (9)$$

Ta metoda nastavljanja parametrov s poskušanjem (*angl. Trial-and-error method*) se uporablja za prilagajanje krmilnih parametrov in te nastavlja skozi posamezni zagon optimizacije [16][18].

2.2 Ogrodje tekmovanja inteligentnih agentov za igranje videoiger

Ker večino tekmovanj izziva inteligentne agente k igranju ene same igre, so se ta tekmovanja izkazala kot prevelika omejitev na določeno področje (domeno) [19].

V ta namen, kot protiutež dosedanjim tekmovanjem, je bilo ustvarjeno *mednarodno tekmovanje inteligentnih agentov za igranje videoiger*, ki za namene tekmovanja uporablja lastno ogrodje.

Agenti so tako izpostavljeni več, do uradnega dela tekmovanja še ne videnim igram, kar pomeni, da se morajo pripraviti na najrazličnejše možne scenarije. Vse različice tekmovanja GVG-AI se držijo enakega reda iger. Videoigre so razdeljene v različne sete (10 iger v setu, vsaka igra pa ima 5 različnih stopenj). Javni nabori, ki so vključeni v ogrodje so namenjeni tekmovalcem za preizkušanje izdelanega agenta.

Ogrodje GVG-AI lahko uporabimo tudi za postopno ustvarjanje vsebine (*angl. Procedural Content Generation*) [11]; ustvarimo lahko poljubne stopnje ali pravila. V obeh primerih ogrodje ponuja vnaprejšnje modeliranje (angl. *Forward Modeling*) [11], kar omogoča agentu, da stanje igre kopira in jo s pomočjo danih potez razvije, da bi s tem po možnosti dosegel naslednje stanje igre.

Ocena uspešnosti avtomatskega igranja se v GVG-AI izračuna po desetih igrah. Dejavniki, ki vplivajo na vrednotenje igre so: število zmag, skupno število točk (rezultat) in porabljen čas. Po vsakem odigranem krogu se tekmovalci ovrednotijo po sistemu točkovanja Formule 1 [19] in sicer po vrsti od prvega do desetega mesta s točkami 25, 18, 15, 12, 10, 8, 6, 4, 2 ter 1. Ostali prejmejo 0 točk [20]. Končen rezultat oziroma ovrednotenje prinese seštevek vseh pridobljenih točk. V primeru neodločenih izidov zmaga agent, ki ima več prvih pozicij, če pa sta tudi v tem segmentu agenta izenačena se štejejo druge, tretje, četrte pozicije in tako naprej [20].

V okviru ogrodja GVG-AI je en od že objavljenih agentnih algoritmov, EA sprotnega razvijanja (angl. *Rolling Horizion EA*, v nadaljevanju RHEA) [21], ki za iskanje uporablja preprost EA. Pri tem je EA z ogrodjem GVGA-AI povezan tako, da ocenitveno funkcijo optimizacijskega algoritma kličemo s pomočjo simulatorja v GVG-AI, kjer agent operira. Ocenitvena funkcija (za klic funkcije simulatorja dodamo vektorju $\boldsymbol{x}_{i,G}$ še trenutno stanje v igri in hevristiko) v EA nam za določen vektor $\boldsymbol{x}_{i,G}$ vrne dosežen rezultat iz simulatorja, pomnožen z $0{,}9^D$ (D je število dimenzij) [22]. V primeru, da je igre konec in je igralec izgubil, ocenitvena funkcija vrne vrednost $-1000 \times 0{,}9^D$ kar pomeni, da smo od zmage zelo oddaljeni. V primeru, da je igre konec in je igralec zmagal, ocenitvena funkcija vrne $1000 \times 0{,}9^D$ kar pomeni, da smo blizu zmagi. Če nismo dosegli nobenega izmed kriterijev, funkcija vrne $trenutni\ rezultat \times 0{,}9^D$. Optimizacijski problem v tej funkciji je predstavljen kot maksimizacija ocene. Ocenitvenih funkcij je posledično toliko, kolikor je možnih scenarijev – pod scenarije štejemo igranje različnih iger, vsaka igra pa ima več stopenj. Pri vsakem scenariju se namreč razlikujejo vhodi in stanje okolja, ki ju kot vektor $\boldsymbol{x}$ vključujemo v ocenitveno funkcijo.

$$f(\boldsymbol{x}_{i,G}) := \begin{cases} 1000 \times 0{,}9^D & \text{če } gameOver \text{ in } playerWins \\ -1000 \times 0{,}9^D & \text{če } gameOver \text{ in } playerLoses \\ rezultat \times 0{,}9^D & \text{sicer} \end{cases} \quad (10)$$

3 Snovanje agenta in implementacija

Snovanje agenta smo pričeli s pregledom obstoječih algoritmov za splošno igranje videoiger. Ugotovili smo, da so EA druga največja zastopana skupina algoritmov na tekmovanju. Odločili smo se za nadomestitev in delno nadgradnjo kode RHEA, v kateri smo dele RHEA zamenjali z algoritmom DE s prilagodljivimi krmilnimi parametri v tistem delu, ki implementira operatorje EA. Agenta smo poimenovali *VorosZamuda* [12]. Osnovan je na agentu *vzorčiRHEA* (angl. *sampleRHEA*), ki kot optimizacijski algoritem za GVG-AI uporablja algoritem RHEA. Ohranili smo dve obstoječi funkciji iz izvorne kode agenta *vzorčiRHEA, evaluate* in *get_best_action*, a spremenili vhodne parametre tako, da so slednji ustrezali našemu agentu. Nato smo implementirali pet novih funkcij, *initialize*, *mutate*, *crossover*, *correction* in *selection*, kamor smo vnesli kodo za računanje z DE iz poglavja 2.1. Vse skupaj smo uporabili v javni metodi *act*, ki je potrebna za zagon agenta. Podrobnejši psevdokodo metode *act* za algoritem *VorosZamuda* je dostopen v [12] na strani 34.

4 Rezultati

Odločili smo se, da lahko nekatere spremenljive parametre, ki jih imata oba agenta – velikost populacije in velikost dimenzije, pustimo na enakih vrednostih, za lažje kasnejše primerjanje rezultatov. Dodati smo morali tudi dve nastavljivi spremenljivki, in sicer skalirni

faktor ter verjetnost križanja, potrebni za izvedbo DE algoritma in dve spremenljivki $B_{\min}$, $B_{\max}$, ki smo ju nastavili v inicializaciji. Razen omenjenih spremenljivk smo ostale spremenljivke pustili nedotaknjene, torej enake kot pri agentu *vzorčiRHEA*.

Za večjo pristnost rezultatov smo se odločili, da oba agenta, tako *VorosZamuda*, kot *vzorčiRHEA*, testiramo na uradni spletni strani GVG-AI, namesto da bi zagone delali sami s pomočjo ogrodja. Tako smo na njihovem strežniku po nekaj urah izvajanja tega bremena hkrati dobili tudi ustrezno rangiranje obeh agentov med ostale agente, ki so na obstoječi lestvici uvrščeni od vključno 1. maja 2016 [23]. Na tem testnem naboru je agent *VorosZamuda* tako dosegel drugo mesto, kar lahko vidimo v tabeli 1. Bil je najboljše uvrščeni slovenski predstavnik. Agent *VorosZamuda* se je uvrstil takoj za agentom *combination*, dosegel pa je vsega 3 točke manj od prvouvrščenega agenta (*combination* je dosegel 114, *VorosZamuda* pa 111 točk). Za agenta *VorosZamuda* se je uvrstilo kar dvajset preostalih agentov.

Tabela 1. Tabela uvrstitve agentov.

Uvrstitev	Uporabniško ime	Država	Točke	Št. zmag (diskvalif.)
1	combination	Nemčija	114	23/50 (0)
2	**VorosZamuda**	Slovenija	111	17/50 (0)
3	fanatax	Nemčija	107	16/50 (0)
4	JACAM	Nemčija	79	12/50 (0)
5	asd592	Nemčija	71	11/50 (0)
6	matze1234	Nemčija	69	11/50 (0)
7	Jaybot	Nemčija	66	16/50 (0)
8	**vzorčiRHEA**	Slovenija	59	15/50 (0)
9	crazypet	Tajska	54	9/50 (0)
10	Cyclus	Nemčija	45	11/50 (0)
11	fraBOT2	Nemčija	44	13/50 (0)
12	JamieHutchison	VB	43	4/50 (9)
13	ToVo1	Slovenija	43	8/50 (0)
14	sampleOLMCTS	VB	42	11/50 (0)
15	sampleRS	VB	34	7/50 (0)
16	YOLOBOT	Nemčija	33	8/50 (0)
17	omnipotent	Nemčija	23	6/50 (0)
18	Damorin	VB	22	6/50 (0)
19	scott877	VB	22	4/50 (5)
20	sampleRHEA	VB	16	5/50 (0)
21	random	VB	14	4/50 (0)
22	treesakul	Tajska	12	3/50 (0)

Nad rangi agentov *VorosZamuda* in *vzorčiRHEA*, iz scenarija prikazanega v tabeli 1 smo izvedli še Wilcoxon Ranked-Sum Test [24]. Agent *VorosZamuda* je pri zagonih dosegel mesta 17, 4, 4, 1, 2, 13, 13, 3, 8 in 1, agent *vzorčiRHEA* mesta 3, 2, 18, 7, 8, 20, 13, 8, 17 in 4. Po izračunu smo dobili vrednost $p = 0{,}0858$. Tako smo ugotovili, da sta agenta z 91 % verjetnostjo signifikantno različna in pokazali, da je nov agent izboljšan.

5 Zaključek

V tem prispevku smo predstavili novega agenta za GVG-AI, ki deluje s pomočjo DE. Agenta smo ocenili in primerjali z drugimi in izkazalo se je, da z verjetnostjo 91 % algoritem deluje signifikantno boljše od osnovnega algoritma *vzorčiRHEA* ter zasede drugo mesto na rangu vseh preostalih algoritmov za planiranje v GVG-AI na podlagi uradnih meritev na strežniku za primerjalne teste GVG-AI.

V nadaljevanju bi algoritem lahko dodatno uglasili in razširili z izboljšanimi mehanizmi za optimizacijo in ga razširili na še več scenarijev v GVG-AI. Ta pristop AL bi lahko uporabili sicer tudi v katerem drugem okolju ML iz realnega sveta na primerih avtomobilske industrije ali globokomorske robotike.

Zahvala

Ta prispevek je nastal s podporo raziskovalnega programa P2-0041 pri ARRS ter CA15140 pri COST.

Literatura

[1] Y. Zheng, S. Yang in H. Cheng, „An application framework of digital twin and its case study,“ *Journal of Ambient Intelligence and Humanized Computing,* izv. 10, št. 3, str. 1141-1153, 2019.

[2] Zamuda, Aleš. "Operacijske raziskave logističnih, transportnih in ekonomskih sistemov: zbrano gradivo." (2020).

[3] J. Madge in T. Oplatek, „Game AI : Agent for GVGAI,“ 2018. http://www.tomasoplatek.com/ index.php/2018/07/09/game-ai/. [Dostop: 9. maj 2019].

[4] M. G. Bellemare, Y. Naddaf, J. Veness in M. Bowling, „The arcade learning environment: An evaluation platform for general agents,“ *Journal of Artificial Intelligence Research,* izv. 47, str. 253-279, 2013.

[5] D. Perez, „The General to Video Game AI Competition,“ [Elektronski]. Available: http://www.gvgai.net/. [Dostop 10. maj 2019].

[6] N. Guid in D. Strnad, Umetna inteligenca, Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko, 2007.

[7] S. J. Russell in P. Norvig, Artificial intelligence: a modern approach, Malaysia: Pearson Education Limited, 2016.

[8] G. N. Yannakakis in J. Togelius, Artificial Intelligence and Games, Springer, 2018.

[9] A. E. Eiben in J. E. Smith, Introduction to evolutionary computing, izv. 53, Berlin: Springer, 2003.

[10] B. Filipič, „Optimizacija in evolucijsko računanje,“ https://dis.ijs.si/filipic/courses/is/50_Optimizacija_in_evolucijsko_racunanje.pdf. [Dostop: 1. april 2019].

[11] D. Perez-Liebana, J. Liu, A. Khalifa, R. D. Gaina, J. Togelius in S. M. Lucas, „General Video Game AI: a Multi-Track Framework for Evaluating Agents, Games and Content Generation Algorithms,“ [Dostop: 18. marec 2019].

[12] Vörös, Matjaž. Evolucijski algoritmi za učenje agenta umetne inteligence pri igranju splošnih videoiger. Diss. Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, 2019.

[13] S. Rainer in K. Price, „Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces,“ *Journal of global optimization,* izv. 11, št. 4, str. 341-359, 1997.

[14] C. A. Coello, G. B. Lamont in D. A. Van Veldhuizen, Evolutionary Algorithms for Solving Multi-Objective Problems Second Edition, izv. 5, New York: Springer, 2007.

[15] A. Zamuda, „Magistrsko delo: Samoprilagajanje krmilnih parametrov pri algoritmu diferencialne evolucije za večkriterijsko optimizacijo,“ Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, 2008.

[16] A. Zamuda in J. Brest, „Self-adaptive control parameters' randomization frequency and propagations in differential evolution,“ Swarm and Evolutionary Computation, izv. 25, str. 72-99, 2015.

[17] M. Mernik, M. Črepinšek in V. Žumer, Evolucijski algoritmi, Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko, Inštitut za računalništvo, 2003.

[18] J. Brest, S. Greiner, B. Bošković, M. Mernik in V. Žumer, „Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems,“ IEEE Transactions on evolutionary computation, izv. 10, št. 6, str. 646-657, 2006.

[19] J. Liu, „GVGAI Single-Player Learning Competition at IEEE CIG17,“ 7 september 2017. [Elektronski]. https://tinyurl.com/y4hyvayh. [Dostop: 18. marec 2019].

[20] D. Perez, S. Samothrakis, J. Togelius, T. Schaul, S. M. Lucas, A. Couëtoux, J. Lee, C.-U. Lim in T. Thompson, „The 2014 General Video Game Playing Competition,“ v *IEEE Transactions on Computational Intelligence and AI in Games*, 2015.

[21] D. Perez, S. Samothrakis, S. Lucas in P. Rohlfshagen, „Rolling horizon evolution versus tree search for navigation in single-player real-time games,“ v *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, 2013.

[22] D. Perez, „Sample Controllers,“ GVGAI.net, http://www.gvgai.net/sampleControllers.php. [Dostop: 9. maj 2019].

[23] D. Perez, „Training Set Games 11 (Continuous Physics),“ http://www.gvgai.net/gvg_rankings.php?rg=11. [Dostop: 27. maj 2019].

[24] H. B. Mann in D. R. Whitney, „On a test of whether one of two random variables is stochastically larger than the other,“ *The annals of mathematical statistics,* izv. 18, št. 1, str. 50-60, 1947.

Uporaba metod strojnega učenja za analizo vpliva lipoproteinov (a) na bolezni srca in ožilja

Tajda Bogovič[1], Peter Kokol[1], Tadej Završnik[3], Jernej Završnik[2], Helena Blažun Vošner[2], David Šuran[3]

[1]*Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor*
[2]*Zdravstveni dom dr. Adolfa Drolca, Maribor*
[3]*Univerzitetni klinični Center Maribor, Maribor*
E-pošta: tajda.bogovic@student.um.si peter.kokol@um.si
tadej.zavrnik@student.um.si jernej.zavrsnik@zd-mb.si
helena.blazun@zd-mb.si david.suran@student.um.s

Using machine learning methods to evaluate lipoprotein (a) and its impact on cardiovascular diseases

Cardiovascular diseases are the number one cause of death and it is expected to remain so. Machine learning has had a significant impact in many areas, including various medical applications. The aim of the study was to analyse and evaluate lipoprotein a - Lp (a) as a cardiovascular risk factor by using regression machine learning techniques. The study was performed on the Anonymous Cardiovascular Database collected in the University Hospital of Maribor. The results indicate that lipoprotein a, cholesterol, triglycerides and salt are closely connected to each other.

1 Uvod

Kardiovaskularne bolezni oz. KVB so eden izmed glavnih vzrokov za smrtnost. Pojavitev bolezni z leti narašča. [1] Najpogostejša kardiovaskularna bolezen je t. i. koronalna arterija, ki jo povzroča premajhen pretok krvi v srčno mišico. Ta nastane zaradi nabiranja maščob, holesterola in ostalih substanc na stenah arterijskih žil. [2]
Dejavniki tveganja vključujejo družinsko anamnezo, visok krvni tlak, kajenje, povišano raven holesterola LDL, sladkorno bolezen, debelost, pomanjkanje rekreacije in prekomerno uživanje alkohola. [3] Številne študije [9, 7] nakazujejo tudi vpliv povišane vrednosti lipoproteina (a) na tveganje za KVB.
Lipoprotein (a) ali na kratko Lp (a) je velik delček lipoproteina, ki ga tvorijo jetra. Visoka raven Lp (a) je bila zdaj opredeljena kot neodvisen dejavnik tveganja za srčno-žilne bolezni, ki je vzročno povezana z aterosklerozo (krčenje arterij), srčnimi napadi, možgansko kapjo, boleznijo aortne zaklopke in srčnim popuščanjem. Meritve vrednosti Lp (a) se na splošno ne izvajajo pogosto, ampak je v veliki meri odvisno od družinskih anamnez oz. stanj, saj je Lp (a) genetsko pogojen. Trenutno znano najboljše zdravljenje visoke ravni Lp (a) je zmanjšanje LDL holesterola. [5] To poteka s pomočjo intravenoznega zdravljenja, ki ima lahko do 90 % uspešnost [4]. Večina dosedanjih študij [6] obravnava zgolj uporabo metod strojnega učenja za namen klasifikacije kardiovaskularnih zapletov, pri tem pa ne vključujejo povezave z vrednostjo Lp (a), saj je merjenje le-te zahtevnejši oz. kompleksnejši proces v primerjavi s pogostejšimi laboratorijskimi krvnimi analizami. [4] Zhao, Juan in drugi [12] navajajo uporabo statistične Pearsonove analize in logistične regresije za iskanje povezav Lp (a) in KVB. Rezultati so pokazali, da tisti posamezniki, ki imajo visoko vrednost Lp (a), bodo imeli večjo verjetnost za nastanek kardiovaskularnih zapletov v primerjavi s tistimi, ki imajo normalno oz. nizko vrednost. V študiji [10] so primerjali izvedbo treh odločitvenih dreves, štirih statističnih algoritmov in dveh nevronskih mrež za napovedanje hipertenzije. Atribut Lp (a) se je izkazal za zanesljiv kazalec pri napovedovanju. Li, Huidong [8] in Xu, Yuan [11] so uporabili metodi logistično regresijsko analizo in XGboost model za diagnosticiranje ateroskleroze oz. napovedovanje možganske kapi pri bolnikih, ki so že predhodno imeli ishemično kap. Bazi, uporabljeni v študijah, sta vsebovali atribut Lp (a).

2 Metode

V naši raziskavi smo uporabili anonimno bazo kardovaskularnega oddelka Univerzitetnega kliničnega centra Maribor v obdobju 1999—2019. Baza vsebuje podatke 4477 bolnišničnih obiskov z 3940 različnimi vnosi in 39 pripadajočimi atributi, ki so podrobneje opisani v tabeli 1.

2.1 Predprocesiranje podatkov

Atribute, kateri so imeli več kot 95 % manjkajočih vrednosti (z izjemo zaporednega števila), smo odstranili, in sicer: zap. št, FT3, FT4, TSH, S-NT-pro BNP, S-magnezij (Mg), S-fosfat (P), S-apolipoprotein A1, S-AST (GOT), S-CRP, S-HDL-holesterol, S-LDH, S-alkalna fosfataza, S-bilirubin celokupni, S-bilirubin direktni, S-holesterol, S-kalij (K), S-kloridi (Cl), S-natrij (Na), S-trigliceridi, S-urea (sečnina), S-Troponin T.

Atributom Ocena GF, S-ALT (GPT), S-AST (GOT), S-CRP, S-alkalna fosfataza, S-bilirubin celokupni, S-bilirubin direktni, S-gama GT S-kalij (K), S-kloridi (Cl), S-kreatinin-E, S-natrij (Na) S-troponin I, S-urea (sečnina), S-HDL-holesterol, S-LDH, S-holesterol, S-trigliceridi, S-Lp (a), S-srčni troponin I in S-Troponin T smo manjkajoče vrednosti dopolnili s povprečno vrednostjo.

Atributa diag1 in sprejemnaDiagnoza smo spremenili v kategorični tip podatka, saj so vpisane le tiste diagnoze, ki se začnejo s šifrantom I25.

Tabela 1: Atributi baze kardovaskularnega oddelka Univerzitetnega kliničnega centra Maribor

Ime atributa	**Tip atributa**	**Opis atributa**
zap. št.	Celo število	Identifikacija pacienta
smrt	Kategorični podatek	0- Ne 1- Da
spol	Kategorični podatek	0 - Ženski 1- Moški
sprejemnaDiagnoza	Niz	Diagnoze I25.X
AH	Kategorični podatek	Prisotnost arterijske hipertenzije 0 - Ne 1 - Da
SB	Kategorični podatek	Prisotnost sladkorne bolezni 0 - Ne 1 - Da
SP	Kategorični podatek	Prisotnost srčnega popuščanja 0 - Ne 1 - Da
KBL	Kategorični podatek	Prisotnost kronične ledvične bolezeni 0 - Ne 1 - Da
D H HTg	Kategorični podatek	Prisotnost hipertrigliceridemije 0 - Ne 1 - Da
diag1	Niz	Diagnoze I25.X
star	Celo število	Starost pacienta izražena v letih
fpri	Kategorični podatek	Št. oddelka
Ocena GF	Decimalno število	Vrednosti laboratorijskih oz. krvnih analiz
S-ALT (GPT)	Decimalno število	
S-AST (GOT)	Decimalno število	
S-CRP	Decimalno število	
S-alkalna fosfataza	Decimalno število	
S-bilirubin celokupni	Decimalno število	
S-bilirubin direktni	Decimalno število	
S-gama GT	Decimalno število	
S-kalij (K)	Decimalno število	
S-kloridi (Cl)	Decimalno število	
S-kreatinin-E	Decimalno število	
S-natrij (Na)	Decimalno število	
S-troponin I	Decimalno število	
S-urea (sečnina)	Decimalno število	
S-HDL-holesterol	Decimalno število	
S-LDH	Decimalno število	
S-holesterol	Decimalno število	
S-trigliceridi	Decimalno število	
S-Lp (a)	Decimalno število	
S-srčni troponin I	Decimalno število	
S-Troponin T	Decimalno število	
FT3	Decimalno število	
FT4	Decimalno število	
TSH	Decimalno število	
S-NT-pro BNP	Decimalno število	
S-magnezij (Mg)	Decimalno število	
S-fosfat (P)	Decimalno število	
S-apolipoprotein A1	Decimalno število	

Po predpocesiranju smo tako dobili 3427 vnosov z 32 različnimi atributi.

2.2 Uporaba metod

V študiji smo uporabili različne regresijske metode za odkrivanje povezav med izbranim atributom Lp (a) in ostalimi atributi. Izbrali smo modele odločitvenih dreves, naključnih gozdov, GradientBoost in AdaBoost, ki smo jih implementirali s pomočjo odprto-kodne Python knjižnice Scikit-learn.

S pomočjo omenjenih regresijskih metod smo ugotovili, kateri atributi imajo največjo korelacijo z atributom Lp (a) in kakšna je bila povprečna kvadratna napaka pri posameznem modelu.

Odločitveno drevo je imelo parameter maksimalne globine enak 4. Model naključnih gozdov je vključeval 500 iteracij, maksimalno globino 5 in minimalno razdelitev 2. GradientBoost in Adaboost sta imela 500 iteracij, stopnjo učenja 0.01. Funkcija izgube je v primeru GradientBoost bila regresija najmanjših kvadratov (ang. least squares regression), medtem ko je pri Adaboost bila linearna.

.

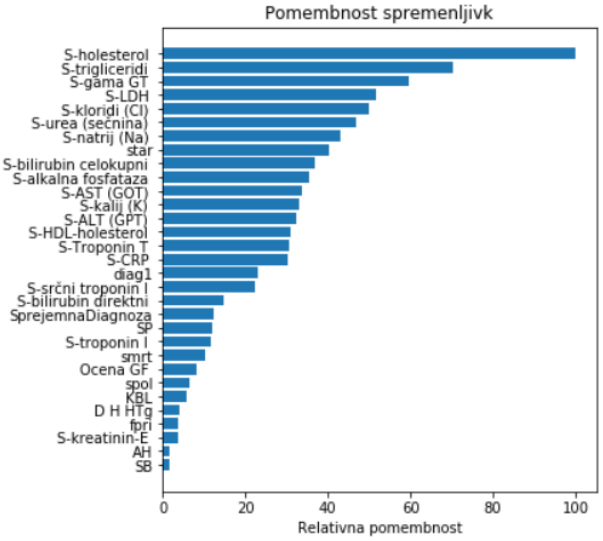
3 Rezultati in diskusija



Slika 1: Pomembnost spremenljivk v primeru regresijskega odločitvenega drevesa

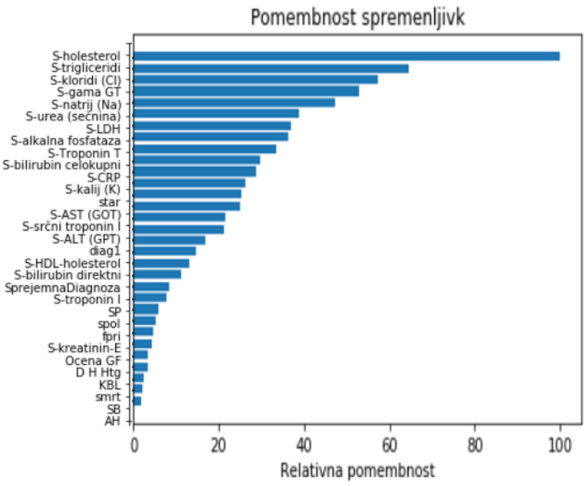
Iz slike 1 lahko razberemo pomembnost posameznih atributov z atributom Lp (a) pri uporabi regresijskega odločitvenega drevesa. Izkaže se, da je atribut S-holesterol eden izmed najpomembnejših, saj znaša njegova relativna pomembnost skoraj 100 %. Sledita atributa SP (67 %) in S-Trigliceridi (57 %). Prav tako ne gre zanemariti atributov spol, KBL in S-Troponin T, ki so dosegli več kot 30 % relativno pomembnost.

Slika 2 prikazuje rezultat pomembnosti atributov pri uporabi regresijskega modela GradientBoost. Opazimo lahko, da je največja povezava med vrednostima Lp (a)



Slika 2: Pomembnost spremenljivk v primeru regresijskega modela Gradientboost

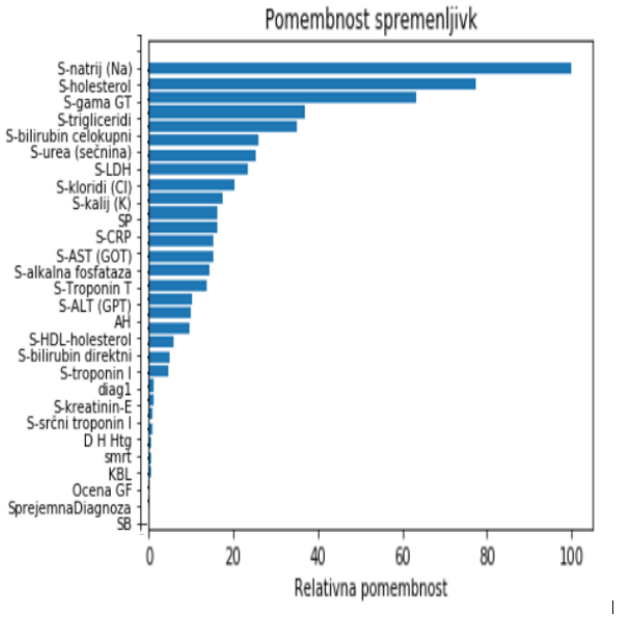
in S-holesterolom. Sledijo atributi S-trigliceridi, S-gama GT, S-LDH in S-kloridi, ki imajo 50 ali več procentno relativno pomembnost. Med 30 in 50 % pomembnost so dosegli S-urea (sečnina), S-natrij (Na), starost, S-bilirubin celokupni, S-lokalna fosfataza, S-AST, S-kalij, S-ALT, S-HDL-holesterol, S-Troponin T in S-CRP. Pri ostalih znaša manj kot 30 %.



Slika 3: Pomembnost spremenljivk v primeru regresijskega modela naključnih dreves

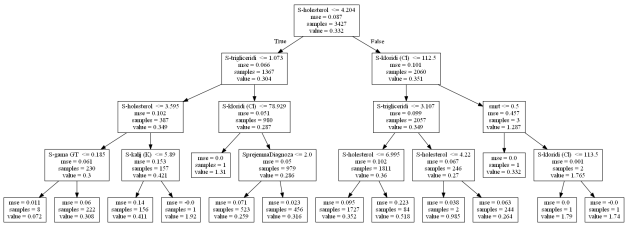
Izris pomembnosti atributov modela naključnih gozdov je viden na sliki 3. Ponovno je S-holesterol dosegel najvišjo vrednost relativne pomembnosti z Lp (a), saj je ta znašala 100 %. Visok delež so dosegli tudi atributi S-trigliceridi, S-kloridi in S-gama GT.

Za razliko od modelov odločitvenega drevesa, GradientBoost in naključnih dreves je AdaBoost dosegel najvišji procent (100 %) relativne pomembnosti z atributom S-natrij, kar lahko razberemo tudi na sliki 4. Sledita S-



Slika 4: Pomembnost spremenljivk v primeru regresijskega modela Adaboost

holesterol in S-gama GT, oba z več kot 50% relativno pomembnostjo.



Slika 5: Izris regresijskega odločitvenega drevesa

Odločitveno drevo, ki smo ga dobili kot rezultat uporabe regresijskega odločitvenega drevesa, vidnega na sliki 5, potrjuje pomembnost atributov S-holesterola, S-trigliceridov in soli, kot sta S-natrij in S-kalij. Listi dreves so vrednosti atributa Lp (a).

Tabela 2: Napaka regresijskih modelov

Regresijski model	Povprečna kvadratna napaka
Odločitveno drevo	0.0898
GradientBoost	0.0869
Naključna drevesa	0.0860
AdaBoost	0.1344

Kot je razvidno iz tabele 2 so vsi modeli imeli majhno povprečno kvadratno napako. Najmanjšo smo zabeležili pri naključnih drevesih, sledita GradientBoost in odločitveno drevo. Najvišjo napako je imel model AdaBoost.

Vsi modeli so pokazali, da je prisotna korelacija med S-holesterolom in Lp (a) vrednostjo. Prav tako se je pokazala pomembnost S-trigliceridov z Lp (a). V treh primerih je slednja znašala (skoraj) 100 %. Le v primeru modela AdaBoost lahko opazimo nekoliko drugačne rezultate, kar pa se tudi kaže v izmerjeni povprečni kvadratni napaki. Ta je bila v primerjavi z ostalimi modeli nekoliko višja.

Glede na to, da je meritev vrednosti trigliceridov in holesterola pogostejša in enostavnejša, lahko povzamemo, da bi v primeru visoke vrednosti teh atributev bilo smiselno izmeriti oz. pridobiti tudi gensko vrednost Lp (a).

4 Zaključek

Prisotnost kardivaskularnih bolezni predstavlja večjo umrljivost in večjo možnost pridružitve zdravstvenih zapletov oz. novih diagnoz. V tej študiji smo podrobneje pregledali in analizirali povezanost vrednosti Lp (a) z ostalimi dejavniki tveganja KVB. Pri tem smo uporabili različne regresijske tehnike strojnega učenja, in sicer odločitvena drevesa, naključne gozdove, GradientBoost ter AdaBoost model. Rezultat uporabe tehnik v skoraj vseh primerih nakazuje povezanost izbranega atributa z atributi S-holesterol, S-trigliceridi in soli.

Literatura

[1] Benjamin, Emelia J., et al. Heart disease and stroke Statistics-2019 update a report from the American Heart Association. Circulation, 2019.

[2] Diseases and Conditions - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions (pridobljeno 19. 6. 2020)

[3] Hajar, Rachel. Risk factors for coronary artery disease: historical perspectives. Heart views: the official journal of the Gulf Heart Association, 2017, 18.3: 109.

[4] HEART, U. K. The cholesterol charity. Bridging the gap. UK: Heart the cholesterol charity, 2013.

[5] High Lipoprotein(a). https://www.heartuk.org.uk/genetic-conditions/high-lipoproteina (pridobljeno 27. 8. 2020)

[6] Kadi, Ilham, Ali Idri, and José Luis Fernandez-Aleman. Systematic mapping study of data mining–based empirical studies in cardiology. Health Informatics Journal 25.3 (2019): 741-770.

[7] Kassner, Ursula, et al. Lipoprotein (a)–an independent causal risk factor for cardiovascular disease and current therapeutic options. Atherosclerosis Supplements, 2015, 18: 263-267.

[8] Li, Huidong, et al. Modeling analysis of the relationship between atherosclerosis and related inflammatory factors. Saudi journal of biological sciences, 2017, 24.8: 1803-1809.

[9] Nordestgaard, Børge G., et al. Lipoprotein (a) as a cardiovascular risk factor: current status. European heart journal, 2010, 31.23: 2844-2853.

[10] Ture, Mevlut, et al. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications, 2005, 29.3: 583-588.

[11] Xu, Yuan, et al. Extreme gradient boosting model has a better performance in predicting the risk of 90-day readmissions in patients with ischaemic stroke. Journal of Stroke and Cerebrovascular Diseases, 2019, 28.12: 104441.

[12] Zhao, Juan, et al. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein (a)(LPA). PloS one, 2019, 14.2: e0212112.

Pregled nekaterih metod na področju analize časovnih vrst

Žiga Stržinar[1,2], Boštjan Pregelj[1], Igor Škrjanc[2]

[1]*Odsek za sisteme in vodenje, Institut "Jožef Stefan", Ljubljana*
[2]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
E-pošta: ziga.strzinar@ijs.si

Overview of some methods in the field of time series analysis

Time series analysis is gaining importance in recent years, as more and more temporal data are being collected. This paper presents the challenges faced in the field. Three analysis steps are further explored: time series representation, time series distance measures and time series clustering. For each step, relevant methods are presented and evaluated. This paper aims to provide an insight into the field and expose some common concepts, challenges and solutions. The methods chosen here, are a basis for further work in the monitoring of dynamic processes.

1 Uvod

Vedno večja razširjenost senzorjev, pojav interneta stvari (*ang. Internet of Things* - IoT), dostopnost orodij za hrambo velikih količin podatkov, pojav nosljive elektronike, digitalizacija poslovnih procesov in drugi tehnološki dosežki v zadnjih letih so omogočili izgradnjo velikih podatkovnih zbirk časovno označenih podatkov - časovnih vrst (ang. *time series*). Področje analize, obdelave in odkrivanja znanj v teh podatkih zato pridobiva na pomenu na različnih področjih življenja: zdravstvo (nosljiva elektronika, elektrokardiogrami, itd.), finančna industrija (nihanja cen delnic, tečajev in drugih finančnih instrumentov), transport (podatki o prometu), proizvodna dejavnost, itd.

Namen tega prispevka je predstaviti pristope primerne za analizo časovnih vrst. Podani so temeljni problemi s katerimi se srečujemo, predstavljene so nekatere pogosto uporabljene metode.

2 Analiza časovnih vrst

Področje analize časovnih vrst je zelo široko, saj dandanes časovne vrste izvirajo iz zelo različnih virov, problemi, ki se rešujejo pa so zelo raznoliki. Literatura ([1], [8]) izpostavlja več namenov analize časovnih vrst. Pogosta naloga je iskanje podobnosti (ang. *similarity search*) in pa ujemanje sekvenc (ang. *subsequence matching*). Tipičen cilj analize časovnih vrst je tudi zaznavanje odstopanj (ang. *anomaly detection*), kjer se znotraj časovne vrste iščejo nepričakovani dogodki, odstopanja od običajnega delovanja. Na področju sta pomembna postopka rojenje (ang. *clustering*) in klasifikacija (ang. *classification*). Pri prvem je želja zbirko časovnih vrst urediti v roje (tudi gruče, ang. *clusters*). Pri klasifikaciji pa se novo časovno vrsto razvršča v enega od vnaprej določenih razredov. Daljše časovne vrste je včasih potrebno segmentirati v krajše segmnete, ki so bolj primerni za nadaljnjo obdelavo. Pri segmentaciji je ključno vprašanje kako določiti prlomne točke.

Pri delu se ukvarjamo s problemom zaznavanja dogodkov na proizvodnih linijah. Na voljo so časovno označene meritve z opazovanega sistema, želja pa je iz teh meritev razpoznati različne faze delovanja ali korake v proizvodnem procesu. Za doseganje tega cilja je potrebno poznavanje metod analize časovnih vrst predstavljeno v tem prispevku.

V naslednjih poglavjih so predstavljene pogoste metode za doseganje omenjenih ciljev. V poglavju 3 so predstavljene metode za predstavitev časovnih vrst. Poglavje 4 se osredotoča na mere razdalje primerne za delo s časovnimi vrstami, poglavje 5 pa obravnava metode za rojenje.

3 Metode za predstavitev časovnih vrst

Pri obdelavi časovnih vrst je opravka z velikimi količinami podatkov (t.i. *big data*) zato je bil predlagan pristop, ki surove podatke (izvirno časovno vrsto) najprej preslika v aproksimirano predstavitev. Komprimirano predstavitev je možno shraniti v pomnilnik. V drugem koraku je dan problem okvirno rešen, podan je en ali več predlogov za končno rešitev - kandidatov. Te kandidate se v zadnjem koraku preveri z dostopom do izvirnih podatkov. Celoten pristop temelji na ustreznosti prvega koraka - dobri predstavitvi podatkov. V tem poglavju so zato predstavljeni uveljavljeni in obetavni pristopi za predstavitev časovnih vrst.

Vzorčenje (ang. *sampling*) je najenostavnejši pristop, pri katerem se časovno vrsto namesto z vsemi vzorci predstavi samo z vsakim n-tim vzorcem.

PAA je alternativa vzorčenju. Metoda poznana tudi kot *Piecewise Aggregate Approximation* ([14]) kot nov vzorec vzame povprečno vrednost n zaporednih vzorcev izvirne časovne vrste. Število vzorcev po PPA je enako, kot se dobi z vzorčenjem, vendar je metoda manj občutljiva na šum in osamelce.

PLA (ang. *Piecewise Linear Approximation*), je metoda predstavljena v [13]. Gre za pristop, ki segment nefiksne dolžine opiše s parametri linearne funkcije. Časovno vrsto se najprej segmentira v K ne-enako-dolgih segmentov, končni točki vsakega segmenta sta elementa osnovne časovne vrste. Konec enega segmenta je začetek naslednjega. Algoritem začne s segmenti dolžine 3. Za vsako premico se izračuna normirana napaka po formuli $e_i = \frac{\sum_{m=1}^{j} d_m^2}{j}$, kjer je j število vzorcev v segmentu, i je indeks segmenta. Izračuna se standardna deviacija e_i čez vse segmente in se označi kot B_k (k je število segmentov). Algoritem se nadaljuje tako, da se združita dva sosednja segmenta, število segmentov se s tem zmanjša za 1. Algoritem se ponavlja, dokler ne ostane le en segment. V vsakem koraku se izbere združitev segmentov (ang. *merging*), ki bo rezultirala v najmanjšem B_k v naslednji iteraciji. Na koncu se uporabi tisto število segmentov k, kjer je B_k najmanjši, torej kjer je napaka najbolj konstantna čez vse segmente.

PIP algoritem predstavljen v [6, 9] časovno vrsto predstavi kot zaporedje vizualno pomembnih točk (*Perceptually Important Points*). Za časovno vrsto P algoritem iskanje točk začne z označitvijo robnih točk kot pomembnih. V naslednji iteraciji se poišče tista točka na P, ki ima največjo razdaljo od obstoječih pomembnih točk. V vseh naslednjih iteracijah se kot pomembna označi točka, ki ima največjo razdaljo do dveh sosednjih pomembnih točk. Algoritem se ponavja dokler vse točke na P niso označene kot pomembne. Točke P so po zadnji iteraciji urejene po pomembnosti.

SAX je v zadnjem času pogosta metoda za predstavitev časovne vrste s simboličnim zapisom. SAX je kratica za *Symbolic Aggregate approXimation)* in je predstavljena v [16]. SAX zahteva, da se na časovni vrsti najprej uporabi PAA. Algoritem v nadaljevanju sloni na predpostavki, da ima osnovna časovna vrsta Gaussovo normalno porazdelitev. Glede na to porazdelitev se nato določi nabor prelomnih točk med katerimi je integral funkcije gostote verjetnosti (PDF) enak. Vsakemu intervalu med dvema prelomnima točkama pripada en simbol uporabljene abecede. Časovno vrsto se predstavi z zaporedjem simbolov glede na to v kateri interval 'pade' posamezen segment dobljen s PAA.

Pristrižena predstavitev (ang. *clipped representation*) je pristop za zelo zgoščeno predstavitev časovne vrste, kjer se le ta predstavi kot zaporedje bitov (npr. 10011000), kjer znak 1 pove, da je vrednost večja od povprečja časovne vrste, 0 pa, da je manjša oz. enaka.

Predstavitev skozi rojenje V [8] je naštetih več primerov, kjer avtorji opisujejo izračun značilk oziroma predstavitev časovnih vrst z rojenjem.

Nekateri avtorji segmentirajo časovno vrsto ter vsak segment predstavijo z enostavno obliko (npr. strmina). Segmente se nato roji, s tem se dobi za vsak segment identifikator roja kateremu pripada. Zaporedje identifikatorjev se uporabi kot nova predstavitev časovne vrste.

Glede na dinamično naravo časovnih vrst je obetaven pristop tudi uporaba samorazvijajočih algoritmov rojenja (angl. *evolving clustering* [17]), ki so prilagojeni za spremljanje in analizo toka podatkov, kar je pogosta zahteva pri analizi časovnih vrst.

4 Mere razdalje

V tradicionalnih podatkovnih zbirkah so mere razdalje med vzorci definirane na osnovi točnega ujemanja med dvema vzrocema, na področju analize in obdelave časovnih vrst pa izbira mere še bolj odvisna od cilja ([8, 1]), ki je lahko:

- podobnost v istih časovnih trenutkih, iskanje časovnih vrst, ki se istočasno enako obnašajo?
- podobnost po obliki, neodvisno od absolutne vrednosti časa (premik o časovni osi)
- podobnost po modelu, kjer se iščejo časovne vrste pri katerih v ozadju veljajo enaka pravila

Evklidska razdalja se v analizi časovnih vrst uporablja na dva načina: 1) za direktno primerjavo dveh enako dolgih časovnih vrst (podobnost v istih časovnih trenutkih) 2) za primerjavo transformiranih zapisov časovnih vrst, na primer primerjava parametrov modelov, DFT ali DWT koeficientov.

DTW je na področju analize časovnih vrst pogosto uporabljena mera, podana v [5]. Glavna značilnost je podana že z imenom metode - dinamično časovno popačenje (ang. *Dynamic Time Warping*). Po tej metodi se za dve časovni vrsti Q in P dolžin n in m zgradi matrika n x m, kjer vsak element predstavlja razdaljo med dvema odčitkoma časovnih vrst. Tipično se uporabi evklidska razdalja [8]. Glede na vnaprej določene omejitve (začetna in končna točka, monotonost) se določi optimalna pot - t.i. pot popačenja (ang. *warping path*).

LCSS , razdalja najkrajšega skupnega izseka (ang. *Longest Common Subsequence*) je definirana kot razmerje med najdaljšim skupnim zaporedjem v dveh časovnih vrstah in dolžino celotne časovne vrste [11]. Parameter izračuna razdalje je toleranca – meja pod katero sta dve točki označeni, da se ujemata. Izračun je podoben kot pri DTW.

MVM je obetavna mera, ki obljublja izboljšavo delovanja napram DTW. Prednost pred DTW so boljši rezultati, ko sta primerjani časovni vrsti različno dolgi, metoda namreč dovoljuje preskakovanje ne-dodeljenih 'repov' časovnih vrst.

5 Rojenje

Rojenje je postopek za nenadzorovano analizo množice vzorcev z razdelitvijo v podmnožice - roje, na način, da

so si pripadniki istega roja med seboj bolj podobni, kot so podobni pripadnikom drugih rojev.

V literaturi ([1, 7, 10, 12, 15]) se metode za rojenje delijo v več skupin:

Delitvene metode (ang. *Partitioning methods*) so metode, ki vse vzorce glede na neko mero razdajle razdelijo v posamezne roje. Globalni optimum je težko zagotoviti, običajno se te metode zadovoljijo že z lokalnim minimumom. Med te metode prištevamo k-Means, k-Medoids, Gustafson-Kessel, fuzzy c-means ([7] našteje številne nadgradnje)...

k-Means (težišče roja je povprečje vzorcev), k-Medoids (težišče roja je centralni vzorec) – oba sta primerna za sferične roje in majhne do srednje velike množice. Algoritma sta pogosto uporabljena tudi kot osnova za druge algoritme ali za primerjavo uspešnosti ([11]). Algoritma sta primerna tudi za uporabo z različnimi merami razdalje. Najpogostejša mera je evklidska razdalja, v posebnih primerih, pa so potrebne druge mere, pri časovnih vrstah je denimo pogosta kombinacija k-means in DTW razdalje.

Gustafson-Kessel je adaptivni algoritem mehkega rojenja primeren tudi za ne-sferične roje [15].

Delitvene metode so zaradi enostavnosti in dobrih rezultatov prav tako zelo pogosti na področju analize časovnih vrst. k-Means je pogosta metoda s katero se primerjajo ostale metode.

Hierarhične metode ustvarjajo hierarhično strukturo primernih rojitev. Obstajata dva pristopa - združevanje (ang. *agglomerative*) ter razdiranje, eljenje (ang. *divisive*). Za hierarhično rojenje je bistvena izbira kriterija za razdelitev oziroma združevanje rojev, saj združenega roja tipično ni več mogoče razdeliti, razdeljenega roja pa ne združiti. Metode lahko temeljijo na razdalji, gostoti ali zveznosti. Izpeljanke tega pristopa se uporabljajo tudi za rojenje časovnih vrst [1, 15]. Pogosto navedena prednost teh metod je zmožnost intuitivne vizualizacije poteka rojenja.

Bistvena je izbira primerne mere razdalje, pogosto je uporabljena DTW (poglavje 4). V novejših člankih se pogosto omenja tudi uporaba SAX za predstavitev in povezano mero razdalje.

Hierarhično rojenje časovnih vrst na težave naleti pri dolgih časovnih vrstah in zelo velikih podatkovnih množicah [1].

Metode na osnovi gostote ne temeljijo neposredno na razdalji med vzorci temveč na gostoti. S tem se izognejo težavi ostalih pristopov, ki so osredotočeni na sferične roje. Te metode so primerne za roje arbitrarnih oblik [10, 7]. Te metode povečujejo roj dokler je gostota okolice (ang. *neighborhood*) večja od nekega praga (ang. *threshold*). Metode so primerne za izločevanje šuma in osamelcev [10]. Primera teh metod sta DBSCAN in OPTICS.

Mrežne metode prostor kvantizirajo v mrežo. Vse nadaljnje operacije se dogajajo v novem prostoru, kar pohitri procesiranje [10]. Ta pristop se lahko uporabi v kombinaciji z ostalimi.

Metode z uporabo modelov za vsak roj privzamejo nek model. Pogosta metoda je SOM - Samo-organizirajoče mreže (ang. *Self-organizing maps*). Tako kot mnogi drugi pristopi SOM zahteva enotno dolžino vzorcev. Raziskujejo se tudi pristopi s polinomskimi modeli, modeli Gaussovih mešanic, verigami Markova itd.

Samorazvijajoče metode (ang. *evolving*) imajo posebno lastnost, da se roji prilagajajo (učijo, razvijajo) toku podatkov sproti, ko ti prihajajo. Algoritmi so zasnovani tako, da omogočajo odmik podatkov od običajne delovne točke. Odmik je lahko postopen (ang. *concept drift*) ali hipen (ang. *concept shift*) [18].

Samorazvijajoči algoritmi pridejo v poštev na več področjih: rojenje, iskanje vzorcev, zaznavanje anomalij, diagnostika [18]. Izvedenih je bilo že več konkretnih implementacij samorazvijajočih algoritmov npr. samorazvijajoč krmilnik [2], samorazvijajoča indentifikacija [4, 3], prepoznavanje dogodkov pri vožnji [17], itd.

Stohastična optimizacija Poleg že opisanih skupin metod za rojenje je za rojenje možno uporabiti tudi metode stohastične optimizacije. Gre za metode namenjene optimizaciji poljubne kriterijske funkcije s stohastičnim pristopom, pogosto z mnogokratnimi evalvacijami kriterijske funkcije pri različnih vhodnih parametrih. Glede na cilje rojenja se lahko zasnuje kriterijska funkcija, ki oceni razdelitev v roje. Z uporabo metod stohastične optimizacije je nato možno najti (glede na kriterijsko funkcijo) primerno razdelitev. Med metodami, ki se lahko uporabijo najdemo PSO (ang. *Particle Swarm Optimization*), genetske algoritme in druge.

5.1 Posebnosti rojenja časovnih vrst

Namen rojenja časovnih vrst je enak kot pri običajnem rojenju, razlika je le v vzorcih, ki se jih roji. Pri običajnem rojenju se rojijo vzorci sami, pri časovnih vrstah pa je to lahko ne-praktično. Časovno vrsto lahko sicer obravnavamo kot visokodimenzionalen vzorec (število dimenzij je dolžina časovne vrste). To lahko, odvisno od primera, rezultira v zelo visoko-dimenzionalnem problemu, ki lahko povzroča težave napačno izbranemu algoritmu rojenja. Prav tako tak pristop ne zaobjame specifike časovnih vrst (pomen oblike, amplitudnega zamika itd.) ter težje obravnava časovne vrste različnih dolžin.

Po [15] se algoritmi za rojenje časovnih vrst delijo:

1. Algoritmi, ki prilagodijo klasičen algoritem tako, da uporabijo mero podobnosti/razdalje, ki je primerna za časovne vrste. Ti algoritmi uporabljajo surovo časovno vrsto, 'klasičen' algoritem ter prilagojeno mero.

2. Algoritmi, ki pretvorijo časovno vrsto v vektor značilk (postopek izluščitve značilk - ang. *feature extraction*) ter uporabijo klasičen algoritem s klasično mero razdalje.
3. Algoritmi, ki kot vektor značilk uporabijo identificirane parametre modela (ang. *model-based approach*).

V [1] avtor predlaga še eno skupino metod za rojenje časovnih vrst:

Več-koračne metode (ang. *multi-step methods*), kjer se časovna vrsta predstavi z več resolucijami, pri vsakem koraku (resoluciji), se izvede rojenje. V [1] je podan pregled in kritična ocena predlaganih večnivojskih metod.

6 Zaključek

V prispevku so predstavljene ključne metode v uporabi na področju analize časovnih vrst. Metode so komentirane glede na primernost za delo s časovnimi vrstami. Obdelane so tri skupine metod: metode za predstavitev časovnih vrst, metode za merjenje razdalj med njimi ter metode za njihovo rojenje.

Povzeto je, zakaj je posebna predstavitev časovnih vrst sploh potrebna. Predstavljenih je osem metod, od najenostajvnejše - z vzorčenjem časovne vrste, do naprednejših algoritmov PLA, PIP in SAX.

V poglavju o merah razdalje je pojasnjena razlika v pojmovanju razdalje med dvema časovnima vrstama in običajnim pojmovanjem razdalje med vektorjema vzorcev, kot ga poznamo na drugih področjih. Predstavljenih je več metod, med njimi pogosto uporabljena DTW.

Področje rojenja časovnih vrst je obravnavano skozi razdelitev na šest skupin metod. Za vsako skupino so opisane glavne značilnosti, uporabnost ter razširjenost pri delu s časovnimi vrstami. Omenjene so posebnosti rojenja časovnih vrst ter omenjeni posebni skupini več-koračnih metod ter samorazvijajočih metod, ki obetata dobre rezultate na tem področju.

Prispevek nudi vpogled v področje ter služi kot osnova za nadaljnje delo na področju. V nadaljevanju avtorji nameravamo tukaj predstavljene metode uporabiti na praktičnem primeru nadzora delovanja dinamičnih procesov.

Zahvala Delo je bilo izvedeno v sklopu programa GOSTOP, ki ga delno financirata Republika Slovenija - Ministrstvo za izobraževanje, znanost in šport ter Evropska Unija - Evropski sklad za regionalni razvoj in v sklopu nacionalnega raziskovalnega programa Sistemi in vodenje, P2-0001.

Literatura

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[2] Goran Andonovski, Bruno Sielly Jales Costa, Sašo Blažič, and Igor Škrjanc. Robust evolving controller for simulated surge tank and for real two-tank plant. *at-Automatisierungstechnik*, 66(9):725–734, 2018.

[3] Goran Andonovski, Gašper Mušič, Saso Blažič, and Igor Škrjanc. On-line evolving cloud-based model identification for production control. *IFAC-PapersOnLine*, 49(5):79–84, 2016.

[4] Goran Andonovski, Gašper Mušič, Sašo Blažič, and Igor Škrjanc. Evolving model identification for process monitoring and prediction of non-linear systems. *Engineering applications of artificial intelligence*, 68:214–221, 2018.

[5] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[6] Fu-Lai Chung, Tak C Fu, Robert Luk, and V Ng. Flexible time series pattern matching based on perceptually important points. 2001.

[7] D. R. Edla, D. Tripathi, V. Kuppili, and R. Cheruku. Survey on clustering techniques. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 696–703, 2018.

[8] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[9] Tak-chung Fu, Fu-lai Chung, Robert Luk, and Chak-man Ng. Representing financial time series based on data point importance. *Engineering Applications of Artificial Intelligence*, 21(2):277–300, 2008.

[10] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[11] Octavian Lucian Hasna and Rodica Potolea. Time series—a taxonomy based survey. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 231–238. IEEE, 2017.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning, Second Edition*, volume 1. Springer series in statistics New York, 2009.

[13] Eamonn Keogh. Fast similarity search in the presence of longitudinal scaling in time series databases. In *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, pages 578–584. IEEE, 1997.

[14] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286, 2001.

[15] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[16] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.

[17] Igor Škrjanc, Goran Andonovski, Agapito Ledezma, Oscar Sipele, Jose Antonio Iglesias, and Araceli Sanchis. Evolving cloud-based system for the recognition of drivers' actions. *Expert Systems with Applications*, 99:231–238, 2018.

[18] Igor Škrjanc, Jose Antonio Iglesias, Araceli Sanchis, Daniel Leite, Edwin Lughofer, and Fernando Gomide. Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: a survey. *Information Sciences*, 490:344–368, 2019.

NAČRTOVANJE IN IZDELAVA PROTOTIPA SOLARNEGA SLEDILNIKA

[1]Domen Gošek, [1,2]Amor Chowdhury, [3]Karl Gotlih, [2]Riko Šafarič

[1]MARGENTO R&D d.o.o., Turnerjeva ulica 17, 2000 Maribor, Slovenija
[2]Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška c. 46, Maribor, Slovenija

[3]Univerza v Mariboru, Fakulteta za strojništvo, Smetanova 17, Maribor, Slovenija

E-pošta: domen.gosek@margento.com

Abstract. *The article presents the process of developing a solar tracker prototype mechanism on the machine, programming, and electrical level. Stages of development and solar tracking algorithms are described. The prototype is a physical system, which tracks sun trajectory by time reference with its algorithm. That means that for tracking it needs 4 inputs which are: number of days in a year, hours, minutes, and geographical latitude. After the development has ended collecting and analysis of result data were done. The results were given by experiments.*

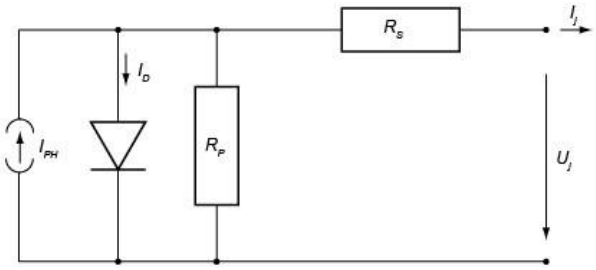
1 Uvod

Izkoriščanje sončne energije je ena najbolj ekološko sprejemljivih oblik pretvorbe energije. Pri sončni energiji uporabljamo energijo svetlobe, ki jo pretvorimo v toploto v primeru termično solarne tehnike ali jo s pomočjo fotovoltaike neposredno pretvorimo v električno energijo [1]. Eden bistvenih problemov pri izrabi energije sonca so relativno nizki transformacijski izkoristki, če k temu dodamo še spremenljiv položaj sevanja na solarni panel, so izkoristki še slabši. Pridobljeno energijo iz sončnih panelov je po nekaterih ocenah mogoče povečati za 20 do 30 odstotkov z uporabo sledilnega sistema namesto fiksne namestitve [2]. Pri razvoju solarnih sledilnikov so se razvile različne metode sledenja soncu. Najbolj priljubljeni so astronomski ali časovni sistemi, poleg njih pa poznamo tudi optične sledilne metode, ki za pozicioniranje sončnega panela uporabljajo različne vrste foto-senzorjev. Cilj našega dela je izdelava cenovno sprejemljivega prototipa solarnega sledilnika, s katerim želimo zagotoviti kvalitetno in zanesljivo sledenje soncu na osnovi koledarja, ure in geografske širine na Zemlji. Projekt obsega zasnovo in trdnostno načrtovanje, izbiro komponent, sestavo, programiranje in testiranje delovanja. S testiranjem prototipa želimo preveriti ekonomičnost delovanja ter povzeti, kaj bi bilo potrebno izboljšati pri nadaljnjem razvoju.

2 Načrtovanje solarnega sledilnega sistema

Pri načrtovanju solarnega sistema je potrebno upoštevati učinkovitost sončnih celic. Učinkovitost sončnih celic je omejena s številnimi faktorji. Tako mora biti energija fotonov zadostna za kreiranje parov vrzel-elektron. Sevanje z večjo valovno dolžino nad 1,15 μm pri silicijevih fotocelicah povzroči le segrevanje celice. Ker lahko en foton kreira le en par vrzel-elektron, se tudi pri valovnih dolžinah manjših od mejne pojavlja višek fotonov, ki prav tako povzročajo le segrevanje celice, ne pa proizvodnjo električnega toka. Tako je zgornja meja pretvorbe sončnega sevanja s silicijevimi celicami približno 23% [3]. Z uporabo večslojnih celic z različnimi materiali je možno pretvoriti širši spekter sončnega sevanja. Take celice dosegajo v eksperimentalni fazi učinkovitosti večje od 30 %. Dodatne izgube predstavlja še zastrtost *n* in *p* sloja zaradi kontaktne mreže na zgornji strani sončne celice, notranja upornost celice in refleksija sončnega sevanja na površini sončne celice.

Karakteristiko sončne celice lahko primerjamo s karakteristiko diode, saj sta tok in napetost odvisna od energije sončnega sevanja in bremena. Na sliki 1 je prikazana poenostavljena električna shema solarnega modula.



Slika 1: Ekvivalentno vezje fotocelice

Solarni modul si lahko predstavljamo kot črno škatlo z dvema sponkama, ki proizvaja tok pri napetosti U. Ekvivalentno vezje se sestoji iz štirih komponent:

1. Izvor foto-toka izrazimo z enačbo:

$$I_{ph} = A_{Cell} H \xi \quad (1)$$

kjer je A_{Cell} površina celice, H intenziteta svetlobe in ξ odzivni koeficient v enotah A/W.

2. Dioda: nelinearni element, ki izraža odvisnost od širine zaporne plasti in izgub z rekombinacijami. Opredeljen je z reverznim tokom, ki odraža prepuščanje elektrine in

rekombinacijo s faktorjem kvalitete q z vrednostmi od 1 do 2. Je empirični faktor.

3. Shunt-ov upor R_p: predstavlja izgube v polprevodniku.
4. Zaporedni upor R_s: prav tako predstavlja izgube neidealnega prevodnika.

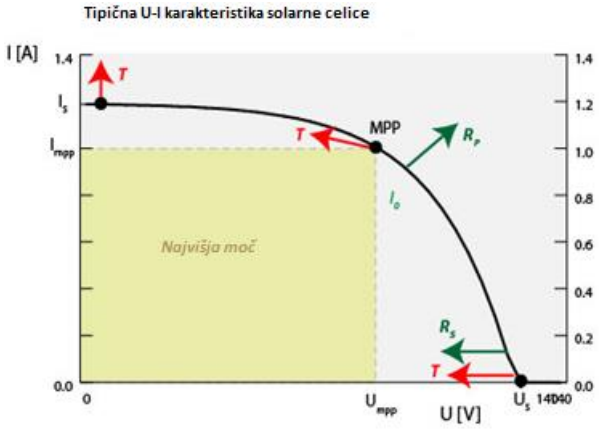
Odvisnost med tokom I in napetostjo U posamezne celice predstavimo z enačbo:

$$I_{Cell} = I_{Ph} - I_0\left[\exp\left(\frac{U_{Cell} + I_{Cell}R_s}{U_T}\right) - 1\right] - \frac{U_{Cell} + I_{Cell}R_s}{R_p} \quad (2)$$

kjer je U_T napetost zaradi toplote

$$U_T = \frac{qkT}{e} \quad (3)$$

Temperatura je izražena v Kelvinih, k je Boltzmannova konstanta in e je elementarni naboj. Z uporabo izraza 3 lahko izračunamo točke maksimalne moči pri različnih temperaturnih vplivih. V modulu je zaporedno povezanih mnogo celic, ki prispevajo k toku in napetosti. Tako moramo pri obravnavi modula upoštevati še število celic.

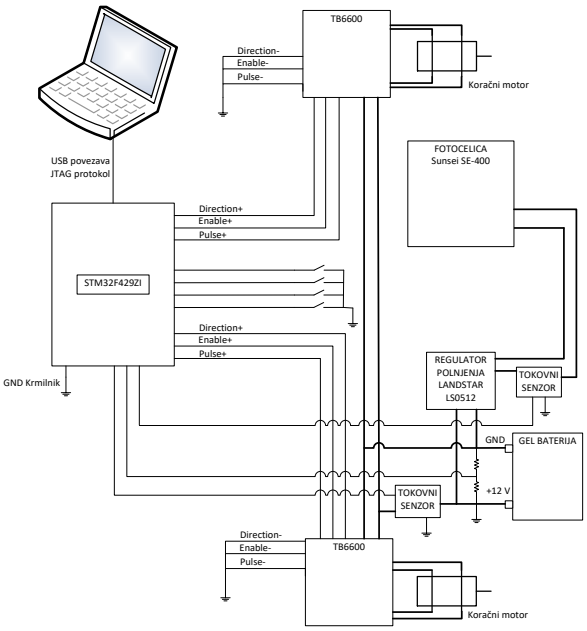


Slika 2: U-I karakteristika foto celice

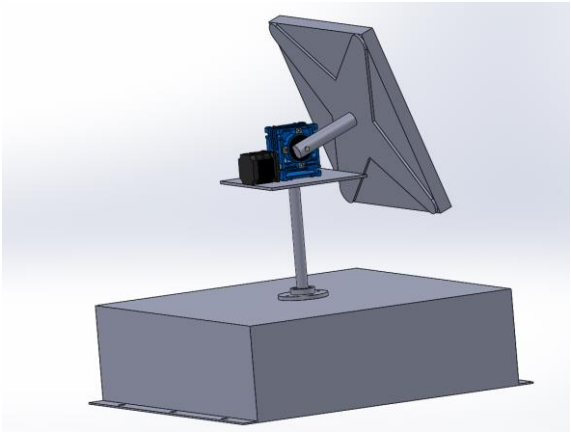
Karakteristika prikazana na sliki 2 je precej občutljiva na spremembe v temperaturi, kar se odraža v nižji napetosti odprtih sponk in višjemu kratkostičnemu toku. Končni rezultat je prehod točke MPP na nižjo vrednost. Nasičenje reverznega toka se odraža v položnejši karakteristiki. Višja zaporedna upornost se odraža v večjih izgubah in manjši končni napetosti na bremenu [3].

3 Pogonske komponente

Struktura sistema je prikazana na sliki 3, vizualni izgled pa na sliki 4. Struktura se deli na del z močnostnimi komponentami in del s signalnimi komponentami. Signalni del se ločuje od močnostnega v gonilnikih, senzorjih in delilniku.



Slika 3: Blokovna shema sistem



Slika 4: Fizični izgled prototipa

V načrtovanem solarnem sledilniku so uporabljeni bipolarni koračni motorji Nema 23 z 1,8° korakom. Motor proizvede 2,83 Nm navora. Masa motorja je 1,3 kg. Dolžina gredi je 21 mm s premerom 6,35 mm. 2 motorja uporabljena v sledilniku sta odgovorna za izvedbo pozicioniranja nosilca panela glede na lego sonca. Napajamo ju s pomočjo baterije preko gonilnikov. Izbrani koračni motorji imajo možnost razširitve z dodatno gredjo, enkoderji, zavorami, reduktorji in multiplikatorji.



Slika 5: Koračni motor Nema 23

Mehanski prenos je zasnovan iz polža, ki oddaja moč, in iz zobnika, ki moč sprejema ter jo oddaja na izhodno gred. Je bistveno manjši od navadnega zobniškega prenosa. Njuni vrtilni osi sta zamaknjeni za 90°.

Razlog v samozapornosti je v naklonu zob zobnika na polž. Ta je tako majhen, da med apliciranjem navora na zobnik, trenje drži polž na mestu, kar je bil bistven razlog za izbiro polžastih gonil v sledilniku.



Slika 6: Polžasto gonilo

4 Sledilni algoritem

Zaradi rotacije in revolucije Zemlje se Sonce na nebu premika po svoji tirnici vsak dan v letu drugače. S pomočjo poznavanja gibanja Zemlje in njene rotacije okoli Sonca in lastne osi, je mogoče izpeljati enačbo, s katero izračunamo azimut in višinski kot Sonca.

Vemo, da glede na specifično lokacijo na Zemlji Sonce vzide na vzhodu in zaide na zahodu. Solarno sledilni mehanizem uporablja informacijo o položaju Sonca na tej lokaciji z namenom nenehnega usmerjanja panela proti Soncu. Shema nebesne oble predstavlja položaj Sonca izražen z azimutom in višinskim kotom glede na geografsko širino.

Algoritem, ki podaja informacijo o položaju Sonca v dveh podatkih, to sta azimut in višinski kot, v odvisnosti od položaja na zemlji, datuma in ure, se imenuje solarni pozicionirni algoritem (ang. Solar position algorithm - SPA). Pri izračunu moramo upoštevati naslednje podatke [4]:

1. geografska širina φ – kot severno ali južno od ekvatorja
2. deklinacija δ – lok časovne krožnice od nebesnega ekvatorja do lege nebesnega telesa na časovni krožnici
3. dan v letu
4. ura – pretvorba v decimalno vrednost na primer 15,30 = 15,5
5. urin kot (t)

Za izračun sončnega vektorja so uporabljene naslednje enačbe:

$$\delta = -23{,}45 \cdot \cos\left(\left(\frac{360}{365}\right) \cdot dan_v_letu + 10\right) \quad (4)$$

$$t = 15 \cdot (Ura - 12) \quad (5)$$

$$Višinski\ kot = arcsin(sin(\delta) \cdot \sin(\varphi) + \cos(\delta) \cdot \cos(t)) \quad (6)$$

Azimut nato izračunamo, če je ura manj ali več kot 12. Če ura tisti dan še ni 12, se azimut izračuna po naslednji enačbi:

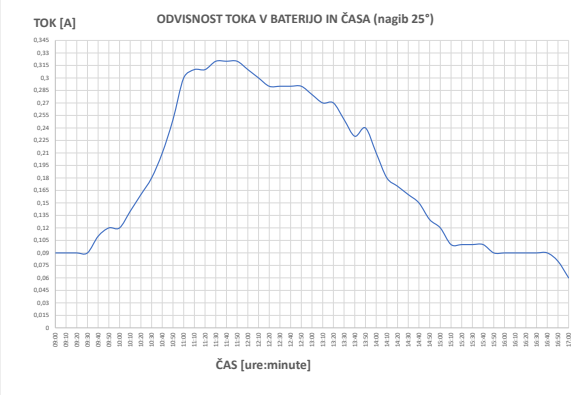
$$Azimut = arccos\left(\frac{sin(\delta) \cdot cos(\varphi) - cos(\delta) \cdot sin(\varphi)}{\sin(90 - Višinski\ kot)}\right) \quad (7)$$

V nasprotnem primeru, ko je ura več ali enaka kot 12, pa se azimut Sonca izračuna kot

$$Azimut = 360 - arccos\left(\frac{sin(\delta) \cdot cos(\varphi) - cos(\delta) \cdot sin(\varphi)}{\sin(90 - Višinski\ kot)}\right) \quad (8)$$

5 Eksperimentalni rezultati

Za korektno primerjavo rezultatov med fiksno nameščenim PV modulom in PV modulom s solarnim sledilnikom bi morali izvesti sočasne meritve ob enakih vremenskih pogojih. Žal zaradi določenih izvedbenih omejitev tega nismo izvedli in so meritve bile opravljene zgolj na prototipu v dveh režimih delovanja ob čim bolj identičnih vremenskih pogojih. Med izvajanjem meritev so bile vremenske razmere sončne z občasno oblačnostjo. Meritev s fiksno nameščenim PV modulom je potekala z namestitvijo sistema v točko, kjer je bilo Sonce vidno čez cel dan. Geografska širina, kjer je bila meritev izvajana je bila 46,55 stopinj. Nagib sončne celice je bil optimalen v zimskih mesecih in znaša 25 stopinj glede na horizont.



Slika 3: Polnilni tok *I* pri fiksno nameščenem PV

Tok v dopoldanskem delu strmo narašča in se približuje najvišji vrednosti, ki znaša okrog 300 mA. Po dosegu vrha opazimo nekoliko položnejše padanje toka I, ki se v zadnjem delu grafa še upočasni.

Skozi opazovani časovni interval lahko izračunamo pridobljeno energijo. Ta je enaka seštevku zmnožkov napetosti baterije *U*, toka I_{vbat} in časovnega intervala t_{med}. Energijo *E*, pridobljeno iz celice splošno zapišemo po naslednji enačbi:

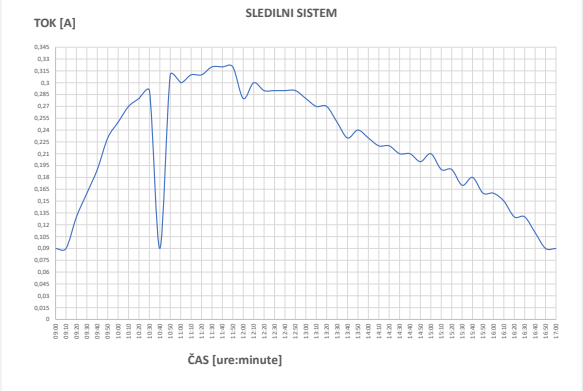
$$E = \sum UI_{vbat}t_{med} \qquad (9)$$

Število seštevkov je enako številu meritev. V danem primeru je bilo opravljenih 49 meritev, izračun posamezne meritve je;

$$E = 13{,}86\ V * 0{,}32\ A * 600\ s$$
$$E = 2661\ Ws = 0{,}74\ Wh$$

Kumulativni rezultat meritev pri mirujočem sistemu pridobljene energije pa znaša 20,40 Wh.

Pri meritvi s sledilnim sistemom smo zaradi objektivnosti primerjave opravili meritve v istem tednu kot za mirujoč sistem ob približno podobnih vremenskih pogojih. Okoliščine, ki so prisostvovale pri meritvah, so opredeljene z opisom vremenskih razmer. Med izvajanjem meritev so bile vremenske razmere sončne z občasno oblačnostjo. Meritev je potekala z namestitvijo sistema v točko, kjer je bilo Sonce vidno čez cel dan. Geografska širina, kjer je bila meritev izvajana je bila 46,55 stopinj.

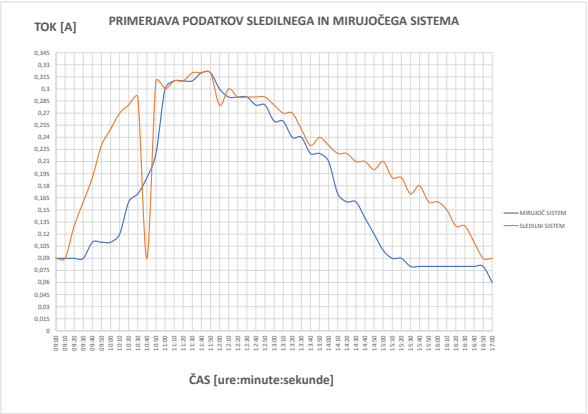


Slika 4: Polnilni tok *I* pri sledilno nameščenem PV

Iz grafa na sliki 8 je razvidno, da merjeni tok *I*, narašča hitreje in pada počasneje. Na grafu opazimo upad merjenega toka *I* med 10:40 in 10:50 uro, kar je posledica mimoidočega oblaka. Kumulativni izračun pridobljene energije *E* pokaže, da je pri sledilnem sistemu pridobljena skupna energija 24,22 Wh.

Pri izvajanju sledenja so energijo porabljali gonilniki motorjev, koračna motorja in mikrokrmilniška enota. Za poravnavo proti soncu je sistem porabil 0,5 s, saj je šlo za majhne premike, sistem pa je bil optimiziran tako, da če so bili potrebni premiki manjši kot 0,05 stopinje, je sistem počakal na naslednji interval. Tako se sistem ni zaganjal po nepotrebnem. Med čakanjem na premik je najmanjšo možno energijo v režimu nizke porabe porabljala krmilniška enota in gonilniki v stanju pripravljenosti.

Sistem je za 49 premikov pri toku 1 A in napetosti 13,6 V pri času 0,5 sekunde porabil 0,0925 Wh, kar je 0,4 % pridobljene energije pri sledilnem sistemu.



Slika 5: Primerjalni graf mirujočega in sledilnega sistema

Kljub pogojno »objektivnim« meritvam, lahko iz dobljenih eksperimentalnih rezultatov pridemo do zaključka, da ob uporabi solarnega sledilnika pridobimo za dobrih 18% več energije v primerjavi s fiksno nameščenim oz. mirujočim PV sistemom.

Na sliki 9 je vidno območje med krivuljama, ki predstavlja dobitek energije zaradi sledenja Soncu.

6 Zaključek

Eksperimentalni projekt izvedbe prototipa solarnega sledilnika je pokazal, da so sledilno solarni sistemi precej učinkovitejši od mirujočih, slaba stran je njihova cena, ki kljub bistvenemu dvigu učinkovitosti vsaj pri manjših solarnih sistemih ne opravičuje dodatnega vložka. Kljub večjemu strošku pa lahko pričakujemo njihovo uporabo predvsem v primerih, ko smo omejeni s prostorom namestitev PV kot na primer med taborjenjem v naravi, kamp prikolicah, v odročnejših krajih za napajanje merilnih sistemov ipd.

Literatura

[1] S. Medved, Solarni inženiring, Ljubljana: Fakulteta za strojništvo, 1993.

[2] S. A. Kalogirou, Solar energy engineering processes and systems, druga izdaja, Amsterdam: Elsevier, 2014.

[3] Theory of solar cells [splet], Dosegljivo: https://en.wikipedia.org/wiki/Theory_of_solar_cells [Datum dostopa 6. 5. 2019].

[4] G. Prinsloo, R. Dobson, Solar Tracking, Stellenbosch, 2015

Sistem za zajemanje, obdelavo in obveščanje uporabnikov o kritičnih situacijah v njihovi okolici

Dalibor Igrec, Uroš Sadek, Amor Chowdhury

MARGENTO R&D d.o.o., Turnerjeva ulica 17, 2000 Maribor, Slovenija
E-pošta: dalibor.igrec@margento.com

System for capturing, processing and informing users about critical situations in their surroundings

Abstract. *Visual and audio stimuli are often used in the transmission of information in human-computer interaction. These interactions can represent more important digital information from the environment, such as information from sensors or information in a public information system. Normally, information transmitted as an audible signal cannot be passed on to all people, especially people with sensory disabilities and the elderly. Given all the above, the presented solution, in addition to visual and audible stimuli, also enables a vibration display of the transmitted information.*

1 Uvod

Vizualni in zvočni dražljaji se pogosto uporabljajo pri prenosu informacij v interakciji človek-računalnik.

Te interakcije lahko predstavljajo pomembnejše digitalne informacije iz okolja, kot so informacije iz senzorjev ali informacije v javnem informacijskem sistemu (železniške, letalske, avtobusne postaje). Še posebej pomembne so interakcije v življenjsko-kritičnih situacijah, kot so požari, poplave ali potresi. Uporaba vibracij kot dodatnih dražljajev skupaj z vizualnimi in zvočnimi oblikami prenosa informacij postane rešitev, ki omogoča dodatne oblike zaznavanja zaradi manj moteče in hitrejše pritegnitve pozornosti.

Obstoječe rešitve na področju sporočanja informacij iz senzorjev uporabnikom ponujajo različna podjetja, a so te zaprte narave in običajno ne dopuščajo priključitev senzorjev drugih proizvajalcev. Na področju javnega informacijskega sistema po dostopnih podatkih ni enotnega standarda, ki bi združeval vse danes uveljavljene tehnologije.

Prav tako je iz obstoječih rešitev možno razbrati, da so te na trgu namenjene predvsem splošni rabi, ki izloča starejšo populacijo in uporabnike s posebnimi potrebami, ki so jih trenutne rešitve na trgu zapostavile.

Iz teh razlogov smo v projektu izkoristili obstoječi potencial s ciljem implementacije inovativne rešitve za predstavitev informacij iz okolja s pomočjo vizualnega, zvočnega in vibracijskega signala, ki bo koristil prav vsem uporabnikom, še posebej tistim s posebnimi potrebami ter starejšim.

Projekt je delno financirala Evropska unija iz Evropskega sklada za regionalni razvoj (ESRR) v okviru Javnega razpisa za »Spodbude za raziskovalno razvojne projekte 2«, ki ga je razpisalo Ministrstvo za gospodarski razvoj in tehnologijo [1].

2 Opis problema in motivacija

Glede na obstoječe študije v času trajanja projekta smo ugotovili, da ne obstaja celovit sistem, ki bi bil vgrajen v prostorih bivanja, na sredstvih javnega transporta, avtobusnih in železniških postajah, v bolnišnicah, domovih starostnikov, v reševalnih vozilih in podobnih objektih oz. ustanovah in zajemal vse tri elemente: vizualne, zvočne in vibracijske komunikacije.

Po raziskavah obstaja problem obveščanja gluhih, slepih, gluhoslepih in starejših ljudi, ki je ena izmed žarečih problematik v svetu. Različne študije kažejo, da obstaja veliko število ljudi z vsaj eno ali več oblik invalidnosti: gluhost, slepota, gluhoslepota. Ugotavljajo, da ima več kot 5% svetovne populacije težave z izgubo sluha. Poslabšanje sluha nastopi takrat, ko je izguba sluha večja od 40 dB pri odraslih in večja od 30 dB pri otrocih. Po podatkih Svetovne zdravstvene organizacije po svetu živi 285 milijonov slepih in slabovidnih, od tega je 39 milijonov slepih in 246 milijonov slabovidnih. Okrog 80% ljudi, ki se soočajo s težavami vida, je starejših od 50 let. Ocenjuje se, da je v Evropi skoraj tri milijone ljudi, ki imajo težave tako z vidom, kot tudi s sluhom [2, 3].

V digitalni dobi, v kateri živimo, ljudje z okvaro ali izgubo sluha in/ali vida, kot tudi starejši, težko zaznavajo in se pravilno odzivajo na informacije iz okolja. Eden od možnih razlogov je v sistemih informiranja, ki temeljijo predvsem na zvočni in vizualni komunikaciji. Vse to lahko privede do življenjsko-kritičnih situacij za gluhe, slepe ali starejše ljudi.

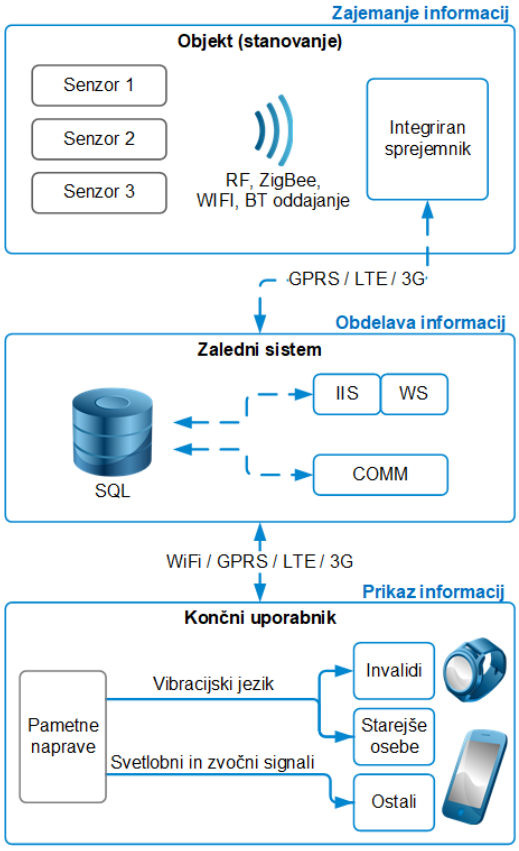
Na primer, gluhe osebe lahko izgubijo svoje življenje, če ne slišijo požarnih alarmov v svojih hišah; slepi, diabetiki se lahko znajdejo v nevarni situaciji, ker ne morejo prebrati številk o stanju njihovega insulina; zaradi hitre rasti zunanje temperature zraka in pomanjkanje vode, lahko starejši trpijo zaradi dehidracije; gluhi ter slepi potniki so pogosto zmedeni na vlaku ali na postaji zaradi nerazumljivega zvočnega obvestila. Omenjene in še veliko podobnih, nevarnih situacij, v katerih se ljudje znajdejo, se zgodi zaradi tega, ker ljudje nimajo možnosti pridobiti ustrezno prilagojene informacije v kritičnih trenutkih.

Prav tako smo iz študij ugotovili, da je trenutno uporaba vibracijskih dražljajev omejena le na osnovne vzorce vibracij. Iz teh razlogov smo v projektu izkoristili obstoječi potencial s ciljem, da bi predstavili

inovativne rešitve za zajem, obdelavo, prenos in predstavitev informacij iz okolja do končnega uporabnika; še posebej za ljudi s posebnimi potrebami in starejšimi.

Kot možno rešitev smo v projektu raziskali možnosti uvedbe vibracijskega jezika s pomočjo obstoječih vibracijskih naprav (pametna ura, mobilni telefon, pametna zapestnica in podobne naprave), ki bi oddajale vibracije v obliki jezika, dovolj preproste in razumljive v dani situaciji. Na ta način lahko uporabnik lažje prepozna vsebino informacij iz senzorjev ali iz javnih informacijskih sistemov. Vibracije tako postanejo osnova za sporočanje (npr. z uporabo enega od obstoječih načinov predstavljenim v [4]).

Tehnična rešitev za prenos informacij vsebuje brezžične okoliške senzorje in sisteme za obveščanje, ki so povezani v oblak, slednji pa je povezan z različnimi napravami, opremljenimi s primernimi brezžičnimi sprejemniki (slika 1).



Slika 1: Prenos informacij med različnimi segmenti sistema

Glede na prikazan prenos informacij je krog sklenjen:

- zgodi se nepredviden dogodek (požar, posebno obvestilo …),
- senzor v bližini zazna dogodek,
- zaznan dogodek pošlje do najbližje komunikacijske naprave (koncentrator, vozlišče …),
- ta preko mobilnega omrežja ali direktne internetne povezave pošlje informacijo do strežniškega sistema,
- strežnik pošlje prejeto informacijo pametnim napravam vsem uporabnikom, ki so na ta tip informacije prijavljeni,
- pametna naprava s pomočjo mobilne aplikacije prejeto informacijo prikaže na zaslonu ter preko vzpostavljene Bluetooth povezave pošlje podatek napravi, kot so pametne ure in zapestnice, ki z določenim načinom (jezikom) vibriranja osebi sporočijo dejansko informacijo alarma in ne le alarmno stanje kot takšno.

Uporabnik, ki je zadnji člen v verigi, na tak način prejme informacije v obliki vibracijskega jezika, ki vsebujejo več različnih vzorcev vibracij. Vsak vzorec ima lasten pomen in vlogo, tako uporabnik takoj razume njegov bistvo posredovane informacije. V kolikor je vibracijska naprava opremljena z zaslonom, je možno izbirati tudi med tekstovnim in/ali grafičnim prikazom vibracij.

3 Komponente sistema

Koncept sistema je inovativen in večnamensko uporaben, saj vključuje raziskovalne možnosti za zajemanje, obdelavo in prenos informacij iz senzorjev ter informacijskih sistemov iz javnih domen, kot so letališča, železniške in avtobusne postaje, informacij javnega značaja, alarmov, požarov ter informacij pridobljenih iz naprav lastnega doma.

Informacije so posredovane do dveh ciljnih skupin:

- splošni uporabniki, katerim so informacije prikazane na mobilnih napravah in
- uporabniki starejše populacije ter tistim s posebnimi potrebami (slepi, gluhi), katerim so informacije posredovane v obliki vibracijskih signalov, kot na primer vibracijskih zapestnih napravah (pametne ure, pametne zapestnice ipd.).

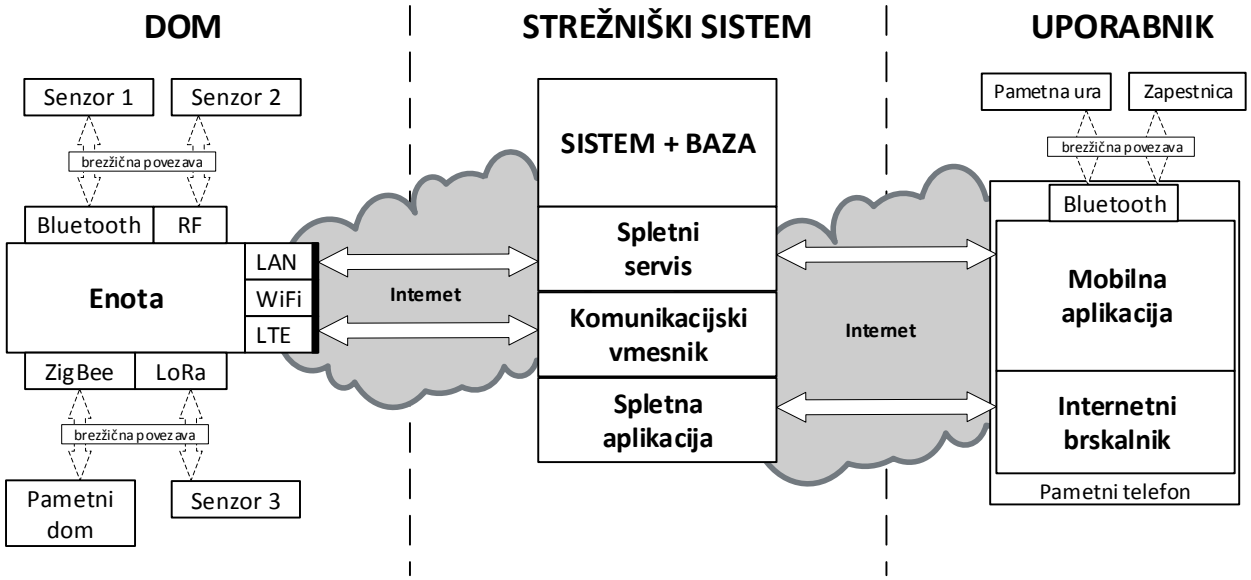
Rešitev vsebuje poleg mobilne aplikacije, ki prejme naročene informacije, tudi končni element v obliki pametne zapestne naprave, ki omogoča sporočanje informacije v svetlobni, zvočni in/ali vibracijski obliki.

V centralnem strežniku se zajete informacije ovrednotijo, obdelajo in posredujejo končnim uporabnikom, ki so naročeni/registrirani za prejemanje določenega tipa informacij (požar, vlom, povečanje ogljikovega monoksida, zamuda avtobusnega prihoda, itd.) tudi glede na njihovo trenutno geolokacijo.

Centralna enota sistema, navedena kot Enota v nadaljevanju, predstavlja multifunkcijsko napravo, ki je nameščena na uporabnikovem domu in predstavlja del rešitve celotnega sistema. Uporablja se kot lokalno vozlišče, »zvezdišče« pametnega doma, ki s pomočjo internetne povezave, prenosom podatkov ter različnih tipov komunikacijskih brezžičnih vmesnikov omogoča učinkovito zajemanje, obdelavo in posredovanje

informacij različnih tipov senzorjev in naprav preko Enote sistema do končnega uporabnika na način, ki v celoti premosti morebitne ovire v uporabnikovem zaznavanju.

Pridobljen podatek na Enoti iz nabora dodeljenih senzorjev je sprejet, obdelan in poslan preko mobilnega omrežja ali lokalne internetne povezave v za to posebej pripravljen strežniški sistem. Ta nato glede na prijavljene uporabnike podatek ovrednoti, obdela in distribuira do njihove pametne naprave (pametni telefon, pametna tablica …), ki s pomočjo vzpostavljene brezžične povezave (npr. Bluetooth) z na primer vibracijsko napravo opozori na posredno nevarnost s pomočjo svetlobnega, zvočnega ali vibracijskega signala.



Slika 2 (zgoraj) prikazuje komponente celotnega sistema:

DOM:
- Enota montirana na domu uporabnika,
- priključena v internet omrežje preko LAN, WiFi ali LTE povezave (odvisno, kaj ima doma uporabnik na voljo),
- nanjo so preko brezžičnih vmesnikov (ZigBee, RF, Bluetooth, LoRa, WiFi) povezane različne (že obstoječe) naprave drugih ponudnikov (senzor požara, javljalnika dima, vloma, odpiranja vrat/okna, potresa, poplave …).

STREŽNIŠKI SISTEM:
- predstavlja celoten zaledni sistem,
- nanj so na eni strani stalno povezane Enote različnih uporabnikov, na drugi strani pa so ti isti uporabniki nanj povezani preko mobilne ali spletne aplikacije,
- ključne komponente:
  - podatkovna baza (data base),
  - spletni servis (web service), preko katerega je omogočen dostop do podatkov tako na strani Enote, kot tudi iz mobilne aplikacije uporabnika,
  - komunikacijski vmesnik (communication interface), preko katerega so Enote stalno povezave in pošiljajo/prejemajo podatke v realnem času,
  - spletna aplikacija (web application), preko katere se izvaja osnovno konfiguriranje iz strani upravitelja in uporabnika sistema.

UPORABNIK:
- uporablja mobilno aplikacijo, delujočo na pametni mobilni napravi uporabnika (iOS in Android),
- mobilna aplikacija preko spletnega servisa pridobiva podatke iz zalednega sistema v realnem času (s pomočjo Web API klicev na eni strani in push-notification storitve na drugi strani),
- mobilna aplikacija, ki je preko Bluetooth vmesnika pametne mobilne naprave povezana s kompatibilnimi Bluetooth napravami (pametna ura, zapestnica, obesek …), preko katerih uporabniku s pomočjo vibracijskih, svetlobnih ali zvočnih efektov sporoča različne informacije glede na sam izvor dogodka.

4 Razvoj sistema

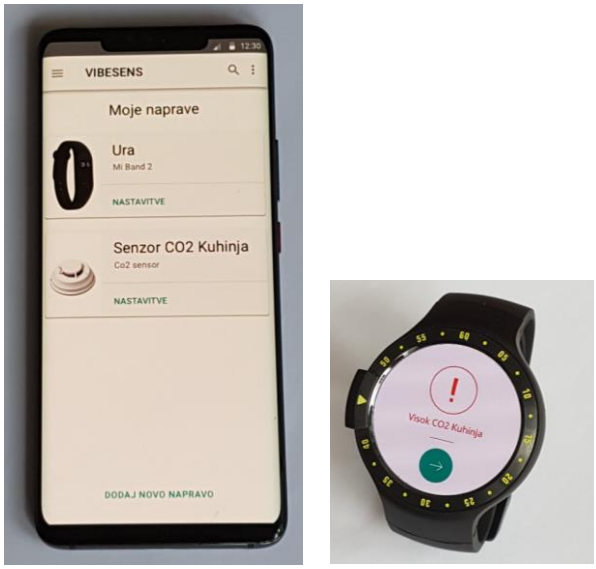
Razvoj je zajemal tri ločene segmente:
- razvoj strojne opreme, kjer se je izvedla nadgradnja obstoječega elektronskega vezja Enote, ki omogoča povezavo s senzorji preko različnih komunikacijskih vmesnikov (Bluetooth, WiFi, ZigBee, RF in LoRa) in zalednim sistemom (LAN, WiFi ali LTE),
- razvoj zalednega sistema, ki je zajemal načrtovanje in vzpostavitev razširjene arhitekture podatkovne baze, razvoj spletne aplikacije za upravljanje s sistemom in razvoj spletnih vmesnikov (za komunikacijo z Enoto na eni strani in mobilnimi aplikacijami na drugi strani),
- razvoj mobilne aplikacije za iOS in Android platformo (po smernicah v [5]), kjer so se implementirali komunikacijski vmesniki med mobilno aplikacijo in zalednim sistemom, implementiral razširjen uporabniški grafični vmesnik, implementiral enotni vmesnik za povezovanje in komunikacijo z novimi napravami

(preko Bluetooth povezane pametne ure, zapestnice ali obesek),

- razvoj mobilne aplikacije na pametni uri za Google Wear OS [6], kjer se je implementiral uporabniški vmesnik za prikaz sporočila v slikovni in tekstovni obliki ter proženje alarma v zvočni in vibracijski obliki. Razvoj aplikacije v Google Wear OS omogoča dostop do vseh potrebnih funkcij za namene obveščanja uporabnika na vizualni, zvočni in vibracijski način.

5 Produkt

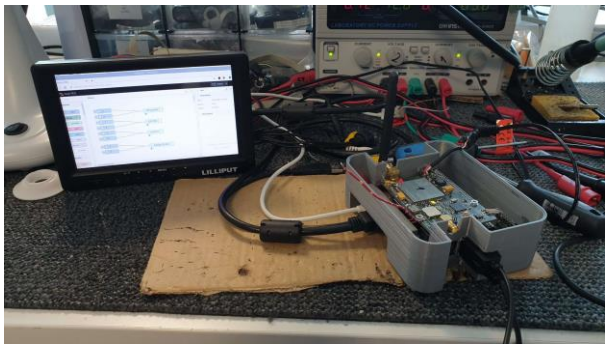
V nadaljevanju je prikazanih nekaj slik iz časa razvoja tiskanine Enote, ohišja in mobilne aplikacije na pametni mobilni napravi ter pametni uri.



Slika 3: Mobilna aplikacija na pametni mobilni napravi (levo) in na pametni uri (desno)



Slika 4: Tiskanina Enote umeščena v prvi prototip 3D tiskanega ohišja



Slika 5: Enota v času razvoja

6 Zaključek

Običajno informacije, ki se prenašajo kot zvočni signal, ne morejo biti zaznane oz. posredovane vsem ljudem, še posebej osebam s senzorno invalidnostjo in starejšim. Zato predstavljena rešitev poleg vizualnega in zvočnega sporočanja omogoča tudi vibracijski prikaz posredovane informacije.

Predstavljena rešitev omogoča uporabo obstoječih naprav na strani zajemanja informacij (senzorji in obstoječi informacijski sistemi), kot tudi na strani obveščanja (mobilna naprava, pametna zapestnica, pametna ura). S tem omogoča nadaljnjo uporabo obstoječe senzorske opreme in uporabnikovih pametnih naprav.

Naslednji korak predstavlja razširitev sistema za potrebe invalidov v javnem prevozu. Invalidi se skupaj s starejšimi ranljivimi skupinami v javnem prevozu soočajo s težavami in ovirami, ki jih tudi običajne osebe težko rešujejo. Zahteve in potrebe teh skupin so pogosto prezrte, sistem znotraj katerega živimo pa še vedno ni v celoti prilagojen in primerno oblikovan v smeri pomoči in oblikovanja prevoznih sredstev do te mere, da bi bili primerno prilagojeni za vse starostne skupine [7]. Vse navedeno je osnovna motivacija za nadaljevanje projekta.

Literatura

[1] Ministrstvo za gospodarski razvoj in tehnologijo: Javni razpis za »Spodbude za raziskovalno razvojne projekte 2«

[2] EBU Central Office: About blindness and partial sight

[3] Hear-it.org: Europe - Hearing loss

[4] Mozilla and individual contributors: Vibration API

[5] Office of the Government Chief Information Officer: Mobile Application Accessibility Handbook

[6] Google: Wear OS

[7] Inclusive City Maker: Making Public Transport Information Accessible to Disabled People

Spletna aplikacija za dostop do medicinskih slik v standardu DICOM

Žan Regoršek[1], Samo Penič[1], Miha Fošnarič[2], Janez Žibert[2]

[1]*Fakuteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, SI-1000 Ljubljana, Slovenija*
[2]*Zdravstvena fakulteta, Univerza v Ljubljani, Zdravstvena pot 5, SI-1000 Ljubljana, Slovenija*
E-pošta: samo.penic@fe.uni-lj.si

Web application for accessing medical DICOM images

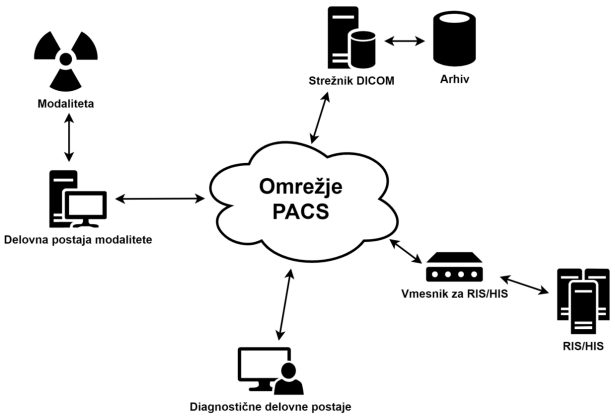
Medical imaging is today mostly digitized. There are many picture archiving and communication systems (PACS) available and are often integrated with hospital information system (HIS) and/or radiological information system (RIS). In practice, however, patients still receive their medical images as a printed hardcopy; on a CD, DVD or USB drive, where they are often saved in a DICOM (Digital Imaging and Communications in Medicine) format. This requires appropriate software and hardware to access them.

In this work we present a lightweight application for remote access to medical images from a DICOM server. It comes with user management and medical image preview and is built on open-source components. It was developed for educational and scientific purposes at the University of Ljubljana, and can be used as a basis for an affordable solution, suitable for medical facilities performing diagnostic imaging, for example small medical clinics with ultrasonography.

1 Uvod

Vse bolj prisotna digitalizacija omogoča tudi v medicinski slikovni diagnostiki posredovanje slik in drugih podatkov o preiskavah pacientom in strokovnemu osebju na hiter, učinkovit in varen način. Pri večini medicinskih slikovnih preiskav se podatki danes zajemajo v digitalni obliki in tudi zdravniki ter inženirji radiologije slike pogosto pregledujejo in obdelujejo na računalniku. V ta namen so na voljo mnogi informacijski sistemi za pošiljanje in arhiviranje slik (*angl.* Picture Archiving and Communication System – PACS). PACS je, predvsem v večjih zdravstvenih ustanovah, pogosto povezan z bolnišničnim informacijskim sistemom (*angl.* hospital information system – HIS) in/ali radiološkim informacijskim sistemom (RIS) (slika 1) [1, 2].

V praksi pacient še vedno večinoma dobi slike kar natisnjene v fizični obliki ali pa shranjene na zgoščenko ali USB ključek. Te so navadno shranjene v formatu standarda digitalnega slikanja in komunikacije v medicini (*angl.* Digital Imaging and Communications in Medicine – DICOM) [3]. Tako za pregled slik še vedno potrebujemo ustrezno strojno in programsko opremo, pri čemer smo praviloma omejeni na uporabo računalnika, saj na



Slika 1: Primer integracije sistema PACS v celostno rešitev, ki nudi zajem medicinskih diagnostičnih slik, njihovo hrambo in nadaljnjo obdelavo ter integracijo v druge informacijske sisteme, kot sta HIS in RIS [1].

mobilnih napravah primerne programske opreme za pregled slik nimamo. Ob tem se postavlja tudi vprašanje hrambe in varnosti podatkov ter obremenjevanja okolja. Nenazadnje prenosni mediji predstavljajo tudi dodaten strošek.

V prispevku predstavljamo preprosto spletno aplikacijo za dostop do medicinskih slik, ki so zbrane v centralnem strežniku v DICOM formatu. Aplikacija je s svojimi odprtokodnimi komponentami primerna za pedagoške in raziskovalne namene Univerze v Ljubljani in je osnova za razvoj cenovno dostopne storitve predvsem v manjših medicinskih diagnostičnih ustanovah, kot so na primer ambulante z ultrazvočno diagnostiko.

Prispevek je organiziran tako, da v naslednjem poglavju opišemo zasnovo aplikacije, njen uporabniški vmesnik in platformo na kateri je zgrajena, čemur sledi kratka razprava o razvoju aplikacije in zaključki.

2 Aplikacija

2.1 DICOM

Standard DICOM organizira medicinske slike po hierarhiji pacient → študija → serija → instanca. Aplikacija obravnava omenjene ravni po naslednji ideji:

- **Pacient**: unikatna identiteta pacienta;

- **Študija**: specifično področje/razlog zdravljenja (npr. zlom kosti, nosečnost, neznani glavoboli ipd.), ki poteka dlje časa in je skupek več individualnih pregledov – serij;
- **Serija**: posamezen pregled v ambulanti, ki je del postopka zdravljenja. Pregled lahko sicer razdelimo na dve seriji, v kolikor sta bili uporabljeni dve različni modalnosti. Tipično poimenovanje bi bilo npr.: 1. pregled celjenja kosti - rentgen , 16. 5. 2020; 2. pregled celjenja kosti - rentgen, 18. 6. 2020; 2. pregled celjenja kosti - MRI, 18. 6. 2020;
- **Instanca**: ob posameznem pregledu se zdravnik odloči posneti sliko iz več zornih kotov ali pa z različnimi nastavitvami slikovne aparature. Vsaka posamezna slika predstavlja eno instanco, ki nadomesti fizično sliko ali posamezno datoteko na prenosnem mediju. V primeru CT posnetka lahko ena instanca predstavlja posamezno rezino.

2.2 Uporabniški vmesnik

Uporabnika najprej pričaka vpisno okence (slika 2). Vsak uporabnik ima unikatno identifikacijsko številko, v primeru pacienta morda kar številko zdravstvenega zavarovanja in osebno geslo. Uporabnik je lahko pacient, ki ima dostop le do lastnih medicinskih slik, ali zdravnik, ki ima dostop do slik vseh svojih pacientov. Tako je zdravniku ob vpisu ponujen seznam pacientov, medtem ko je pacient takoj usmerjen na zavihek študije.

Uporabnik izbere ime pacienta (študije, serije, instance) in izbiro potrdi s klikom na gumb (slika 3). S strežnika Orthanc[4] se prenesejo ustrezni podatki, nato pa se odpre nov zavihek naslednje ravni. Za večino uporabnikov je najbolj zanimiv zavihek serija, ki omogoča tekoč pregled vseh instanc posamezne študije. S klikom na "Poglej preiskavo v novem zavihku" se odpre spletni prikazovalnik Orthanc, v katerem lahko s koleščkom na miški drsimo skozi instance. Če miško s pritisnjenim levim gumbom vodimo levo ali desno po sliki, spreminjamo kontrast in svetlost posamezne instance.

Vsako instanco, serijo ali študijo (vse serije) lahko prenesemo v obliki arhiva zip. Prenešene datoteke so datoteke dcm (format DICOM), za katere potrebujemo ustrezno programsko opremo (npr. pregledovalnik medicinskih slik Weasis [5]). Prenešene datoteke dcm so nestisnjena oblika medicinske slike, ki jo ustvaril posamezen slikovni aparat. Posamezna datoteka je zato lahko velika do nekaj deset MB. Za paciente večinoma ne prinaša dodatne vrednosti, zdravniku pa omogoča naknadno prilagajanje več parametrov posnetka.

Zavihek "Instanca" prikaže posamezno instanco v formatu png. Na levi strani je ves čas prisotna orientacijska razpredelnica, ki uporabnika seznani s trenutno izbranimi ravnmi hierarhije DICOM (slika 4).

2.3 Platforma R Shiny

Aplikacija za oddaljen dostop do digitalnih izvidov po standardu DICOM je napisana v statističnem programskem jeziku `R`, razširjenim s knjižnico `Shiny` [6]. Slednja omogoča postavitev spletne aplikacije brez poglobljenega znanja programiranja spletnih strani in vsebuje uporabniški vmesnik (*angl.* front-end) in strežniški del kode (*angl.* back-end).



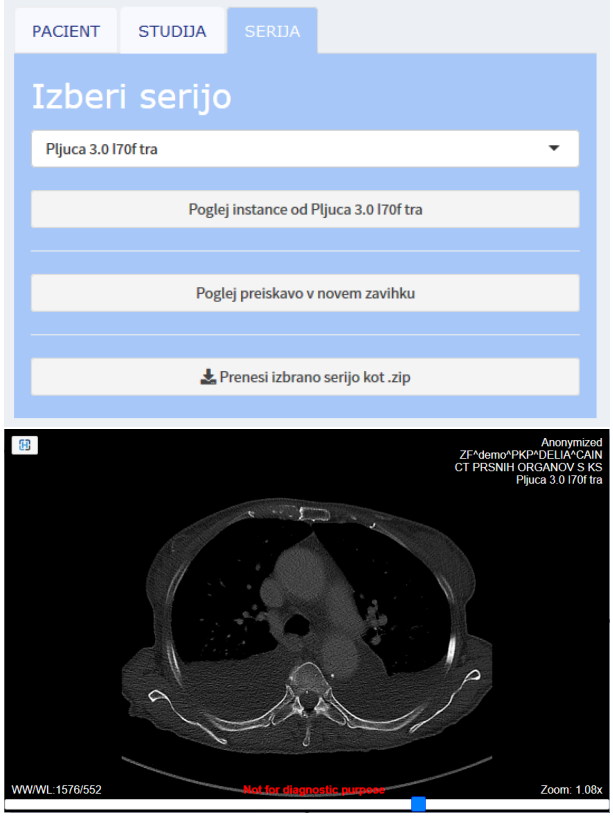
Slika 2: Zgoraj: Prijavno okno. Ker je aplikacija DEMO različica, so objavljeni tudi uporabniški podatki. Spodaj: Izbira pacienta. Seznam pacientov je v obliki padajočega seznama.

Osnovna struktura aplikacije Shiny R:

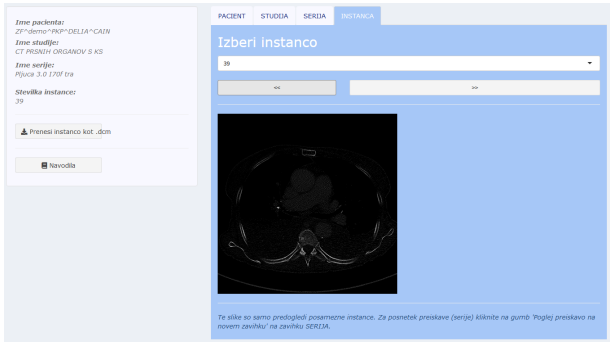
```
library(shiny)
# Define UI for application
ui <- fluidPage()
# Define server logic
server <- function(input, output) {}
# Run the application
shinyApp(ui = ui, server = server)
```

Spletna aplikacija v `R` (`Shiny`) je razdeljena na dva dela – na strukturi `server` in `ui`, ki se kot argumenta podata funkciji `shinyApp()`. Slednja zažene strežnik, ki nudi dostop do aplikacije. Na osebnem računalniku v času razvoja za to poskrbi kar razvojno okolje `RStudio`, na javnem strežniku pa potrebujemo `RStudio Server`.

V `ui` je koda za uporabniški vmesnik. Določiti moramo osnovno postavitev strani, morebitno razdelitev na zavihke in položaje modulov s katerimi uporabnik upravlja (gumbi, meniji, lokacije slik,...). Vsak modul se



Slika 3: Zgoraj: Zavihek serija. Spodaj: spletni prikazovalnik Orthanc.



Slika 4: Trenutno izbrane DICOM ravni na zavihku instanca.

določi v obliki funkcije, npr. `tabPanel(args)`, `actionButton(args)`, `selectInput(args)`. Funkcija `ui` se izvede le enkrat, ko se stran naloži, nato pa se ne izvaja več. To pomeni, da se moduli kasneje ne spreminjajo glede na interakcije z uporabnikom, temveč je njihov izgled (slika), funkcija (gumb) ali vsebina (meni) določena ob naložitvi aplikacije. To omogoča izdelavo strani z le izrazito deterministično strukturo in podatki, ki so na voljo že ob zagonu aplikacije ter se nato ne spreminjajo več. Za prikaz dinamičnih podatkov je rešitev uporaba funkcije `uiOutput()`, ki osrednji strukturi `ui` pove, da bo na določenem mestu nek element, ki se bo spreminjal glede na uporabnikove vnose. Tako se ob naložitvi strani za določen element le rezervira geometrijski prostor v aplikaciji, vsebina pa se dinamično oblikuje in spreminja med uporabe aplikacije.

Uporaba funkcije `uiOutput()`:

```
tabPanel(
  title = div("PACIENT",
    style="font-family:verdana;"),
  value = "1",
  column(
    width = 12,
    uiOutput("selectPatient"),
    uiOutput("toStudyButton"),
    hr(),
    uiOutput
      ("launchPatienthWeasisButton"),
  ),
  fluidRow(),
  br()
)
```

V naši aplikaciji smo se odločili še za dinamično dodajanje oz. odstranjevanje zavihkov glede na napredovanje uporabnika po hierarhiji standarda DICOM. Funkcije, ki jih proži uporabnik s klikom na gumb, kličejo funkcijo `addPanel()`, ki znotraj aplikacije naloži dodaten zavihek, o katerem pred klikom ni bilo nič znanega. Tako smo iz strukture `ui` praktično vso kodo, ki določa izgled, preselili v strukturo `server`, saj je stran izjemno dinamična in nedeterministična.

V strukturi `server` se poleg funkcij za izgradnjo omenjenih dinamičnih elementov nahaja vsa logika aplikacije. Ta vključuje funkcije, ki se odzivajo na klike gumbov in izvajajo klice funkcij, ki iz Orthanc strežnika prenesejo podatke o pacientu, študiji, seriji oz. instanci in jih nato implementirajo v padajočih menijih in gumbih oz. prikažejo kot slike, ter dodajo ali odstranijo zavihke. Struktura `server` tako predstavlja 95% vse programske kode.

Aplikacija v okviru razširitve `shinydashboard` vsebuje tudi vpisno okence, ki je pravzaprav neodvisna predaplikacija, ki ob uspešni avtentikaciji naloži osrednjo aplikacijo. Ob vpisu se glede na uporabniško ime določi tudi stopnja dostopa v obliki spremenljivke `permission`, ki se uporablja kot indikator v osrednji aplikaciji za omejitev modulov, ki se naložijo. Uporabnikovi prijavni podatki so v našem primeru definirani kar v programski kodi, vendar omenjena razširitev omogoča tudi preverjanje iz podatkovne baze.

Programiranje z vgrajenimi funkcijami je sicer precej omejeno, saj je razširitev `Shiny` namenjena predvsem predstavitvi statističnih podatkov in ne toliko splošnih podatkov, prenesenih iz drugih virov. Omogoča tudi dodajanje kode CSS za oblikovanje in kode JavaScript za dodajanje manjših funkcionalnosti.

Uporaba HTML `style` značke za dodajanje kode CSS:

```
div(paste(
  reacVals$instances$instanceNumber
    [instIdx(input,reacVals)]),
style="font-family:verdana;
```

```
  font-style: italic;" ),
hr(),
downloadButton(
  outputId = "downloadInstance",
  label = "Prenesi instanco
    kot .dcm",
  style = "vertical-align: middle;
          font-family:verdana;
          width:50%;" )
```

Baza podatkov za aplikacijo je strežnik Orthanc. Orthanc ima pripravljen aplikacijski vmesnik REST (iz angl. "Representational state transfer"), tako imenovan REST API, ki omogoča komuniciranje s strežnikom. Za uporabo zahtev REST moramo v skripto R jezika dodati knjižnico `httr`. Naša aplikacija podatke le prenaša, tako da uporablja le GET zahtevo v obliki funkcije GET(). Funkcija vrne odgovor iz strežnika, ki je v primeru Orthanca zapakiran v formatu JSON (skrajšano od angl. "JavaScript Object Notation"). V tem primeru zna R zapisati vsebino odgovora JSON v spremenljivko, do vsebine posameznih značk znotraj odgovora JSON pa se enostavno dokopljemo tako, da zaporedoma pišemo značke. Alternativa je, da uporabimo funkcijo, ki nam želeno značko izlušči iz odgovora, npr. s `status_code(response)`.

GET zahteva za prenos podatkov o pacientu:

```
response = GET(
  paste("https://pacs.zf.uni-lj.si/
    orthanc/patients/",
    patientIDs[i],sep=""),
  authenticate(username,password)
)
patientName =
  content(response)
  $MainDicomTags$PatientName
patID =
  content(response)
  $MainDicomTags$PatientID
status =
  status_code(response)
```

Funkcije iz `httr` knjižnice namenjene komuniciranju sprejmejo tudi opcijski argument `authenticate()`, kjer podamo podatke za dostop do zaščitenega strežnika. Prenos datotek se opravi z uporabo funkcije jezika `R download()`.

Prenos instance v obliki slike v formatu png:

```
outfile <- tempfile(fileext = '.png')
src=paste("https://",
      username,":",
      password,
      "@pacs.zf.uni-lj.si
        /orthanc/instances/",
        instanceID,"/preview",
      sep = "" )
download.file(src, outfile,
  mode = 'wb')
```

3 Razprava in zaključek

Tekom razvoja aplikacije smo naleteli na nekaj težav, saj je `R Shiny` zasnovan predvsem za prijazen prikaz podatkov generiranih v jeziku `R` in ne toliko za oblikovanje splošnih spletnih aplikacij. Tudi dokumentacija za knjižnico `Shiny` je precej skopa. So pa toliko bolj bogati spletni forumi, kar daje vedeti, da je `Shiny` pogosto uporabljano okolje tudi za izdelavo spletnih aplikacij. Hkrati omogoča oblikovanje s kaskadnimi stilskimi podlogami (CSS) in dodajanje JavaScript kode. Zaradi strukture jezika `R (Shiny)` daljša koda sicer hitro postane nepregledna, hkrati pa deljenje aplikacije na več datotek ni preveč razvijalcu prijazno. Povzeli bi lahko, da je `R Shiny` primeren za oblikovanje enostavnejših aplikacij za prikaz večje količine podatkov in manj za specifične aplikacije z veliko interakcije z uporabnikom. V našem primeru je sicer `R Shiny` več kot zadostoval in na koncu je nastala konceptualna aplikacija, ki izpolnjuje vse na začetku zastavljene cilje.

Razvili smo preprosto odprtokodno spletno aplikacijo za dostop do medicinskih slik v formatu DICOM, ki je lahko tudi osnova za razvoj cenovno dostopne storitve, ki morda najde svoje mesto predvsem v manjših medicinskih ustanovah s slikovno diagnostiko.

4 Zahvala

Pričujoče delo je v določeni meri del projekta *Odprtokodni sistem za zajem in obdelavo medicinskih slik iz ultrazvočnih diagnostičnih naprav po DICOM standardu*, ki smo ga med marcem in junijem 2020 izvedli v okviru programa *Po kreativni poti do znanja*. Zato se avtorji zahvaljujemo *Javnemu štipendijskemu, razvojnemu, invalidskemu in preživninskemu skladu RS*, *Ministrstvu za izobraževanje, znanost in šport RS*, *Evropskemu socialnemu skladu EU*, *Zdravstveni fakulteti UL* ter *Univerzi v Ljubljani* za finančno in logistično podporo. In nenazadnje seveda tudi vsem ostalim sodelujočim na projektu: podjetju *ECHOSON d.o.o.* in študentom Manci Pišek, Pii Pikelj, Sari Kidrič, Ninu Kleindienstu in Nejcu Klanjščku.

Literatura

[1] A. Gerečnik, *Implementacijao doprtokodnih sistemov PACS IN RIS*. Ljubljana: Zdravstvena fakulteta, 2020.

[2] G. Smith, "Introduction to RIS and PACS," in *PACS* (K. Dreyer, J. Thrall, D. Hirschorn, and A. Mehta, eds.), pp. 9–25, Springer, 2006.

[3] Wikipedia, "DICOM — Wikipedia, the free encyclopedia." `http://en.wikipedia.org/w/index.php?title=DICOM&oldid=955326347`, 2020. [Spletna različica; dostop 19. julija 2020].

[4] S. Jodogne and Orthanc community, "Orthanc." `https://www.orthanc-server.com/`, 2012–now. Odprtokodna programska oprema. [Spletna različica; dostop 19. julija 2020].

[5] University Hospital of Geneva, "Weasis medical viewer." `https://nroduit.github.io/en/`. Odprtokodna programska oprema. [Spletna različica; dostop 19. julija 2020].

[6] R Studio, “Shiny from R Studio.” https://shiny.rstudio.com/. Odprtokodna programska oprema. [Spletna različica; dostop 19. julija 2020].

Sistem prikazovanja eksponatov muzeja

Jan Breznar, Bogdan Lipuš, Simon Kolmanič

Laboratorij za geometrijsko modeliranje in algoritme multimedije, Inštitut za računalništvo,
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerze v Mariboru,
Koroška cesta 46, 2000 Maribor, Slovenija
E-pošta: jan.breznar2@um.si

Abstract. *We introduce a small information system that will help to improve the user experience in museums by interactive presentation of information about the exhibits. The system consists of three modules. The first module is a display application that takes care of displaying structured data about the exhibits. The second module is an editor that allows the user to edit and add information about new exhibits. The last module is a server application, an interface between the database and the first two modules. The goal of our work was to find an effective replacement for the classical information panels that are part of every museum exhibition. By replacing the classical panels with touchscreen computers, the whole new type of information about the exhibits can be included in the presentation. The users can take control over the amount of information they wish to take in, while the exhibition space can be used more efficiently comparing to the use of classical panels.*

1 Uvod

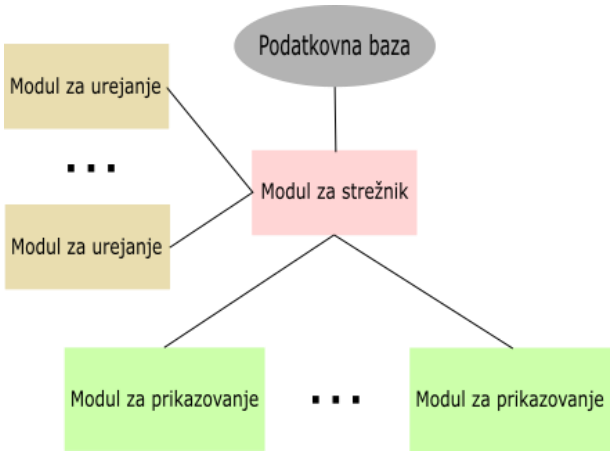
Obisk muzeja je skozi zgodovino eden od najučinkovitejših načinov spoznavanja krajev in njihove preteklosti. Tudi danes so obiski muzejev še vedno del turistične ponudbe nekega kraja. Da je muzej zanimiv, hkrati pa tudi informativen, morajo muzejske zbirke razen zanimivih razstavnih eksponatov ponujati tudi množico informacij, ki eksponate postavijo v ustrezen prostor in čas. Tako so informacijske table vedno del muzejskih razstav, ki pa zahtevajo veliko prostora, pa tudi ustrezno postavitev. Zaradi učinkovitejše izrabe razstavnega prostora, so se zelo zgodaj pojavile ideje, da bi v razstave vključevali računalnike [1, 2, 4], ki pa še vedno niso izrinili tiskanih informativnih tabel. To se lahko spremeni z uporabo računalnikov tipa vse v enem (angl.: All in one PC) z velikimi zasloni na dotik.

Za tak način predstavitve podatkov so se odločili tudi v Pokrajinskem muzeju Maribor, kjer bodo klasične informacijske table zamenjali z zasloni na dotik in računalniki tipa vse v enem. Razstava bo razdeljena v posamezne sobe, vsaka soba pa bo imela svoj zaslon s podatki o eksponatih v njej. Vsak eksponat bo predstavljen s fotografijo in množico povezanih podatkov. Ko bo uporabnik pritisnil na določeno fotografijo eksponata, se bo odprlo okno z osnovnimi informacijami o njem in povezavami do dodatnih informacij. Na ta način bo lahko uporabnik kontroliral, koliko podatkov bo pregledal, pripravljavci razstav pa lahko s pomočjo statistike ogledov ocenijo, katere informacije so bolj in katere manj zanimive. Uporabnik bo lahko izbiral med slovenskim in angleškim jezikom. Statistika pa bo kustosom omogočala analize, ali so informacije enako zanimive tujim in domačim obiskovalcem.

V članku je opisan razvoj manjšega informacijskega sistema, ki ga uporabljamo za urejanje in prikaz podatkov, kot tudi modulov za urejanje in prikazovanje podatkov o muzejskih eksponatih. Čeprav je predvideno, da se bodo podatki prikazovali na zaslonu na dotik, pa lahko z minimalnimi spremembami prikazovanje podatkov preselimo na katerokoli mobilno napravo.

2 Razvoj aplikacije

Aplikacija se deli na tri različne module. Prvi modul je namenjen strežniku in skrbi za vstavljanje in branje podatkov iz baze. Naslednji modul je prikazovalnik, ki prikazuje fotografije in besedila. Zadnji modul je urejevalnik podatkov, s katerim lahko urejamo in spreminjamo besedila ali slike eksponatov, glej sliko 1. Tako modul za vpisovanje, kot ogled podatkov je lahko hkrati inštaliran na več računalnikih hkrati, pri tem, pa morajo biti uporabniki pri vpisovanju podatkov pazljivi, da urejajo podatke o različnih sobah, saj aplikacija avtomatično ne skrbi za preprečevanja prepisovanja podatkov o eksponatih v primeru hkratnega urejanja.



Slika 1: Predstavitev in povezava med moduli.

2.1 Moduli za prikazovanje, urejanje in strežnik

Modul za prikazovanje bil razvit v programskem jeziku JavaScript. Za osnovno delovanje uporabljamo ogrodji

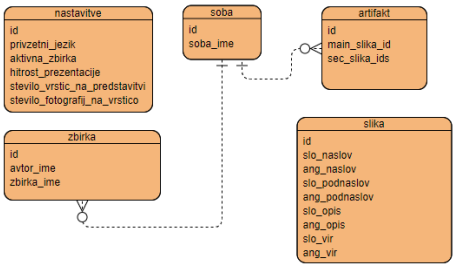
React[1] in Electron[2] [3]. Ogrodje React nam omogoča izdelovanje dinamičnih spletnih strani s pomočjo komponent, Electron pa nam pretvori spletno aplikacijo v namizno in je platformsko neodvisen.

Modul za urejanje je napisan v programskem jeziku C++. Pri tem smo tudi uporabili ogrodje Qt Creator[3], s katerim je bil izdelan videz in funkcionalnost aplikacije.

Modul za strežnik je napisan v programskem jeziku JavaScript in uporablja ogrodje »Express« in je platformno neodvisen, torej je primeren za strežnike z operacijskimi sistemi Linux ali Windows Server. Pri tem modulu je bila uporabljena relacijska podatkovna baza »MySql«. Baza je prav tako kot ogrodje neodvisna glede na platformo, na kateri se izvaja.

2.2 Podatkovna baza

V podatkovni bazi se nahaja pet tabel, ki jih prikazuje ER diagram na sliki 2.



Slika 2: E-R model podatkovne baze.

V tabeli »Nastavitve« hranimo privzeti jezik predstavitve. Ker je informacijski sistem zasnovan tako, da lahko v njem hranimo podatke o večih zbirkah, moramo hraniti tudi podatek, katera zbirka je trenutno aktivna. Fotografije eksponatov se prikazujejo v trakovih z drsečimi fotografijami, glej sliko 3, zato je potrebno shraniti število trakov in hitrost drsenja slik. Takšen način prikazovanja nam omogoča, da je predstavitev enako zanimiva, ne glede na število eksponatov v zbirki. V tabeli »Zbirka« hranimo podatke o imenu zbirke in imenu njenega avtorja. Na zbirko se navezujejo sobe (tabela »Soba«), v katerih shranimo ime sobe in primarni ključ. V tabeli »Artifakt« beležimo podatke o tem, v kateri sobi je eksponat lociran, in katere slike pripadajo temu eksponatu. Ena izmed fotografij je osnovna fotografija objekta, druge slike pa predstavljajo povezave do sekundarnih podatkov, ki se navezujejo na eksponat. V zadnji tabeli "Slika" hranimo podatke o naslovu, podnaslovu, opisu in viru. Vse te podatke hranimo za slovenski in angleški jezik.

3 Delovanje aplikacije

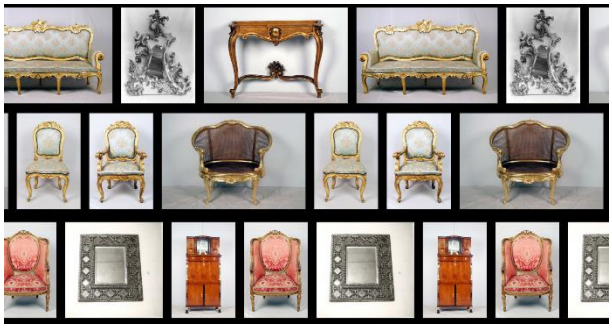
V nadaljevanju opisujemo delovanje obeh modulov, modul za prikazovanje, poimenovan Prikazovalnik ter modul za urejanje eksponatov, imenovan Urejevalnik.

Prikazovalnik in Urejevalnik se lahko nahajata na ločenih računalnikih, vendar morata imeti oba dostop do strežnika, na katerem se nahaja strežniški modul. Prikazovalnik je neodvisen od platforme, urejevalnik je pa primeren le za platformo Windows. Na več različnih računalnikih lahko teče več prikazovalnikov, kjer vsak prikazuje drugačne eksponate, glede na to na katero sobo so nastavljeni, da se prikazujejo.

3.1 Delovanje prikazovalnika

Prikazovalnik lahko prikazuje istočasno enega do tri trakove z drsečimi fotografijami eksponatov, kar je prikazano na sliki 3. Fotografije drsijo z vnaprej določeno hitrostjo, ki jo lahko nastavimo v Urejevalniku.

Ob zagonu Prikazovalnika se iz strežnika prenesejo vse glavne fotografije eksponatov, lociranih v določeni sobi, kar lahko pri veliki količini fotografij ali pri slabši povezavi do strežnika zahteva precej časa. V kolikor imamo prikazovalnik in strežnik v istem omrežju, začetno nalaganje slik deluje relativno hitro in dovolj gladko za nemoteno rabo. V vsakem primeru pa se prenos slik izvede zgolj ob prvem zagonu modula.



Slika 3: Prikazovalnik z drsečimi slikami eksponatov.

Uporabnik ima tudi možnost, da premakne vsebino vrstice v levo ali desno smer. Ob pritisku na željeni eksponat, se odpre novo okno katero prekrije delno drseče vrstice, ki ga prikazuje slika 4. Ob pritisku se v ozadju izvede zahteva na strežniški modul, tako aplikacija pridobi vse potrebne podatke (opis, naslov, …) o izbranem eksponatu. Teh podatkov je relativno malo, zato uporabnik ne zazna nobene zakasnitve pri odpiranju okna z informacijami o eksponatu.

Okno prikazuje podrobne informacije o izbranem eksponatu. Najprej je prikazana primarna fotografija in besedilo, ki opisuje vsebino slike. Pod opisom so še prikazane sekundarne fotografije eksponata, če jih le-ta ima. Ob pritisku na katero iz sekundarnih fotografij se ta

[1] https://reactjs.org/, zadnji obisk 19. 7. 2020
[2] https://www.electronjs.org/, zadnji obisk 19. 7. 2020
[3] https://www.qt.io/, zadnji ogled 19. 7. 2020.

poveča na mesto, kjer je prikazovanje fotografije in opis se prav tako spremeni. Jezik napisa lahko spremenimo ob pritisku na napis za drugi jezik.

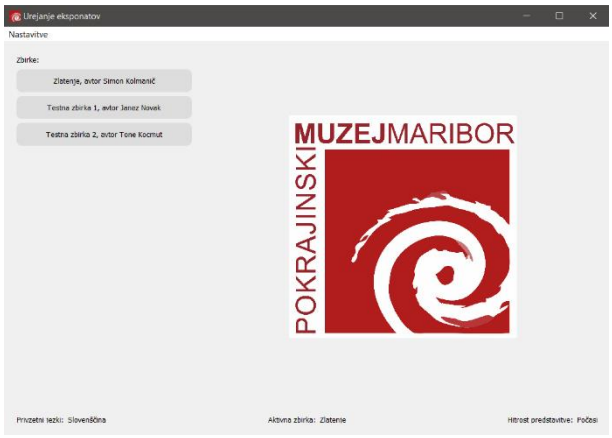


Slika 4: Okno z fotografijami in besedilo izbranega eksponata

Uporabnik ima možnost fotografijo povečati s kretnjo s prsti ali s pomočjo gumbov »+« ali »-«. Ker želimo, da se kvaliteta slike ob povečavi ne spremeni, morajo biti slike v dovolj veliki ločljivosti, ki pa zaradi hitrosti prenosa fotografij ob začetnem zagonu aplikacije, ne sme biti prevelika. S pomočjo eksperimentov smo določili, da je najprimernejše če je vsaj ena stranica slike večja od 2000 pik, druga pa večja od 1000 pik. V kolikor slika ne ustreza tem zahtevam, sistem o tem obvesti uporabnika, hkrati pa ne dovoli vnosa fotografij večjih od 10 MB.

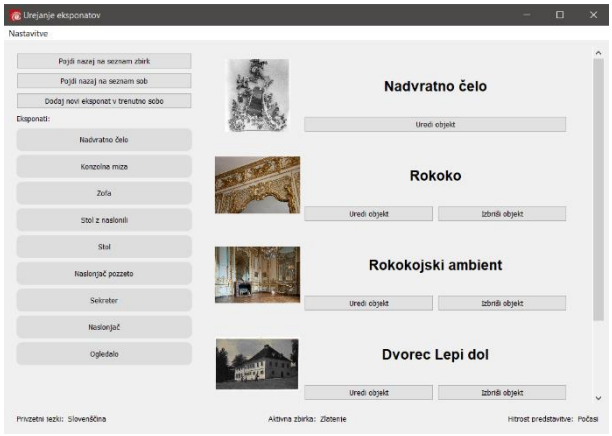
3.2 Delovanje urejevalnika

Urejevalnik ob zagonu najprej prikaže vse obstoječe zbirke, glej sliko 5. Tukaj imamo možnost urediti obstoječe zbirke, lahko jih izbrišemo, nastavimo določeno kot aktivno zbirko, ali pa ustvarimo novo.



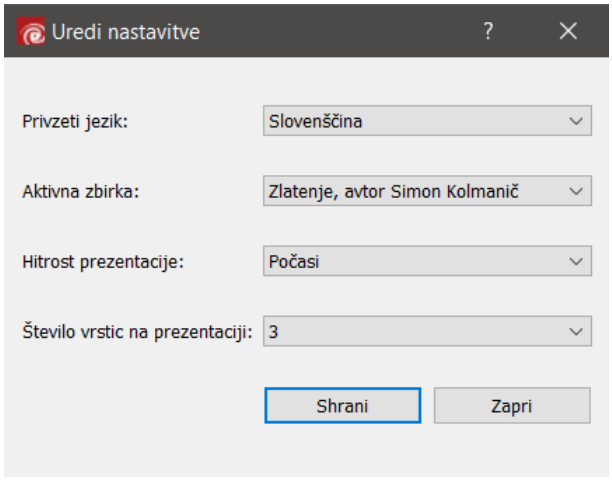
Slika 5: Urejevalnik z prikazom zbirk.

Ob izbiri ukaza »Uredi zbirko« se seznam zbirk zamenja s seznamom sob, ki pripadajo izbrani zbirki. Ko izberemo sobo, v kateri želimo urejati eksponate, se nam na levi strani seznam sob spremeni v seznam eksponatov izbrane sobe, glej sliko 6. Ob kliku na eksponat imamo možnost eksponat urediti, ga izbrisati, ali v njega dodati novo fotografijo. Ob izbiri eksponata, se na desni strani pojavi seznam vseh fotografij, ki pripadajo eksponatu. Ena izmed slik je glavna slika eksponata, ostale pa predstavljajo povezave do dodatnih informacij o objektu in se prikažejo v posebnem traku, ki ga lahko vidimo na sliki 4. Vse slike in podatke, vezane na njo obravnavamo na enak način. Tako imamo možnost urejati vsako fotografijo posebej in spremeniti njen opis, naslov in ostale podatke vezane na fotografijo. Ob dodajanju ali zamenjavi obstoječe fotografije, program sam preveri, ali je velikost datoteke s slike primerna. V kolikor datoteka presega dovoljenih 10 MB, jo sistem zavrne, medtem ko sliko s premajhno ločljivost na željo uporabnika shrani v bazo.



Slika 6: Prikaz seznama fotografij za določeni eksponat.

V Urejevalniku imamo tudi možnost urejati nastavitve za Prikazovalnik, slika 7. Te lahko odpremo s pomočjo gumba »Nastavitve«, kjer se odpre novo okno, v katerem lahko nastavimo privzeti jezik, določimo novo aktivno zbirko, hitrost predstavitve in število prikazanih trakov na prikazovalniku.



Slika 7: Okno za urejanje nastavitev.

3.3 Delovanje modula za strežnik

Modul za strežnik čaka na zahteve od Prikazovalnika in Urejevalnika. Njegova glavna vloga je, da v podatkovno bazo shranjuje podatke, ki jih dobi od Urejevalnika,

medtem, ko pri zahtevi Prikazovalnika iz baze pridobi vse potrebne podatku in jih pošlje nazaj Prikazovalnik.

4 Rezultati in razprava

Sistem za prikazovanje eksponatov smo testirali s strežnikom s 24 procesorji Xeon E5-2640 z 224 GB delovnega pomnilnika, na katerem smo namestili virtualni računalnik z operacijskim sistemom Ubuntu Server, ki smo mu dodelili 1 procesor in 1 GB pomnilnika, s čemer smo simulirali realni strežnik. Nanj smo namestili modul za strežnik in podatkovno bazo »MySql«. Urejevalnik smo namestili na več različnih računalnikov z različnimi procesorji in različno velikostjo delovnega pomnilnika, vsi pa so uporabljali operacijski sistem Windows 10. Enako smo testirali tudi Prikazovalnik, le da smo v tem primeru še dodatno uporabili računalnik z zaslonom, občutljivim na dotik. Računalniki so bili povezani tako v isto omrežje, kot tudi v oddaljeno omrežje, v vseh primerih je predstavitev tekla gladko in brez napak, le nalaganje slik ob prvem zagonu je bilo v oddaljenem omrežju občutno počasnejše.

S pomočjo Urejevalnika smo izdelali testno zbirko, v kateri so štiri sobe, zaradi majhnega števila eksponatov, ki smo jih dobili na razpolago pa smo napolnili le eno sobo. Dodali smo 9 eksponatov, kjer ima vsak eksponat 5 fotografij s potrebnima naslovom in opisi v slovenskem in angleškem jeziku.

Rezultati, ki smo jih dobili s pomočjo testiranja, so bili zadovoljivi, če upoštevamo, da je večina uporabljenih fotografij velikosti med 2 in 5 MB. V primeru, ko so bili vsi računalniki v istem omrežju, so bili časi nalaganja zelo hitri, na oddaljenih omrežjih, pa je bilo potrebno počakati od 30 sekund do 2 minut, da so se prikazale slike. Pričakujemo, da bi se ob šibkem strežniku lahko pojavile težave s hitrostjo nalaganja, kadar bi bilo v zbirki več sto eksponatov, saj lahko postane strežnik preobremenjen s prenosom fotografij na Prikazovalnik. Tako je v celotni strukturi prav strežnik ozko grlo, od katerega je posledično odvisen čas nalaganja predstavitve v Prikazovalniku.

Težavo z nalaganjem fotografij pri Prikazovalniku, bi lahko v prihodnje rešili tako, da bi se s strežnika najprej prenesle fotografije v nižji ločljivosti, tako, da bi se Prikazovalnik lahko zagnal, nato pa bi se med delovanjem v ozadju prenesle slike polne ločljivosti, ki bi postopoma zamenjale tiste z nižjo ločljivostjo.

V prikazovalnik bi ob izkazani potrebi lahko implementirali sistem za beleženje statistike. S tem bi lahko ugotovili, kateri eksponati so med uporabniki bolj zanimivi in kateri manj, koliko uporabnikov povečuje fotografije in si ogleda podrobnosti in podobno, kar bi lahko dodatno izboljšalo hitrost delovanja, oziroma zmanjšalo potrebe po moči strežnika. Določili bi lahko tudi, katere vsebine so bolj zanimive za slovenske uporabnike in katere bolj za tuje.

5 Zaključek

V članku je bil opisan razvoj sistema za prikazovanje eksponatov, ki bi v muzejih doprinesel veliko večjo stopnjo interaktivnost pri ogledu razstav. Prikazovalnik je zelo razširljiva komponenta, ki bi jo lahko z minimalno dela pretvorili v spletno aplikacijo, ali pa morda v mobilno aplikacijo. Tako bi v muzejih še povečali interaktivnost in popestrili uporabniško izkušnjo, uporabniki pa bi za predstavitev lahko uporabili lastne naprave. Lahko pa bi izdelali tudi povsem virtualne razstave, kar bi bilo še posebej uporabno, če bi morali muzeje zaradi morebitnih karanten zapreti za obiskovalce. Seveda bi ob taki spremembi morali povečati tudi zmogljivost modula za strežnik, saj bi se z večjim številom povezav povečale tudi zahteve po prenosih fotografij in s tem tudi časi nalaganja fotografij.

Zahvala

To delo je podprla Slovenska raziskovalna agencija v okviru programske skupine P2-0041.

Literatura

[1] Bay, Herbert, Beat Fasel, and Luc Van Gool. "Interactive museum guide." In *The Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage*. 2005.

[2] Yura, Shunsuke, and Ken Sakamura. "Real-time browser for the Digital Museum available with low-cost terminals and low-bandwidth networks." In *Proceedings 13th TRON Project International Symposium/TEPS'96*, pp. 70-80. IEEE, 1996.

[3] spletni vir: Building a React Desktop App with Electron, https://blog.bitsrc.io/building-an-electron-app-with-electron-react-boilerplate-c7ef8d010a91
(dostopano dne: 14.8.2019)

[4] Walczak, Krzysztof, Wojciech Cellary, and Martin White. "Virtual museum exbibitions." *Computer* 39, no. 3 (2006): 93-95.

Mehanizmi izmenjave zaupanja in ugleda

David Jelenc

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Laboratorij za e-medije
Večna pot 113, 1000 Ljubljana, Slovenija
E-pošta: david.jelenc@fri.uni-lj.si

Abstract

Trust and reputation systems are systems that estimate the trustworthiness of entities as they roam virtual communities and applications. But while entities can traverse application boundaries their trust and reputation information cannot: it remains locked to the application in which it was generated which forces entities to build trust and reputation anew in every application they join. We describe the issue of trust and reputation system information exchange. We define the problem, outline the set of desired features that a exchange mechanism ought to have and analyze the existing solutions through the lens of required features. We conclude that only a few solutions exist that address the issue of trust and reputation information exchange and all of them are in some aspect limited.

1 Uvod

Z nenehno rastjo omrežja Internet postajajo aplikacije, v katerih raznovrstne entitete—bodisi ljudje, naprave ali avtonomne storitve—zasledujejo svoje cilje s sodelovanjem, čedalje bolj razširjene. A v tovrstnih aplikacijskih okolij je sodelovanje težko vzpostaviti, saj ne vemo, ali bo entiteta, s katero smo v interakciji, dejansko spoštovala svoje obveze. Takšni pomisleki so še posebej izraziti v odprtih okoljih, ki nimajo razsodnikov, ki bi tovrstne spore razreševali, ali v primerih, kadar so si entitete neznane. V takih primerih se pogosto uporabijo modeli zaupanja in ugleda, ki zbirajo, ocenjujejo in razpečujejo podatke o zaupanju in ugledu entitet. Ti podatki se lahko nato uporabijo v raznovrstnih odločitvah, kot so denimo s kom stopiti v kupo-prodajno interakcijo, od koga pridobiti podatke in podobno. In čeprav se zavedamo, da imata besedi zaupanje in ugled različen pomen, bomo v tem prispevku uporabljali pojma model zaupanja in model ugleda kot sinonima; razlog je v tem, da sta v znanstveni literaturi pogosto pomešana.

Primerov uporab sistemov za obvladovanje zaupanja in ugleda je precej in posledično se taki sistemi zelo razlikujejo. Z vidika arhitekturnega pristopa poznamo centralizirane in porazdeljene storitve. Primer prvih so ocenjevalni sistemi, s katerimi kupci ocenjujejo prodajalce na centraliziranih elektronskih tržnicah [18, 17], primeri drugih pa porazdeljena okolja P2P (angl. peer-to-peer) za izmenjavo računskih virov [14]. Nadaljnje lahko sisteme za obvladovanje zaupanja in ugleda ločimo glede to, komu so v pomoč: bodisi podpirajo odločanje ljudi bodisi prispevajo vhodne podatke v povsem avtomatizirane procese, kot je denimo usmerjanje omrežnega prometa [9]. A vsi taki sistemi imajo enak cilj, ki je neodvisen od arhitekturnega pristopa ali končnega uporabnika: oceniti stopnjo zaupanja entitet in izboljšati odločanje.

Eden izmed odprtih problem sistemov za obvladovanje zaupanja in ugleda je njihova nezmožnost izmenjave podatkov. Medtem ko entitete prosto prehajajo med različnimi platformami, so njihovi podatki o zaupanju in ugledu vanje tesno zasidrani in dostopni le aplikaciji, ki jih je ustvarila. In čeprav lahko včasih poslovna odločitev povzroči nastanek takšnih podatkovnih silosov (primeri neuspelih poskusov združitve podatkov s platforme Amazon in eBay [19]), lahko velik del zaslug pripišemo tudi pomanjkanju rešitev in standardiziranih pristopov, ki bi takšno izmenjavo omogočali. Denimo, v preglednem prispevku o sistemih zaupanja in ugleda Hendrikx et al. [5] ugotavljajo, da je zmožnost uvoza in izvoza tovrstnih podatkov eden izmed odprtih izzivov.

Celovita rešitev izmenjave podatkov o zaupanju in ugledu bi prinesla trojno korist [4]. Prvič, omogočila bi aplikacijam, da ocenijo stopnjo zaupanja in ugleda z večjo mero natančnosti; tako kot vsak algoritem za podajanje napovedi, tudi sistemi za zaupanje in ugleda delujejo bolje, če imajo na razpolago več podatkov. Drugič, izmenjava podatkov bi omogočila entitetam, da prenašajo pridobljen ugled in zaupanje med različnimi sistemi. Tako entitetam ne bi bilo treba le-teh znova vzpostavljati v vsakem sistemu posebej. In tretjič, izmenjava podatkov bi omogočila hitrejšo rast in razvoj novih sistemov. Tak primer je paradigma Internet Stvari (angl. Internet of Things), v kateri se večje število vsakodnevnih naprav (značke RFID, senzorji, aktuatorji, mobilne naprave) povezuje in omogoča nove primere uporabe. Če želimo, da bodo interakcije med tovrstnimi entitetami uspešne, se bodo morale take *stvari* naučiti medsebojno zaupati in mnogi vidijo rešitev v sistemih za obvladovanje zaupanja in ugleda [12, 23].

2 Funkcionalne zahteve za izmenjavo

Rešitev za izmenjavo informacij o zaupanju in ugledu mora nasloviti štiri vidike: vsebino in strukturo sporočil,

vmesnik storitve, oceno zaupanja podatkov in prevajanje podatkov.

2.1 Določiti vsebino in strukturo sporočil

Prvi vidik zadeva vsebino in strukturo sporočil, ki kodirajo informacije o zaupanju in ugledu. Ob tem je velika raznolikost obstoječih sistemov težavna. Pri tovrstni analizi so pomembni vhodni podatki, iz katerih se izračun zaupanja vrši, in tip podatka, v katerem je izračun zaupanja izražen. Večina sistemov za obvladovanje zaupanja in ugleda vrši izračune na podlagi ocen, ki jih je entiteta pridobila v preteklih interakcijah, in ocen izpeljanih iz mnenj, ki so jih o tej entiteti podale druge entitete [15, 2, 5, 22]. Nekateri modeli še dodatno v izračun vključijo sistemske informacije, kot so struktura družbenega omrežja [20], kateremu entiteta pripada, tip vloge, ki jo entiteta opravlja [6] in podobno. V tem prispevku se bomo omejili le na najpogostejše modele tj. take, ki vršijo izračune na podlagi ocen: bodisi takih iz preteklih informacij bodisi takih, ki so izpeljane iz mnenj.

Podobno se sistemi razlikujejo tudi pri načinih, kako je ocena zaupanja izražena. Nekateri definirajo stopnjo zaupanja kot skalarno vrednost z nekega intervala, denimo število z intervala $[0,1]$, celo število z lestvice med 1 in 10, ali kvalitativna vrednost z urejenostne lestvice. Drugi podatke o zaupanju in ugledu kodirajo kot terko, kjer posamezna komponenta označuje del ocene: denimo sistemi, ki temeljijo na ogrodju subjektivne logike [8] uporabljajo trojico števil (b,d,u), ki označujejo količino prepričanja, dvoma in negotovosti zaporedoma; vsaka od komponent je z intervala $[0,1]$ in njihova skupna vsota znaša 1. Idealna rešitev bi bila taka, ki se jo da aplicirati na čim več obstoječih sistemov.

2.2 Določiti vmesnik storitve

Rešitev mora opredeliti jasen vmesnik storitve in načine za realizacijo. Potrebno je definirati, kako sistemi medsebojno zahtevajo podatke, v kakšnem obsegu ter s kakšno pogostostjo. Prav tako mora rešitev imeti mehanizem za sporočanje napak.

2.3 Določiti zanesljivost izmenjanih informacij

Tretji vidik rešitve zadeva oceno zanesljivosti izmenjanih informacij. Vsak sistem ni enako zaupanja vreden in dobra rešitev mora informacije o tem, kako zaupanja (ne)vredni posamezni sistemi so, pri integraciji upoštevati. V nasprotnem primeru se lahko zgodi, da so izračuni po izmenjavi slabši. Pri tem dodajmo, da večina sistemov za obvladovanje zaupanja že ima osnovne mehanizme za naslavljanje potencialno lažnih informacij, saj gre za temeljno funkcionalnost tovrstnih sistemov.

2.4 Prevesti podatke ob izmenjavi

Zadnji vidik pokriva mehanizme za prevajanje (angl. translation) podatkov med sistemi. Velika raznolikost med obstoječimi modeli zaupanja in ugleda pomeni, da ti operirajo z zelo različnimi tipi podatkov. Te razlike je potrebno pri integraciji upoštevati. Pri tem gre lahko bodisi zgolj za manjša odstopanja kot je denimo drugačen interval, na katerem so vrednosti izražene (denimo uporaba intervala $[0,1]$ ali $[0,100]$) bodisi za večje razlike, kot je denimo sistemska pristranskost (angl. bias) tj. nagnjenost sistema, da uporablja pretirano pozitivne (ali negativne) ocene.

3 Pregled obstoječih rešitev

Znanstvena literatura o pričujoči tematiki je precej skopa; dosedanje delo je večinoma skoncentrirano na razvoj modelov zaupanja in ugleda, integracija sistemov pa je bila zapostavljena. Eden izmed prvih poskusov izmenjave podatkov o zaupanju in ugledu prihaja s strani Trčka [21], ki predlaga definicijo tipov v jeziku XML (XML DTD) za opis vmesnika izmenjave. Predlagani sta dve sporočili, `trustResponse` in `trustRequest`, ki se uporabita za pošiljanje zahtevkov in podajanje odgovorov. Rešitev je dokaj osnovna in specifična za v istem članku predlagan model zaupanja in se ga ne da uporabiti širše. Prav tako rešitev ne vsebuje mehanizmov, s katerimi bi lahko omejili količino izmenjanih podatkov (rešitev implicira, da vsak `trustResponse` vsebuje vse podatke sistema)

V sorodnem delu Kovač in Trček [10] predlagata spletno storitev, ki implementira vmesnik za izmenjavo. Vmesnik omogoča agentom pridobiti oceno stopnje zaupanja drugih agentov kot tudi shranjevanja ocen iz preteklih interakcij. Čeprav se omenjeno delo ne osredotoča na integracijo, kljub temu podaja definicijo vsebine in strukturo sporočil ter vmesnik storitve. A prav tako so definicije prilagojene specifičnemu modelu zaupanja in v rešitvi manjkajo poizvedovalni mehanizmi. Vmesnik storitve in sporočila so definirana s pomočjo sheme XML.

Marienfield et al. [13] predlagajo ontologijo za podajanje ocen. Avtorji podajo strukturo sporočila, ki vsebuje šest komponent, in sicer `about`, ki pove, na katero entiteto se ocena navezuje, `submittedBy`, ki pove, kdo je oceno podal, `creationTime`, ki poda čas oddaje ocene, `hasAspect` govori o kontekstu oz. storitvi, za katero je bila ocena podana, ter komponenti `hasScale` in `hasValue`, s katerima je ocena definirana. Prva komponenta govori o tipu ocene oz. o ocenjevalni lestvici (nominalna, urejenostna, intervalna, razmerostna), druga pa o sami vrednosti. Takšna definicija ocene je precej bolj splošna od prejšnjih, a pokriva zgolj modele, kjer so ocene skalarji; večvrednostnih ocen s tem sporočilom ni mogoče izraziti. Na podoben način so ocene podane tudi v okolju za testiranje modelov zaupanja Alpha Testbed [7], kjer posamezna ocena sestoji iz petih komponent: `source`, `target`, `service`, `date` in `value`, ki zaporedoma označujejo ocenjevalca, ocenjenca, tip storitve, datum ocenjevanja in podano oceno, ki je izražena kot decimalno število z intervala $[0,1]$.

Grinshpoun et al. [4] predlagajo mehanizem za deljenje ugleda v virtualnih skupnostnih, ki so ga poimenovali CCR (angl. cross-community reputation). Delo primarno naslavlja problem določanja zanesljivosti in prevajanja podatkov ob izmenjavi. Delovanje sistema lahko opišemo sledeče: ko sistem A pošlje poizvedbo v sistem B in ko slednji vrne odgovor, se nad izmenjanimi podatki opravi zaporedje transformacij. Najprej se prejeti podatki preslikajo iz domene, ki jo uporablja sistem

Tabela 1: Ocena splošnosti obravnavanih rešitev glede na kriterije iz sekcije 2. Pomen ocen je sledeč: 0 - rešitev ni podana; 1 - rešitev je primerna le za specifične modele; 2 - rešitev je širše uporabna; 3 - rešitev je povsem splošna.

Delo	Sporočila	Vmesnik	Zanesljivost	Prevedba
Trček [21]	1	1	0	0
Kovač in Trček [10]	1	1	0	0
Marienfeld et al. [13]	2	0	0	0
CCR in TRIC [4, 3]	1	1	2	2

B, v domeno, ki jo uporablja sistem A. Zatem se podatki preslikajo glede na kontekst, v katerem so nastali: sprva se preslikajo iz kontekstov, ki jih uporablja sistem B, v univerzalne kontekste, nakar se ocene iz univerzalnih kontekstov preslikajo v kontekste, ki jih uporablja sistem A. Avtorji predpostavljajo, da vse ocene vsebujejo še podatek o zanesljivosti; sama ocena je par vrednost-zanesljivost. Zanesljivost ocene se ob vsaki taki transformaciji zniža glede na obseg spremembe; uporabljena je različica Pinyolovega prevajanja ocen [16].

Gal-Oz et al. [3] podajo implementacijski vidik pristopa CCR imenovan sistem TRIC (angl. Trust and Reputation In virtual Communities). Avtorji zasnujejo centralizirano arhitekturo, v kateri strežnik TRIC igra vlogo posrednika za izmenjavo informacij. Posledično je strežnik TRIC zadolžen za vsa prevajanja in za izvajanje varnostnih politik. V delu avtorji obravnavajo tri vidike izmenjave: kdo jo prične, kdaj se zgodi in kakšen je njen obseg.

Sistema CCR in TRIC sta obetavna poskusa za izmenjavo informacij o zaupanju in ugledu v centraliziranih okoljih kot so denimo spletne tržnice. Žal pa avtorji predpostavljajo vnaprej določeno obliko ocen (par vrednost-zanesljivost) ter obstoj univerzalno določenih kontekstov, kar omejuje splošnost rešitve.

Z vidika funkcionalnih zahtev iz sekcije 2 lahko povzamemo, da Trčkov in Kovačev [21, 10] predlog le delno naslavljata vsebino in strukturo sporočil ter določata vmesnik storitve, Marienfield et al. [13] določa srednje napredno vsebino in strukturo sporočil, predloga CCR in TRIC [4, 3] pa v glavnem naslavljata zanesljivost in prevedbo pri izmenjavi; povzetek primerjav je podan v Tabeli 1. Pri tem poudarjamo, da nobena od obravnavanih rešitev ni v celoti splošna in da problem izmenjave ostaja odprt.

Pri koncu pregleda podajamo še lasten pogled na vsebino in strukturo sporočil ter mehanizme poizvedovanja (definicijo vmesnika). Na izmenjavo med različnimi sistemi zaupanja lahko gledamo tudi kot na integracijo podatkov iz porazdeljenih podatkovnih baz, pri čemer ima vsaka podatkovna baza svoje specifike, kot je struktura ocen. Rešitev se tako ponuja v domeni semantičnih tehnologij, kjer lahko uporabimo tehnologiji Resource Description Framework (RDF) [11] in SPARQL Protocol and RDF Query Language (SPARQL) [1]. Prvo lahko uporabimo za definicijo sporočil oz. ocen, drugo pa za pošiljanje prilagojenih poizvedb. Pri tovrstni rešitvi vsak sistem skrbi za svoj imenski prostor, v katerem definira sporočila in določi atribute, po katerih lahko ostali sistemi poizvedujejo, poizvedbe pa se vršijo preko tehnologije SPARQL

4 Zaključek

V prispevku smo obravnavali izziv izmenjave podatkov, ki jih uporabljajo sistemi za obvladovanje zaupanja in ugleda. Identificirali smo funkcionalne zahteve, ki jih mora rešitev imeti, in skozi njihovo prizmo obravnavali obstoječe rešitve. Ugotovili smo, da obstaja zgolj nekaj parcialnih rešitev in da izziv izmenjave podatkov med sistemi za obvladovanje zaupanja in ugleda ostaja odprt.

Literatura

[1] SPARQL 1.1 overview. W3C recommendation, W3C, March 2013. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

[2] Gennaro Costagliola, Vittorio Fuccella, and Fernando A Pascuccio. Towards a trust, reputation and recommendation meta model. *Journal of Visual Languages & Computing*, 2014.

[3] Nurit Gal-Oz, Tal Grinshpoun, Ehud Gudes, and Ingo Friese. Tric: An infrastructure for trust and reputation across virtual communities. In *Fifth International Conference on Internet and Web Applications and Services (ICIW)*, 2010.

[4] Tal Grinshpoun, Nurit Gal-Oz, Amnon Meisels, and Ehud Gudes. Ccr: A model for sharing reputation knowledge across virtual communities. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, 2009.

[5] Ferry Hendrikx, Kris Bubendorfer, and Ryan Chard. Reputation systems: A survey and taxonomy. *Journal of Parallel and Distributed Computing*, 2015.

[6] Ramón Hermoso, Holger Billhardt, Roberto Centeno, and Sascha Ossowski. Effective use of organisational abstractions for confidence models. In *Proceedings of the 4th European Workshop on Multi-Agent Systems EUMAS*, 2006.

[7] David Jelenc, Ramón Hermoso, Jordi Sabater-Mir, and Denis Trček. Decision making matters: A better way to evaluate trust models. *Knowledge-Based Systems*, 2013.

[8] Audun Jøsang. *Subjective logic*. Springer, 2016.

[9] Rida Khatoun, Youcef Begriche, Juliette Dromard, Lyes Khoukhi, and Ahmed Serhrouchni. A statistical trust system in wireless mesh networks. *Annals of Telecommunications*, 2016.

[10] Damjan Kovač and Denis Trček. Qualitative trust modeling in soa. *Journal of Systems Architecture*, 2009.

[11] Markus Lanthaler, Richard Cyganiak, and David Wood. RDF 1.1 concepts and abstract syntax. W3C recommendation, W3C, February 2014.

http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[12] Margaret L Loper and Brian Swenson. Machine to machine trust in smart cities. In *IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, 2017.

[13] Florian Marienfeld, Edzard Höfig, Andrea Horch, Maximilien Kintz, and Jan Finzen. Making sense of ratings: a common quantitative feedback ontology. In *Proceedings of the 7th International Conference on Semantic Systems*, 2011.

[14] Xianfu Meng, Tianjiao Li, and Yu Deng. prefertrust: An ordered preferences-based trust model in peer-to-peer networks. *Journal of Systems and Software*, 2016.

[15] Isaac Pinyol and Jordi Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 2011.

[16] Isaac Pinyol, Jordi Sabater-Mir, and Guifre Cuni. How to talk about reputation using a common ontology: From definition to implementation. In *Ninth Workshop on Trust in Agent Societies*, 2007.

[17] Ansley Post, Vijit Shah, and Alan Mislove. Bazaar: Strengthening user reputations in online marketplaces. In *Proceedings of NSDI'11: 8th USENIX Symposium on Networked Systems Design and Implementation*, 2011.

[18] Kevin Regan, Pascal Poupart, and Robin Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the National Conference on Artificial Intelligence*, 2006.

[19] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Commun. ACM*, 2000.

[20] Jordi Sabater, Mario Paolucci, and Rosaria Conte. Repage: Reputation and image among limited autonomous partners. *Journal of artificial societies and social simulation*, 9:3, 2006.

[21] Denis Trček. Towards trust management standardization. *Computer Standards & Interfaces*, 2004.

[22] Sokratis Vavilis, Milan Petković, and Nicola Zannone. A reference model for reputation systems. *Decision Support Systems*, 2014.

[23] Zheng Yan, Peng Zhang, and Athanasios V Vasilakos. A survey on trust management for internet of things. *Journal of network and computer applications*, 2014.

Razpoznavanje vzorcev

Pattern Recognition

Štetje objektov na slikah z uporabo genetskega algoritma

Gregor Babnik, Luka Šajn

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko
E-pošta: gregor.babnik@student.uni-lj.si, luka.sajn@fri.uni-lj.si

Counting objects in images using a genetic algorithm

The work deals with the automatic counting of objects in images. A genetic algorithm is used as a learning method to find appropriate operations used to process the images. The success of an individual solution is measured as a difference between the number of counted objects and the real object count. ARes algorithm is used to adjust the resolution of input images. The image processing part is implemented using two libraries TensorFlow and OpenCV. The work is tested against various sets of images in different domains.

1 Uvod

Naše delo [1] obravnava implementacijo samodejnega štetja objektov na slikah. Obstaja veliko domen, ki imajo množico slik, te pa vsebujejo veliko količino določenih objektov. Primeri so lahko slike celic, insektov, rastlin, zrn in podobno. Ročno štetje vsega tega je lahko zelo zamudno. Cilj je bil implementirati program, ki se bo z uporabo manjše, učne množice slik iz neke specifične domene naučil oziroma poiskal rešitev za avtomatsko štetje objektov na vseh slikah v tej domeni. Iskanje rešitev na podlagi učne množice je potekalo z uporabo genetskega algoritma. Učna množica slik je bila procesirana z zaporednimi operacijami, te naj bi ustvarjal genetski algoritem in jih nato postopoma izboljševal. Program smo implementirali v programskem jeziku Python, z uporabo programskih knjižnic TensorFlow in OpenCV, ki sta bili uporabljeni za obdelavo in procesiranje slik. Program naj bi poiskal potrebne operacije in prav tako parametre teh operacij samodejno. Uspešnost našega programa smo testirali na slikah iz več domen po naraščajoči kompleksnosti.

2 Implementacija

Celotni program smo napisali v programskem jeziku Python. Za uporabo Pythona smo se odločili, ker ga hkrati podpirata TensorFlow in OpenCV.

Naš program se deli na tri glavne dele:

- prvi del opravlja predprocesiranje, torej pripravo vhodnih slik na procesiranje,
- drugi del programa izvaja genetski algoritem in razvija nove osebke rešitev,
- tretji del, ki sprejme novonastali osebek in po inštrukcijah, vsebovanih v osebku, obdela množico slik in tako ovrednoti podani osebek. Ovrednoteni osebek je nato ali vstavljen v populacijo ali zavržen, če je slabši od vseh, ki so trenutno v populaciji. Tretji del programa je torej evalvacija osebkov, ki se zgodi v treh fazah: procesiranje z operacijami TensorFlow, procesiranje z operacijami OpenCV in nato še štetje.

2.1 Predprocesiranje

Preden program lahko začne obdelavo slik, kot mu narekuje genetski algoritem, je treba te slike primerno pripraviti. Vhodne slike so lahko različnih oblik in velikosti. Prav tako so slike običajno barvne, naš algoritem pa obravnava slike v sivinah.

2.1.1 Pretvorba v sivine

Prvi korak je sprememba vseh slik v sive. Najprej se slike iz RGB-barvnega prostora konvertirajo v barvni prostor CIELAB [6]. Prednost CIELAB je namreč njegova komponenta L, ki se nanaša na lightness oz. osvetljenost. Konvertiranje prek vrednosti osvetlitve prinaša najboljše rezultate pri aplikacijah računalniškega vida [3].

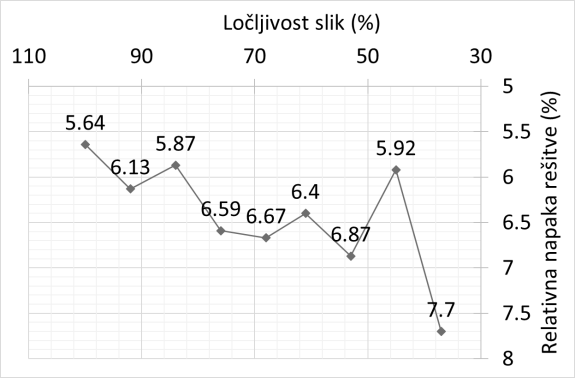
2.1.2 Izenačevanje slik

Ko so slike v sivinah, preidemo na drugi korak predprocesiranja. Zagotoviti je treba, da se uporabi celoten razpon sivin, ki so na voljo, oziroma izenačevanje histograma sivin posamezne slike. Ker običajno izenačevanje ne dosega dovolj dobrih rezultatov, uporabimo namenski algoritem CLAHE [7], ki močno izboljša kontrast in vidne podrobnosti na slikah. V primerjavi z običajnim izenačevanjem, ki deluje globalno, CLAHE izenačuje vrednosti sivin lokalno v obsegu 8×8 slikovnih točk in obenem omeji nastanek šumov, ki se utegnejo pojaviti ob tem.

2.1.3 Nastavitev ločljivosti - ARes

Tretji korak predprocesiranja je nastavitev vseh slik na skupno velikost in izbiranje primerne ločljivosti. Ta postopek opravimo z algoritmom *ARes* [8]. Algoritem *ARes* poišče primerne ločljivosti, ki so nižje od začetne. Pridobljena nižja resolucija nam omogoča hitrejše procesiranje slik. Prav tako nižje ločljivosti potrebujejo manj grafičnega pomnilnika, ki pa je omejen.

Primerna ločljivost se smatra za ločljivost, ki je manjša ali enaka začetni in pri tem ohrani oziroma izboljša kakovosti rešitve. Zmanjšana primerna resolucija lahko pripomore k izboljšanju rešitve (slika 1) tako, da odstrani nepotrebne in moteče podrobnosti slike.



Slika 1: Graf kakovosti rešitve v odvisnosti ločljivosti slik najdenih z algoritmom ARes. Primer štetja celic. [1]

2.2 Implementacija genetskega algoritma pri našem programu

Genetski algoritem deluje na populaciji osebkov, ki jo upravlja in nad njo izvaja selekcijo in križanje, ti pa pripeljeta do novega osebka. Novonastali osebek se nato še mutira.

2.2.1 Opis in zgradba osebka

Osebek vsebuje zapis oziroma navodila, katere operacije izvesti nad tenzorjem slik v našem programu. Osebek je sestavljen iz dveh delov, vsak del pa vsebuje zaporedje določenih operacij. V prvem delu so operacije, ki so implementirane v TensorFlowu. V drugem pa je zaporedje operacij, ki so implementirane v OpenCV.

2.2.2 Kriterijska funkcija

Kriterijska funkcija določa kvantitativni kriterij uspešnosti osebkov. Zmožnost pridobitve dobrih rezultatov pri genetskem algoritmu je zelo odvisna od kakovosti kriterijske funkcije. Kriterijska funkcija usmerja iskanje genetskega algoritma po prostoru in tako močno vpliva na konvergenco populacije. Pri uporabljeni kriterijski funkciji (enačba 1), x_i predstavlja resnično vrednost, y_i pa pridobljeno vrednost. Vsako posamezno odstopanje x_i od y_i se potencira, saj tako kaznujemo osebke, ki so dobri le v povprečju, obenem pa vsebujejo posamezne zelo nadpovprečno slabe rezultate.

$$f(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{100 * (x_i - y_i)}{x_i}^{e} \quad (1)$$

2.2.3 Postopek križanja osebkov

Križanje osebkov deluje v dveh nivojih. Na prvem deluje na nivoju vseh operacij, na drugem pa na nivoju genov posamezne operacije.

Križanje na nivoju vseh operacij v osebku uporablja dve različni strategiji: enostavno enotno križanje in križanje z ujemanjem zaporedja.

Enostavno enotno križanje (uniform crossover) Pri enostavnem enotnem križanju se novi osebek sestavlja tako, da se istoležeče operacije staršev zaporedno križajo druge z drugo.

Križanje z ujemanjem zaporedja Križanje z ujemanjem zaporedja (slika 2) v nasprotju z enostavnim enotnim križanjem upošteva tipe operacij in poskuša ohraniti obstoječe zaporedje, vsebovano v obeh starših, pri novem osebku. Križanje z ujemanjem omogoči prenos možne dobre podrešitve v novi osebek, kjer je podrešitev daljša od posamezne operacije in je pravzaprav zaporedje operacij. Cilj križanja z ujemanjem je hitrejše konvergiranje genetskega algoritma k rešitvi. Algoritem z ujemanjem zaporedja je predstavljen s psevdokodo (algoritem 1). Križanje parametrov operacij se zgodi le pri istoležnih operacijah istega tipa. Novi parameter se določi z naključnim izborom enega izmed parametrov staršev. Verjetnost, kateri parameter bo izbran, določimo z vhodnim parametrom programa *paramter_crossover_weight*. Če istoležni operaciji nista istega tipa, se parametri prepišejo le iz tistega starša, ki je istega tipa kot otrok. Če se tip otroka razlikuje od obeh staršev, se parametri naključno generirajo.

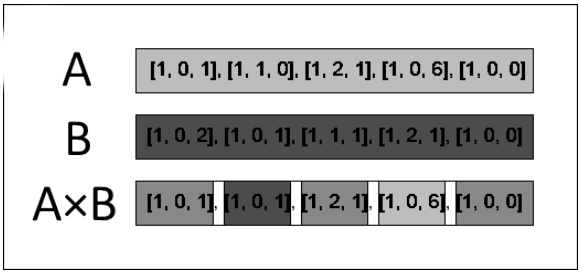
Algoritem 1 Najdi podobno zaporedje

```
indexA ← 0
indexB ← 0
for all operationA ∈ parentA do
  if indexB ≥ length(parentB) then
    indexB ← 0
  end if
  while indexB < length(parentB) do
    operationB ← parentB[indexB]
    if ¬sameType(operationA, operationB) then
      indexB ← indexB + 1
    end if
    append the pair {operationA, operationB} to
    the list commonOPs
  end while
  if operationA ∉ commonOPs then
    opeationB ← parentB[indexA]
    append the pair {operationA, operationB} to
    the list commonOPs
  end if
  indexA ← indexA + 1
end for
```

Zaporedne pare operacij nato križamo med seboj na nivoju genov.

2.2.4 Postopek mutacije osebkov

Mutacija osebkov se zgodi na dveh nivojih. Mutacija na prvi stopnji ima možnost spremembe enega ali več genov vsake operacije, kar spremeni operacijo v neko drugo



Slika 2: Primer križanja z ujemanjem zaporedja.

operacijo. Mutacija na drugi stopnji pa ima možnost spremeniti parametre vsake operacije v osebku. Mutacija parametrov se zgodi neodvisno od mutacije genov. Mutacija parametrov poteka tako, da se jih enostavno nadomesti z drugim naključno generiranim parametrom. Ali se zgodi mutacija, je odvisno od verjetnosti, ki jo določimo kot parameter programa *paramter_muatation_chance*. Ko je novi osebek ustvarjen, se uporabi pri procesiranju slik, pri čemer pridobi tudi oceno.

2.2.5 Opis operacij

V našem programu imamo dva tipa operacij. Prvi tip so operacije, ki so implementirane v TensorFlowu, drugi pa operacije, ki so implementirane v OpenCV. Prvi tip operacij zato imenujemo operacije TensorFlow, drugega pa operacije OpenCV. V program smo vključili pogosto uporabljene operacije procesiranja slik s področja avtomatskega štetja objektov [2]. Operacije Tensorflow (tabela 1) in operacije OpenCV (tabela 2) so podrobneje predstavljene v diplomskem delu [1].

	Koda	Kratek opis
1.	0 0 0	Ničelna operacija
2.	1 0 0	Inverzija
3.	1 0 1	Linearna točkovna op.
4.	1 0 2	Potenčna točkovna op.
5.	1 0 3	Logaritemska točkovna op.
6.	1 0 4	Polinomska točkovna op.
7.	1 0 5	Odstranitev intervala vrednosti
8.	1 0 6	Zamenjava vrednosti z drugo
9.	1 1 0	Linearna konvolucija
10.	1 1 1	Minimalni filter
11.	1 1 2	Maksimalni filter
12.	1 1 3	Povprečni filter
13.	1 1 4	Dvojna linearna konvolucija
14.	1 2 0	Sobelov operator
15.	1 2 1	Gaussova zameglitev
16.	1 2 2	Ostrenje

Tabela 1: Tabela operacij, implementiranih v Tensorflowu

2.3 Procesiranje slik in štetje

Procesiranje slik poteka z uporabo TensorFlowa in OpenCV. Prvi del procesiranja, ki je implementiran s TensorFlowom, predstavlja glavni del obdelave slik. Drugi del, implementiran z OpenCV, pa se osredotoča predvsem

	Koda	Kratek opis
1.	0	Ničelna operacija
2.	1	Inverzija
3.	2	Zameglitev z mediano
4.	3	Dvostranski filter
5.	4	Morfološka operacija
6.	5	Binarizacija
7.	6	Watershed algoritem

Tabela 2: Tabela operacij, implementiranih v OpenCV

na izboljšanje opravljenega dela prvega dela. Na koncu preštejemo skupke slikovnih točk, ki so se prikazali na sprocesiranih slikah in rezultate podamo kriterijski funkciji za končno oceno osebka.

2.3.1 Procesiranje s TensorFlowom

Na začetku izvajanja programa ustvarimo TensorFlow graf operacij. Graf vsebuje implementacije vseh operacij Tensorflow našega programa. Razlog za to je, da graf ustvarjamo samo enkrat na sejo, namesto ob vsakem procesiranju slik. Pri procesiranju pa izvajamo le tisti del grafa, ki izvede želeno operacijo. Pri vsakem ocenjevanju osebka naložimo tenzor neobdelanih slik v spremenljivko TensorFlow, nad katero nato izvršimo podane operacije.

2.3.2 Procesiranje z OpenCV

Ko je procesiranje s TensorFlowom opravljeno, pride na vrsto procesiranje z operacijami OpenCV. To je v nasprotju s TensorFlowom preprostejše, saj ne potrebuje vnaprej definiranega grafa operacij in ustvarjanja posebne seje. V primerjavi s TensorFlowom se slike obdelajo zaporedno.

3 Testiranje in rezultati

Za testne probleme smo izbrali množice slik, ki so vsebovale veliko določenih objektov. Množične slike smo razdelili na učno in testno množico. Program z uporabo genetskega algoritma izdela rešitev na podlagi učne množice in jo na koncu preizkusi na testni množici.

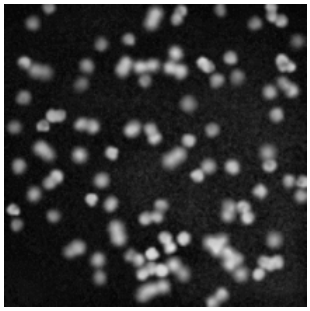
3.1 Test štetja celic

Test vsebuje umetno generirane realistične slike mikroskopskih celic bakterij [4]. Učna množica vsebuje 140 primerov slik, testna množica pa 60. Vhodne slike (slika 3) vsebujejo veliko število objektov celic, ki so neenakomerno razpršene po posamezni sliki in se tudi precej prekrivajo.

Rezultat testa ima poprečno relativno napako 5,88 %, standardna deviacija je 4,57 %, 95 % interval zaupanja je 5,88 % $\pm$ 1,16 %. Rezultat kvarijo predvsem slike, ki vsebujejo veliko prekrivajočih se celic. Problem prekrivajočih se celic program rešuje z uporabo algoritma Watershed, vendar ima tudi ta svoje omejitve in more pomagati, ko je prekrivanje previsoko.

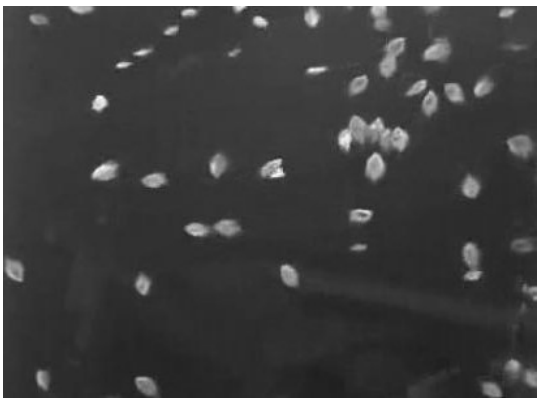
3.2 Test štetja rib

Test vsebuje slike rib [5]. Slike (slika 4) imajo globino in različno orientacijo posamezne ribe. Globina oziroma oddaljenost je sicer majhna, vendar je razlika v ostrini rib.



Slika 3: Celice: primer vhodne slike

Med ribami ni veliko prekrivanja. Ozadje je enobarvno in monotono. Odsev stekla akvarija nekoliko popači sliko. Velikost učne množice je 90 slik, velikost testne pa je 39 slik.



Slika 4: Primer slike rib.

Program ni imel večjih težav pri štetju posameznih rib. Napaka štetja je le pri ribah, ki se prekrivajo ali pa so nagnetene v večje gruče. Rezultat testa ima poprečno relativno napako 4,87 %, standardna deviacija je 3,3 %, 95 % interval zaupanja je 4,87 % $\pm$ 1,04 %.

3.3 Test štetja čebel

Zadnji test štetja vsebuje slike čebel [5]. Učna množica vsebuje 93 slik, testna množica pa 25.

Slike (slika 5) vsebujejo roj čebel na neenakomerni podlagi – travi in travnatih rastlinah. Na slikah skorajda ni prekrivanja med čebelami. Čebele so v zraku in tako različno orientirane, kar pomeni, da se njihovi obrisi razlikujejo po velikosti in obliki.

Povprečna relativna napaka štetja čebel je 8,29 % s standardno deviacijo 6,39 %. 95 % interval zaupanja je 8,29 % $\pm$ 2,50 %. To je bil najtežji izmed testov, ki smo jih dali v program. Del težavnosti izvira tudi iz tega, da imajo pri teh slikah barve pomembno vlogo - rumenkaste čebele med zelenkastim rastlinjem. Pri pretvorbi iz barvnih slik v sivini se je precej podatkov izgubilo.

4 Zaključek

Implementirali program, ki se nauči šteti objekte na slikah. Za učenje potrebuje množico slik, na kateri poteka



Slika 5: Primer slike čebel.

učenje. Učenje poteka tako, da program z uporabo genetskega algoritma išče uspešno zaporedje operacij, ki obdelajo slike. Zaporedje operacij je uspešno takrat, ko je razlika med preštetimi in dejanskimi objekti na slikah zmanjšana do zadovoljive mere. Ta razlika služi tudi kot vhodna spremenljivka v kriterijski funkciji. Pri določanju primernih ločljivosti slik smo si pomagali z algoritmom ARes. Za izvedbo operacij nad slikami smo uporabili programski knjižnici TensorFlow in OpenCV. Uporaba TensorFlowa nam je pospešila izvajanje programa s procesiranjem podatkov (obdelavo slik) na GPE. Delovanje programa smo testirali na različnih domenah slik, z različnimi težavnostmi za učenje. Prednost programa je samodejno iskanje rešitev.

Literatura

[1] Gregor Babnik. Štetje objektov na slikah z uporabo genetskega algoritma. Diplomska naloga, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2020.

[2] Jayme Garcia Arnal Barbedo. A review on methods for automatic counting of objects in digital images. *IEEE Latin America Transactions*, 10(5):2112–2124, 2012.

[3] Christopher Kanan and Garrison W Cottrell. Color-to-grayscale: does the method matter in image recognition? *PloS one*, 7(1):e29740, 2012.

[4] V. Lempitsky and A. Zisserman. Vgg cell dataset from learning to count objects in images. 2010.

[5] Zheng Ma, Lei Yu, and Antoni B Chan. Small instance detection by integer programming on object density maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3689–3697, 2015.

[6] K McLaren. Xiii—the development of the cie 1976 (l* a* b*) uniform colour space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92(9):338–341, 1976.

[7] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[8] Luka Šajn and Igor Kononenko. Multiresolution image parametrization for improving texture classification. *EURASIP Journal on Advances in Signal Processing*, 2008:137, 2008.

Semi-automated correction of MOBIUS eye region annotations

Ožbej Golob **Peter Peer** **Matej Vitek**

Computer Vision Laboratory, Faculty of Computer and Information Science,
University of Ljubljana, Večna pot 113, 1000 Ljubljana
E-pošta: ozbej.golob@gmail.com, {peter.peer, matej.vitek}@fri.uni-lj.si

Semi-automated correction of MOBIUS eye region annotations

MOBIUS is a publicly available dataset of ocular images with manually created annotations of the various eye regions. However, manual markups are prone to human error, and in this work we explore the semi-automatic approach we utilised to correct flaws in the MOBIUS dataset annotations. This improves the dataset's usability for segmentation and training deep segmentation models. The program we wrote removes areas that are outside of the required area (outliers), areas that are inside of other areas (inliers) and missing or blurred edges from annotations. This results in fixed annotations and a better deep learning model. Evaluation was performed on a model built on new annotations and a model built on original annotations from MOBIUS dataset. The model trained on fixed annotations achieved significantly better results in all metrics than the one trained on the original data. Thus, for a better learning model, the distortions on original annotations should be removed.

1 Introduction

Ocular biometrics refers to a branch of biometrics that studies the use of various eye parts for recognition. Ocular biometrics for recognition is receiving significant attention from researchers in recent years. Popular biometrics traits are iris, retina, sclera and the periocular region, where iris is the most popular [1, 2].

Sclera biometrics is not as popular as iris or periocular biometrics [3], despite the fact that it has some unique advantages. Sclera is a white region in the eye and contains blood vessel patterns that can be used for human recognition [4]. Sclera area is a highly secured part of the eye, so it is impossible to steal. It is a vessel pattern, so it is unique for each individual (even for twins) and it is stable throughout a persons lifetime [5, 6]. Sclera is reliable for recognition in case of potential eye redness and also in the presence of contact lenses [2]. Because of this properties, sclera biometrics is a reliable approach for human identification. Although it is not the most popular type of ocular biometrics, its popularity is rising. This is mainly due to a fact that some interesting recent research was performed on multi-modal eye recognition (using iris and sclera). This research reported that iris information fusion with sclera can increase the applicability of iris biometrics in off-angle and off-axis eye gaze [4].

Recognition systems based on the vasculature of sclera consist of (i) a vasculature segmentation stage, which extracts the vascular structure from sclera and (ii) a recognition stage, where the vascular structure is represented with image descriptors, which are then used for recognition. The first stage consists of two steps. The first step locates the sclera and the second step extracts the vasculature required for recognition [2].

In the field of sclera segmentation an important and extensive source of approaches and information is the annual Sclera Segmentation Benchmarking Competition (SSBC) [4, 7, 8, 9, 10]. The goal of these competitions was to set a common platform for evaluation of sclera segmentation and to record recent developments in sclera segmentation and recognition in visible spectrum.

Before deep learning most segmentation methods used handcrafted features and filter based methods. Only few of them were sclera specific. One example of those methods was Unsupervised Sclera Segmentation (USS) [11]. Most recent research in sclera segmentation uses general purpose deep models, usually based on convolutional encoder-decoder (CED) architecture. Some of those models are U-Net [2, 12], SegNet [13], ScleraSegNet [14], RITnet [15], RefineNet [16], and DeepLab [17].

Most of the existing research is focused on images taken in the near-infrared (NIR) spectrum, which requires special equipment to capture. The latest research is leaning towards images captured in visible spectrum, which can be taken with mobile phone cameras [18].

MOBIUS [19] is currently the only publicly accessible dataset for development of deep models for mobile ocular biometrics (including sclera segmentation and recognition). Annotations in the dataset are manually produced. Original images are taken with mobile phone cameras and are consequently captured in the visible spectrum. They also contain sclera, pupil and iris ground truth information, which is especially important for the development of deep models.

Because the annotations were manually produced, a large number of them contain distortions. In this work we present our method that removes distortions from annotations and fixes them, thus improving the quality of the dataset. This improves the dataset's usability both for

the evaluation of general segmentation methods as well as for learning deep segmentation models.

The rest of the paper is structured as follows: in section 2 we look at the different types of defects in the annotations and the methods we used to fix them. In section 3 we show both qualitative and quantitative results that point to the impact of our fixes. Finally, we conclude the paper in section 4 with a look at the implications of our work.

2 Methods

Our program is written in Python in combination with OpenCV library [20]. The annotations in the dataset consist of the sclera (green), iris (red), and pupil (blue). Our program therefore detected red, green and blue contours with the help of the OpenCV function `findContours`.

The annotations had three main types of distortions. In this section we look at each of the three types of distortions and the methods used to address them.

2.1 Outliers

In theory, only the sclera region should have been green, but some annotations had green spots (outliers) in areas other than the sclera. Outliers appeared with iris and pupil as well. An example of green outliers is shown in Figure 1.

To fix this problem with the iris, our program selected the biggest red contour, but only if its center was between the top and bottom of the biggest green contour's Y coordinates and if it had sufficient size (contour area bigger than 2500). Then it did the same for pupil, only with the biggest blue contour. The sclera was somewhat more complex since it wasn't necessarily one-piece like the iris and pupil were. To address this, the program first selected the biggest green contour, if it had sufficient size (contour area bigger than 5000). Then it processed each green contour and selected it only if it partially matched the Y coordinates of the biggest green contour.

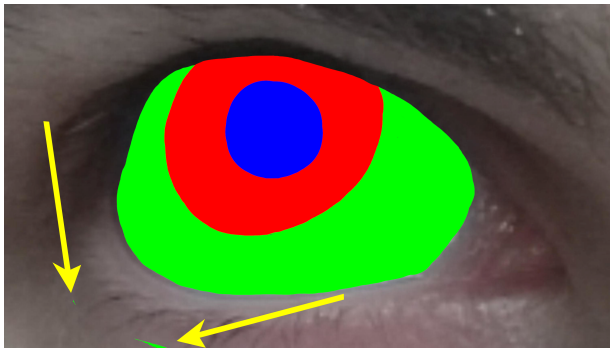


Figure 1: Annotation with green outliers (best viewed with zoom online).

2.2 Inliers

In some annotations there were blue and red spots (inliers) inside of the sclera region, which was supposed to be green. Inliers appeared on iris and pupil as well. An example of red inliers in the pupil is shown in Figure 2.

For iris, the program selected the biggest red contour and filled it with red. It did the same for the pupil, only with the biggest blue contour. For the sclera, program selected all contours that were left after the process of removing outliers (see the previous section) and filled them with green.

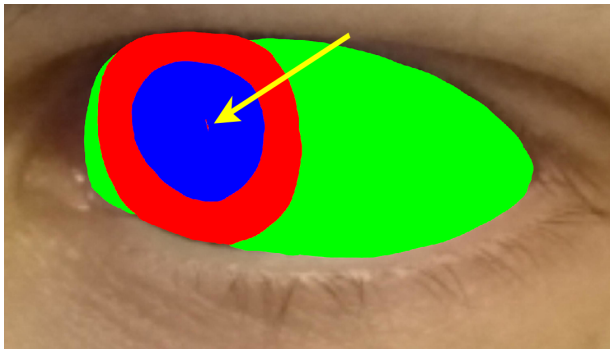


Figure 2: Annotation with red inliers (best viewed with zoom online).

2.3 Blur and missing edges

Some annotations had edges between sclera/iris or iris/pupil that weren't processed correctly. They were either missing or blurred. Example of missing edges between the sclera and iris regions is shown in Figure 3.

To fix this issue, our program created a Convex Hull of each green contour. Then it filled those contours with green color. After the process of fixing green missing edges, it created a Convex Hull of the biggest red contour and filled it with red. After that, it filled the biggest blue contour with blue to remove inliers.

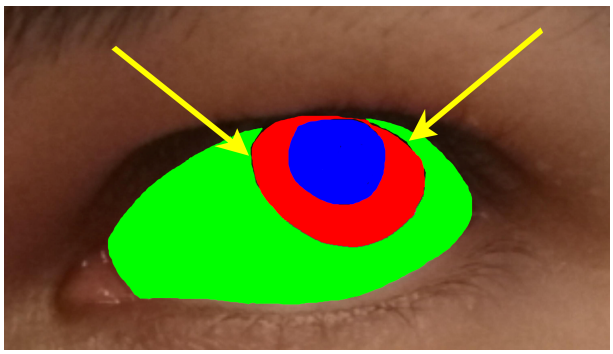


Figure 3: Annotation with missing edges (best viewed with zoom online).

3 Results

In this section we first look at a qualitative comparison between the original annotations from section 2 and the annotations corrected with our method. We also perform a quantitative analysis, using a state-of-the-art deep model for image segmentation, DeepLab [17].

Figure 4 shows an example of how our method removed green outliers from a faulty annotation. We can see the original annotation and the processed annotation. We also highlight the differences to clarify just what was removed/added in the fixed annotation. As we can see, the green outlier from the bottom of the before image has been removed.

Figure 5 shows images of annotation with red inliers before and after the program processed it and the differences. Notice that the red inliers from the pupil have been removed.

Figure 6 shows images of annotation with missing edges before and after the program processed it and differences. Missing edges between sclera and iris have been removed.

Because we filled each green and red contour using their Convex Hulls, the jagged edges present in some annotations were also smoothed out as an added bonus.

To demonstrate the impact of our annotation fixes, we trained and evaluated DeepLabV3+ [17] (which is currently state-of-the-art in many segmentation tasks) in the task of sclera segmentation on both the original annotations from the MOBIUS dataset and the annotations that our program fixed. We report the results in the form of several standard metrics in image segmentation: the mean and frequency-weighted intersection-over-union (fwIoU, mIoU), the F_1-measure and the average precision. In order to get a better idea of the data distribution and model generalisation, we report the results as means and standard deviations over all the images in the testing part of the dataset.

Table 1: Comparison of accuracy measurements, reported as mean $\pm$ standard deviation over the images in the test dataset.

Measurement	**Old** ($\mu \pm \sigma$)	**New** ($\mu \pm \sigma$)
mIoU	0.797 ± 0.159	0.835 ± 0.147
fwIoU	0.919 ± 0.041	0.936 ± 0.031
F_1-measure	0.771 ± 0.13	0.814 ± 0.119
Average precision	0.755 ± 0.17	0.827 ± 0.12

As we can see from Table 1, the deep model trained and evaluated on the fixed annotations achieved better results with smaller standard deviations in all metrics than the same deep model trained and evaluated on the original annotations.

4 Conclusion

Our program successfully removed several different types of distortions present in the manual annotations in the MOBIUS dataset. The evaluation results show the model trained on the corrected annotations achieved better overall performance (higher metric means), as well as more consistent performance across different images (lower metric standard deviations). This implies that the new annotations are more consistent across the entire dataset and allow for better training of the model, as well as for better generalisation to previously unseen images. Our approach therefore resulted in fixed annotations that noticeably improved the quality of the MOBIUS dataset.

References

[1] Peter Rot, Žiga Emeršič, Vitomir Štruc, and Peter Peer. Deep multi-class eye segmentation for ocular biometrics. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–8, 07 2018.

[2] Peter Rot, Matej Vitek, Klemen Grm, Žiga Emeršič, Peter Peer, and Vitomir Štruc. Deep sclera segmentation and recognition. In A. Uhl, C. Busch, S. Marcel, and R. Veldhuis, editors, *Handbook of Vascular Biometrics*, pages 395–432. Springer, 2020.

[3] Peter Rot, Matej Vitek, Blaž Meden, Žiga Emeršič, and Peter Peer. Deep periocular recognition: A case study. In *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 000021–000026. IEEE, 2019.

[4] Abhijit Das, Umapada Pal, Miguel Ferrerc, and Michael Blumensteina. Ssbc 2015: Sclera segmentation benchmarking competition. pages 1–6, 09 2015.

[5] Matej Vitek, Peter Rot, Vitomir Štruc, and Peter Peer. A comprehensive investigation into sclera biometrics: A novel dataset and performance study. *Neural Computing & Applications*, pages 1–15, 2020.

[6] V. Patil and A. M. Patil. Human identification method: Sclera recognition. *International Journal of Computer Science and Network (IJCSN)*, 6(1):24–29, 2017.

[7] Abhijit Das, Umapada Pal, Miguel Ferrer, and Michael Blumenstein. Ssrbc 2016: Sclera segmentation and recognition benchmarking competition. pages 1–6, 06 2016.

[8] Abhijit Das, Umapada Pal, Miguel Ferrer, Michael Blumenstein, Dejan Štepec, Peter Rot, Žiga Emeršič, Peter Peer, Vitomir Štruc, S V Aruna kumar, and B S Harish. Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition. pages 742–747, 10 2017.

[9] Abhijit Das, Umapada Pal, Miguel Ferrer, Michael Blumenstein, Dejan Štepec, Peter Rot, Žiga Emeršič, Peter Peer, and Vitomir Štruc. Ssbc 2018: Sclera segmentation benchmarking competition. pages 303–308, 02 2018.

[10] Abhijit Das, Umapada Pal, Michael Blumenstein, and zhun sun. Sclera segmentation benchmarking competition in cross-resolution environment. 06 2019.

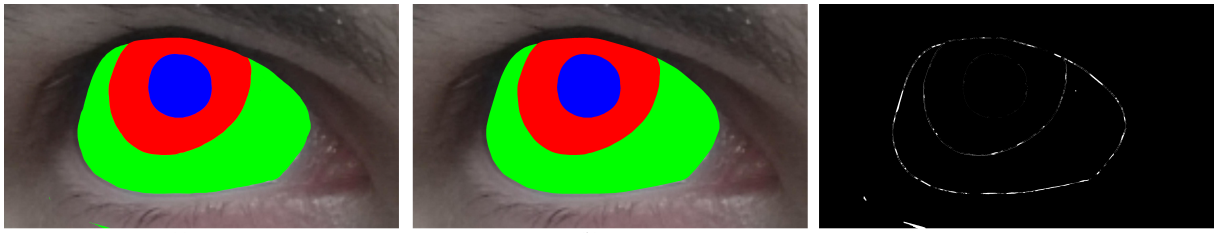
[11] Daniel Riccio, Nadia Brancati, Maria Frucci, and Diego Gragnaniello. *An Unsupervised Approach for Eye Sclera Segmentation*, pages 550–557. 01 2018.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

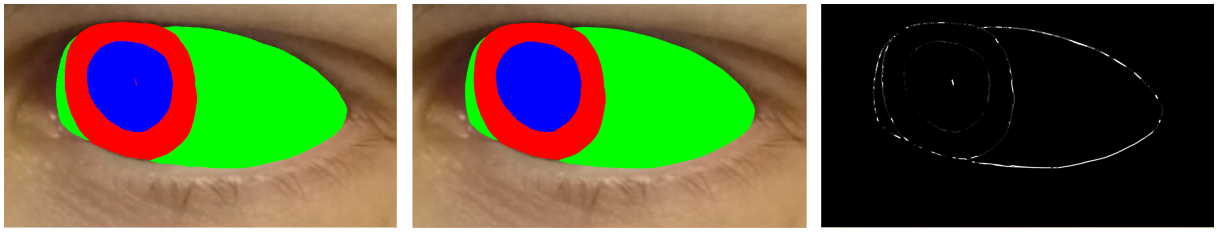
[14] C. Wang, Y. He, Y. Liu, Z. He, R. He, and Z. Sun. Scleraseg net: an improved u-net model with attention for accurate sclera segmentation. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[15] Aayush K Chaudhary, Rakshit Kothari, Manoj Acharya, Shusil Dangi, Nitinraj Nair, Reynold Bailey, Christopher



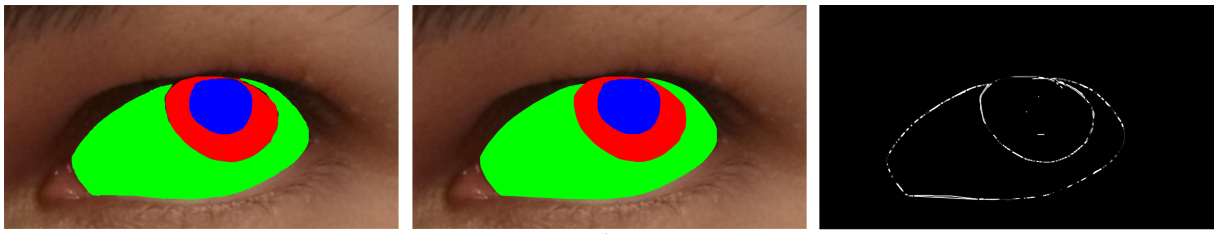
(a) (b) (c)

Figure 4: Annotation with green outliers (best viewed with zoom online). (a) Original annotation. (b) Fixed annotation. (c) Differences.



(a) (b) (c)

Figure 5: Annotation with red inliers (best viewed with zoom online). (a) Original annotation. (b) Fixed annotation. (c) Differences.



(a) (b) (c)

Figure 6: Annotation with missing edges (best viewed with zoom online). (a) Original annotation. (b) Fixed annotation. (c) Differences.

Kanan, Gabriel Diaz, and Jeff B Pelz. Ritnet: Real-time semantic segmentation of the eye for gaze tracking. *arXiv preprint arXiv:1910.00694*, 2019.

[16] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[18] Ishan Nigam, Mayank Vatsa, and Richa Singh. Ocular biometrics: A survey of modalities and fusion approaches. *Information Fusion*, 26:1 – 35, 2015.

[19] Matej Vitek, Abhijit Das, Yann Pourcenoux, Alexandre Missler, Calvin Paumier, Sumanta Das, Ishita De Ghosh, Diego R. Lucio, Luiz A. Zanlorensi Jr., David Menotti, Fadi Boutros, Naser Damer, Jonas Henry Grebe, Arjan Kuijper, Junxing Hu, Yong He, Caiyong Wang, Hongda Liu, Yunlong Wang, Zhenan Sun, Daile Osorio-Roig, Christian Rathgeb, Christoph Busch, Juan Tapia Farias, Andres Valenzuela, Georgios Zampoukis, Lazaros Tsochatzidis, Ioannis Pratikakis, Sabari Nathan, R Suganya, Vineet Mehta, Abhinav Dhall, Kiran Raja, Gourav Gupta, Jalil Nourmohammadi Khiarak, Mohsen Akbari-Shahper, Farhang Jaryani, Meysam Asgari-Chenaghlu, Ritesh Vyas, Sristi Dakshit, Sagnik Dakshit, Peter Peer, Umapada Pal, and Vitomir Štruc. SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2020.

[20] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Measuring the induction of affect using facial expression analysis technology: a pilot study

Marko Meža[1], Andrej Košir[1], Gregor Strle[1,2]

[1]*User-adapted Communication and Ambient Intelligence Lab, Faculty of Electrical Engineering, University of Ljubljana*
[2]*Research Centre of the Slovenian Academy of Sciences and Arts*
E-pošta: marko.meza@fe.uni-lj.si

Abstract. *The paper presents a pilot study investigating usefulness of facial expression analysis technology for measuring emotion induction. The study involved six male participants and a subset of 85 images from the International Affective Picture System dataset (IAPS), rated on the valence and arousal dimensions of affect.*

The results show low reliability and agreement, both when comparing the mean induction ratings gathered from the face reader to the ground truth (the IAPS ratings), as well as when comparing the induction ratings between the six participants. We conclude that the presented setup does not adequately measure the induction of affect and that the time interval for measuring participant responses from the stimulus onset should be carefully selected.

1 Introduction

Non-verbal cues, particularly affect and emotions, are critical to maintaining, interpreting and understanding social signals, communication, and behavior among humans [1]. One of the challenges of the current state-of-the-art is the analysis and prediction of affective cues using sensor technology and machine learning in different setting, from researches in the wild or based on the existing datasets [2–6].

One area of research in our laboratory is a development of sensor technology framework that would support reliable measurements of emotional, behavioral, and social cues. To this end, the paper presents a pilot study investigating usefulness of facial expression analysis technology for measuring emotion induction.

2 Method

2.1 Participants

The pilot study was conducted on 6 young male participants enrolled in the curriculum at the The Faculty of Electrical Engineering, University of Ljubljana. The tasks given to the participants were a part of a laboratory assignment created to investigate how affect can be measured with sensor technology.

2.2 Materials

Materials from the International Affective Picture System dataset [5] were used in the study. The IAPS dataset is a widely used stimulus set for emotion research, based on the dimensional model of affect [7]. It contains 1194 images with short descriptions, rated along the dimensions of valence and arousal. To make this pilot study feasible, a subset of 85 images was chosen by manually selecting samples with high, medium and low arousal ratings. In order to achieve this, we ordered the database in the descending order of arousal ratings and selected top, medium and bottom part of the samples for each group from the database. Some images were further removed due to their explicit content.

2.3 Recording apparatus

The recording apparatus consisted of a laptop computer with an image playback software, a real-time face expression analysis software, and a control script.

The *image playback application* (Figure 1) was developed for the purpose of this study and used to present the participant the 85 IAPS images, in the same order, one after another, with each image presented for 5 seconds. The application stores exact timestamps of when each individual image was shown to the user.

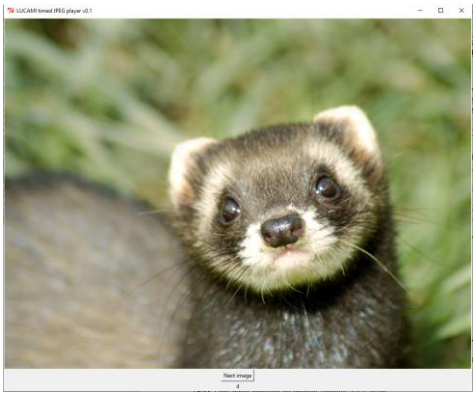


Figure 1. Image playback application showing an IAPS image from the test session.

Noldus FaceReader [8], a *facial expression analysis software*, was used for measuring the induction of affect. The Noldus face reader computes facial emotion expressions in real time, from the video recording of the participant. The software records the valence and arousal, as well as timestamps (Figure 2). For our setup, we recorded between 10 and 12 data samples per second. The Noldus API was used to control the recording sessions and analysis of the data. A *data logging application* was developed in the course of the study to obtain the face reader data and store it into a .csv file for further analysis.

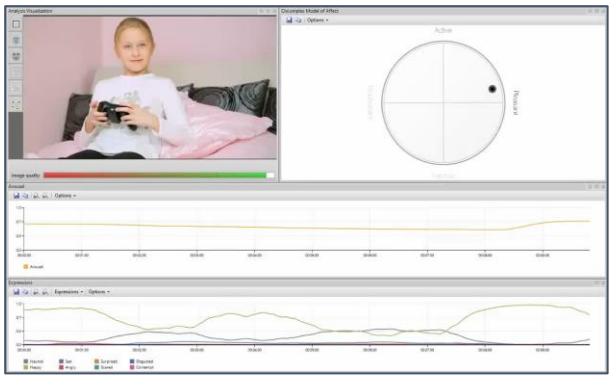


Figure 2. A screenshot of Noldus Face reader software

All components of the recording setup were controlled by a single *control script*. The script started the Noldus face reader software, initiated the analysis, started the data logging application and the image playback application, as well as finished the recording sessions by terminating all software components used during the sessions.

The output for each recording session includes two .csv files, one containing the timestamps of the presented images and the other containing the face reader timestamps and valence and arousal ratings sampled approximately 10 times per second.

2.4 Recording procedure

The participants were first informed about the aim and details of the pilot study and the experimental procedure. They were instructed to act normally and to not exaggerate and play their expressions. Each participant was subjected to a test session, during which they became accustomed to the recording apparatus and the experiment procedure. After the test trial, each participant was set in a quiet environment behind a laptop computer used for the experiment. The participant was given time to relax and told to start the recording session when ready. After each session, the recorded data were first archived, then the experimental setting was prepared for the next participant.

2.5 Statistical analysis

The IAPS ratings were first normalized to correspond to the valence-arousal measuring scale (-1,1) used by the Noldus face reader. For the Noldus face reader data, the mean valence and arousal ratings were taken for each IAPS image the participant was exposed to, in order to assess the reliability among the participants. The overall mean valence and arousal Noldus ratings were also calculated for each IAPS image, in order to evaluate the reliability of the face reader ratings with the ground truth (IAPS ratings).

Independent Samples *t*-test (Welch's *t*-test for unequal variances) was used to test the means of both groups of ratings (IAPS vs. Noldus), and a one-way Anova was used to test the difference between the face reader participants for the valence and arousal dimensions. The intraclass correlation (ICC) was used to assess the reliability of the ratings, both between the IAPS (the ground truth) and the Noldus ratings, as well as between the participants using the Noldus face reader.

3 Results

A comparison of the mean ratings from IAPS and the mean ratings from the Noldus face reader shows significant differences, both for valence and arousal. The IAPS ratings for the valence (M = 0.21, SD = 0.43) are significantly different than the Noldus ratings (M = 0.06, SD = 0.03), with t = 3.19, p < .002. Similarly, significant difference was found for the arousal, between the IAPS ratings (M = 0.01, SD = 0.34) and the Noldus ratings (M = 0.32, SD = 0.01), with t = 8.3, p < .001.

The significant difference in the distribution of the ratings between both groups is shown also in Figures 4 and 5. The low value of the ICC shows there is no agreement in the ratings between the IAPS and the Noldus face reader. The $ICC(1,1)$= -0.004 for valence is extremely poor, with the 95% confidence interval for the ICC (-0.2, 0.2). Similarly, the agreement for the arousal is also extremely poor, with the $ICC(1,1)$= -0.27 and the 95% confidence interval for the ICC (-0.45, 0.06).

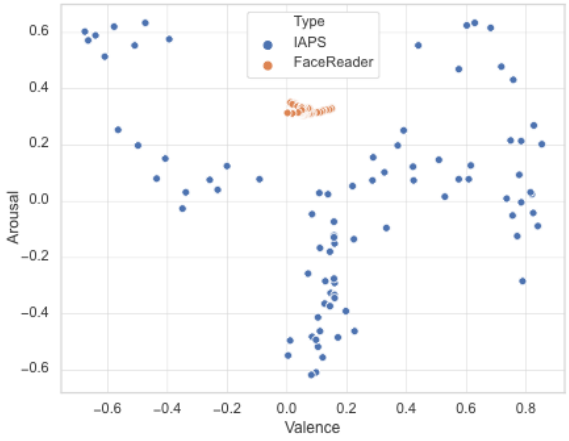


Figure 4. Valence-arousal space of the mean IAPS (blue) and face reader ratings (orange).

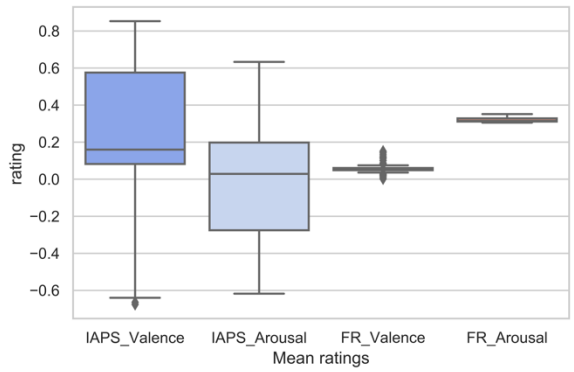


Figure 5. Distribution of mean valence/arousal ratings for IAPS and Noldus (FR).

The agreement is also low between the Noldus participants. The *ICC*(2,1)= 0.04 for valence is extremely poor, with the 95% confidence interval for the ICC (0.002, 0.08). Similarly, for the arousal, with the *ICC*(2,1)= 0.02 and the 95% confidence interval for the ICC (-0.002, 0.05). The one-way Anova on the Noldus participants' ratings shows significant differences among the participants for both valence and arousal, with $p < .001$. The differences in valence and arousal ratings between the participants are shown in Figure 6. We can observe that the valence ratings are consistently lower and with several outliers, as compared to the arousal ratings.

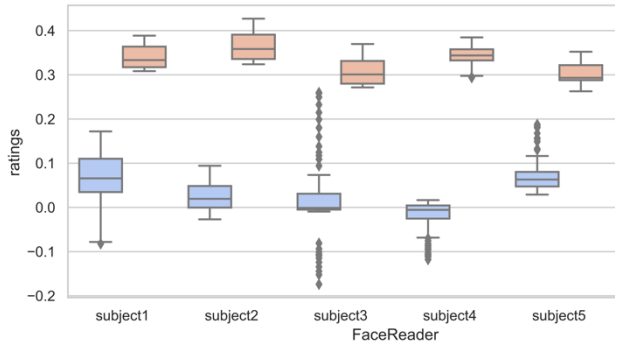


Figure 6. Distribution of the participants' mean valence/arousal ratings (Noldus face reader).

4 Discussion

The presented pilot study investigated the induction of affect as measured by the Noldus face reader. The results show low reliability and agreement, both compared to the ground truth (the IAPS dataset), as well as among the participants of the study.

For the presented setup, the results generated by the Noldus face reader do not adequately measure the emotional induction from images (e.g., see Figures 4 and 5). We speculate the main reason for this are the mean ratings taken of the entire time interval each image was presented to the participant, as the distributions of the valence and arousal ratings are small for all the participants. It might be more appropriate to sample the ratings in a narrower time interval starting from the onset an image is shown, as the face reader recordings show higher values for the initial responses to the stimuli, for both dimensions, which then decrease over the five second timespan.

It is also important to note that, in part, the ratings are influenced by the individual's personality and their facial expressions which vary among the participants, with some participants being more expressive than others. However, as shown by the results (see also Figures 4 and 5), the low reliability and agreement with the ground truth (IAPS) cannot be ascribed to differences in emotional expressivity among the six participants, as the distribution of the ratings is extremely low.

In our future work, the presented study will be extended with a larger sample and, particularly, with in-depth analysis of the time interval that could potentially better capture the induction of affect.

Literature

[1] Vinciarelli A, Pantic M, Bourlard H. Social signal processing: Survey of an emerging domain. Image and vision computing. 2009 Nov 1;27(12):1743-59.

[2] Sawyer R, Smith A, Rowe J, Azevedo R, Lester J. Enhancing student models in game-based learning with facial expression recognition. InProceedings of the 25th conference on user modeling, adaptation and personalization 2017 Jul 9 (pp. 192-201).

[3] Mollahosseini A, Hasani B, Mahoor MH. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing. 2017 Aug 21;10(1):18-31.

[4] Kollias, Dimitrios, et al. "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond." International Journal of Computer Vision 127.6-7 (2019): 907-929.

[5] Lang, PJ.; Bradley, MM.; Cuthbert, BN. International affective picture system (IAPS): Technical manual and affective ratings. University of Florida, Center for Research in Psychophysiology; Gainesville: 1999.

[6] Stöckli, S.; Schulte-Mecklenbeck, M.; Borer, S. & Samson, A.C. Facial expression analysis with AFFDEX and FACET: A validation study. Behavior Research Methods, 50 (4), (2018): 1446-1460.

[7] Russell JA. A circumplex model of affect. Journal of personality and social psychology. 1980 Dec;39(6):1161.

[8] Noldus FaceReader: Tool for automatic analysis of facial expression: Version 6.0. Wageningen, the Netherlands: Noldus Information Technology B.V. (2014).

Semantic Face Editing with GAN Latent Code Optimization

Martin Pernuš, Vitomir Štruc, Simon Dobrišek

Faculty of Electrical Engineering, University of Ljubljana
E-pošta: {martin.pernus, simon.dobrisek}@fe.uni-lj.si

Abstract

Recent advances in generative modeling of images have reached new heights in the task of image generation. The generative model that produces the most photorealistic and high resolution images is Generative Adversarial Network (GAN). In our paper we propose a multi-task latent vector optimization procedure that combines local GAN inversion with face attribute constraints for the task of semantic face editing. The experiments show that the method produces visually pleasing and semantically corresponding face images.

1 Introduction

With the advancements in generative modelling with Generative Adversarial Networks (GANs) [5], image generation has achieved unprecedented image photorealism. The latent space of GAN models provides a compressed representation of high-resolution image, which can be useful for semantic image manipulation. By manipulating the latent codes of GAN generated images, several useful image editing applications have been proposed [1, 2, 19]. The advancements in this field could have an important impact toward automatic image editing tasks. Such algorithms could enable image manipulation by simply specifying the desirable visual attributes. This could have an important impact in art, media and entertainment industry.

The process of obtaining a latent code of an image and performing various operations has been popularized by the variational autoencoder [12]. Research showed that combining latent codes produced a learned image manifold that did not just correspond to pixelwise averaging, but had a sense of the intrinsic data distribution. Since the GAN architecture lacks a direct way to obtain an image's latent code, GAN inversion optimization method [1, 2] is usually applied.

Generative modeling and image editing tasks have been largely focused on face modeling and editing as generative modeling relies on vast quantities of training examples. For example, state-of-the-art StyleGAN model [9] is trained on 70,000 high resolution face images. Several studies used the pretrained StyleGAN model to edit face images. Our work is inspired by the work of [19], where the latent code vector space of GANs was analyzed and linear movements in vector space were proposed to edit face images. Instead of analyzing latent space, we directly perform gradient-based optimization on latent code with spatial and semantic constraints.

Although studies have analyzed latent code manipulation for various semantic image operations, disentanglement of semantic variations of latent codes remains challenging. In our work, we propose to use a local GAN inversion technique to preserve the person's identity while modifying the selected face attributes. We test the proposed method and visually analyze the results in comparison with a competing method.

2 Related Work

Generative Adversarial Networks (GANs) are one of the most used generative models for image modelling task. Since the original proposal in [5], the majority of advances stem from better architectures and loss functions. The architecture design was improved in [18], which proposed convolutional GAN design. Karras et al. [10] first managed to produce megapixel images using progressive learning of GANs. The model was further improved in StyleGAN [8, 9], where the architecture is inspired by style transfer architecture, resulting in state-of-the-art unconditional image generation. Loss function beyond the one proposed in the vanilla GAN model were explored in [6, 14, 16, 17].

Due to resource intensive training of GAN models, recent research focuses on the analysis of pretrained GAN models. Pretrained parametric space of GAN weights was manually tuned in [3] to achieve localized deletion and addition of objects in the output image. In [7] linear and non-linear walks in latent space were learned that achieved some basic image manipulation, such as brightness and zoom change. In [19], linear subspaces of latent facial semantics were identified to edit face images.

These methods are based on GAN generated images. GAN models are inherently constrained by the probability distribution of the data they are trained on, which limits the usability of latent codes for image editing. Abdal et al. [1, 2] projected face images using GAN inversion technique in an extended latent space before editing face images in various ways. In our work, we also use a version of GAN inversion with an extended latent space, but we also consider various spatial and semantic constraints.

3 Methods

Our method is based on StyleGAN model [9]. StyleGAN is the current state-of-the-art GAN method for unconditional image generation. We use the pretrained StyleGAN model trained on Flickr-Faces-HQ dataset [8].

StyleGAN's generator is defined by two main components: initial latent code non-linear mapping and the generator G that generates the final image. Non-linear latent code mapping is defined as $f : \mathcal{Z} \rightarrow \mathcal{W}$ that maps the initial Gaussian sampled latent code z to code w of the same 512-d dimensionality. The purpose of mapping f is the disentanglement of factors of variation that could be present in the Z vector space due to its Gaussian distribution. The generator G maps latent space $\mathcal{W}$ to images. The w codes are mapped by 18 learned affine transforms as an input to the convolutional layers of the generator. Each convolutional layer also receives a stochastic component in the form of spatial noise $n \in \mathcal{N}_S$ that slightly affects the output face image. See [9] for more details about the model.

The goal of GAN inversion is retrieving the latent code w and optionally n that best matches the image I given pretrained generator G. As shown in [2], a wide variety of images, including non-face images, can be embedded in the extended latent space W^+. The extended latent space W^+ consists of a concatenation of 18 different 512-dimensional w vectors. We also optimize the noise component of StyleGAN n.

To achieve the presence of selected face attribute on the final image, we also introduce a pretrained classifier C that predicts the presence of the selected facial attribute. The predicted probability of the selected attribute is denoted as $\hat{y}(w, n) = C(G(w, n))$.

Starting from a suitable initialization of w and n, we search for their optimized versions w^* and n^*. We only optimize a subset of noise n, the portion that affects the face portion of the image. Our loss function is defined as

$$\begin{aligned}\mathcal{L}(w, n) =& \lambda_{\text{mse}} ||M \odot (G(w, n) - I)||_2^2 + \\ & \lambda_{\text{ce}}(-y \log \hat{y}(w, n) - (1 - y) \log(1 - \hat{y}(w, n)))\end{aligned} \tag{1}$$

where M is the spatial mask that defines the region where mean squared error part of the loss is defined, I is the target image, λ_{mse} is the weighting constant for mean squared error loss and λ_{ce} is the weighting constant for the binary cross entropy loss.

A special consideration must be taken with regard to the spatial mask M. It must be large enough to preserve enough face in the image to retain the face identity, while also allowing the rest of the image to change enough to satisfy the selected face attribute constraint. We define M with face segmentation model that is based on the DeepLabv3 segmentation model [4]. We trained the model to spatially predict the face components of interest on the image I based on the selected facial attribute. We then binarize the prediction based on predefined threshold and blur it with Gaussian filter to allow smooth changes to the face image.

4 Experiments

4.1 Implementation

The classifier C is based on the state-of-the-art multi-task neural architecture [20] and is trained on CelebA dataset [15] for 23 epochs. The starting learning rate is 0.05 and it decays to one tenth its current value every 40,000 steps. The CelebA dataset contains 200,000 celebrity images with 40 annotated facial attributes per image as well as their identity information.

The face segmentation model is set as DeepLabv3 model [4] that is trained on CelebAMaskHQ dataset [13]. This dataset contains 30,000 images with the size of 512×512 and 19 facial components and accessories such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth. Each image is annotated with a segmentation mask of facial components. After grouping several semantically similar attributes, we ended up with 'background', 'mouth', 'eyebrows', 'eyes', 'earrings', 'hair', 'nose' and 'skin' facial components that we use to train the face segmentation model. The model is trained for 5 epochs with a learning rate of $3 \cdot 10^{-4}$.

The latent code w is initialized with the latent code mean, calculated by passing 10,000 Gaussian sampled z vectors and calculating the mean of the StyleGAN's feedforward network. The optimization of w^* and n^* is run for 2000 iterations. The learning rate is first linearly increased from 0 to 10^{-2}, then decayed back to 0 using cosine schedule during the last 500 iterations. The threshold for binarizing the image is set to 0.9 and the size and the standard deviation of Gaussian filter are set to 51 and 31, respectively. λ_{mse} is set to 100 and λ_{ce} is set to 10.

The optimization algorithm of choice for all the models is the Adam optimization method [11].

4.2 Evaluation

For comparison with the existing face editing techniques, we implemented the Interface method [19]. The main idea behind the Interface method is finding a hyperplane that optimally separates latent space based on support vector machine and moving the latent code in the selected direction. It operates in W space, which guarantees that latent code will generate a face image. However, W space is often lacking when performing GAN inversion. Thus, the reconstructions do not match the original image as well as they would with W^+ space.

To find the individual facial attribute hyperplane we followed the Interface method procedure. First we generated 500,000 StyleGAN images. We picked the 10,000 most positive and 10,000 most negative images per facial attribute according to our classifier C. The hyperplane was then identified using linear support vector machine on the latent codes of these images. For visualization we move the latent codes for 0.5 the norm of the latent vector as opposed to 3 times the norm in the original proposal, since in our experiments the face images failed to preserve any identity information in the latter case.

In Figure 1 we show the results of the compared Interface method and our proposed method for several facial

attributes. The results visually indicate that our proposed method retains the identity information better than the Interface method.

The advantages of our method in comparison with InterFace method can be summarized as better visual results and no pretraining requirement. However, once the InterFace learns the optimal hyperplane, its method allows immediate calculation of new latent vectors (and new face images), while our method requires optimization of latent code. Our optimization procedure takes approximately 10 minutes.

5 Conclusion and future work

In this paper, a method for changing facial attributes while preserving identity is proposed. The method is based on local GAN inversion technique with facial attribute constraints. Experimental results visually suggest that real face images can be modified according to selected attributes while preserving their identity.

Additional improvements could be made using disentanglement techniques, which could modify latent space in a way that changing a single facial attribute wouldn't affect other attributes, while allowing the pose of the person to change. That is the topic of our further research.

Acknowledgments

This work was supported in parts by the ARRS Research Programmes P2-0250 (B) Metrology and Biometric Systems, the ARRS young researcher program and the ARRS research project J2-2501 - Deep generative models for beauty and fashion (DeepBeauty).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. "Image2StyleGAN++: How to Edit the Embedded Images?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8296–8305.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. "Image2stylegan: How to embed images into the stylegan latent space?" In: *Proceedings of the IEEE international conference on computer vision*. 2019, pp. 4432–4441.

[3] David Bau et al. "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2018.

[4] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[5] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[6] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems*. 2017, pp. 5767–5777.

[7] Ali Jahanian, Lucy Chai, and Phillip Isola. "On the" steerability" of generative adversarial networks". In: *International Conference on Learning Representations*. 2019.

[8] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4401–4410.

[9] Tero Karras et al. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8110–8119.

[10] Tero Karras et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations*. 2018.

[11] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[12] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[13] Cheng-Han Lee et al. "Maskgan: Towards diverse and interactive facial image manipulation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 5549–5558.

[14] Jae Hyun Lim and Jong Chul Ye. "Geometric gan". In: *arXiv preprint arXiv:1705.02894* (2017).

[15] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[16] Xudong Mao et al. "Least squares generative adversarial networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2794–2802.

[17] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. "f-gan: Training generative neural samplers using variational divergence minimization". In: *Advances in neural information processing systems*. 2016, pp. 271–279.

[18] A. Radford, L. Metz, and S. Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[19] Yujun Shen et al. "Interpreting the latent space of gans for semantic face editing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9243–9252.

[20] Simon Vandenhende et al. "Branched multi-task networks: deciding what layers to share". In: *arXiv preprint arXiv:1904.02920* (2019).

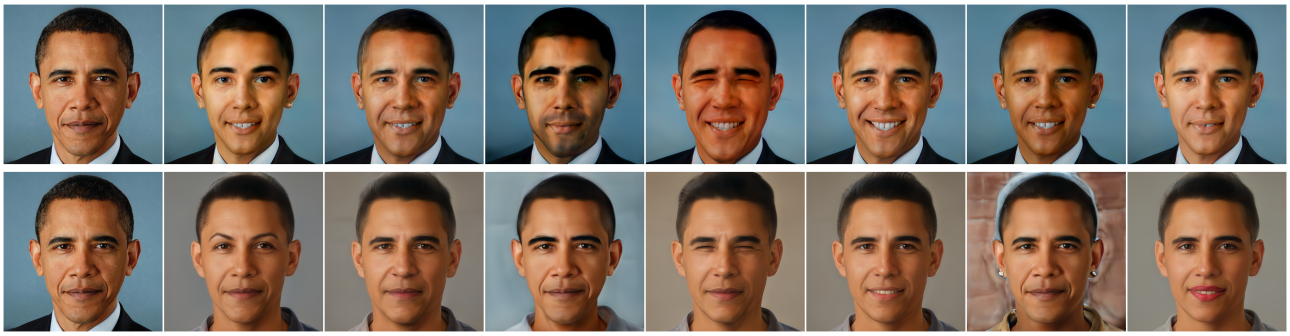


Figure 1: Comparison of InterFace [19] performance (first row) to our method (second row) on several face editing tasks for an example face image. The selected facial attributes from left to right are defined as follows: original image, arched eyebrows, big nose, bushy eyebrows, narrow eyes, smiling, wearing earring and wearing lipstick.



Figure 2: Comparison of InterFace [19] performance (first row) to our method (second row) on several face editing tasks for an example face image. The selected facial attributes from left to right are defined as follows: original image, blond hair, brown hair, gray hair, straight hair, wavy hair.

O klasifikaciji slik v ne-enolično določljive razrede

Jon Natanael Muhovič, Domen Tabernik, Danijel Skočaj

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
E-pošta: {jon.muhovic,domen.tabernik,danijel.skocaj}@fri.uni-lj.si

About classification of images into non-uniquely identifiable classes

Image classification is one of the most basic and frequently addressed computer vision tasks. The usual formulation of this tasks requires classification of an image into the one of several possible classes. The most common metric for measuring the classifier's performance is classification accuracy, defined as a percentage of correctly classified images. However, such formalisation of the classification problem relies on a strong assumption that for every image a category is uniquely identifiable and assigned by the domain expert. In this paper we address scenarios where this assumption does not hold. In particular, we present an analysis of the results obtained by the convolutional neural network and twelve participants who were tasked to classify the images of planks into eight classes and discuss the label ambiguity problem.

1 Uvod

Klasifikacija slik je ena najbolj osnovnih in uporabljanih funkcionalnosti v računalniškem vidu. Zelo pogosto se namreč srečamo s problemom, ko je potrebno razpoznati kaj je prikazano na sliki, oz. prikazano uvrstiti v primeren razred. Zato se raziskovalci s tem problem ukvarjajo od najzgodnejših dni računalniškega vida naprej.

1.1 Definicija problema

Običajno je problem klasifikacije formuliran na zelo enostaven način. Vsebino slike je potrebno klasificirati natanko v enega izmed n razredov. Na podoben način nato merimo tudi uspešnost klasifikatorja. Uveljavljena mera je klasifikacijska točnost, ki predstavlja odstotek slik, ki so bile pravilno klasificirane. Včasih to zahtevo celo malce omilimo in smatramo klasifikacijo kot uspešno, če klasifikator vrne pravilni razred med prvimi k napovedmi. Pri tem seveda tudi predpostavimo, da je kategorija za vsako sliko oz. njeno vsebino enolično določljiva ter pravilno določena s strani poznavalca problema. Celoten postopek učenja klasifikatorja ter njegove evalvacije sloni na tej predpostavki.

Pri realnih problemih pa ta predpostavka ne drži vedno. Nekatere kategorije so dvoumne, nekatere vsebujejo druge kategorije, nekatere slike se lahko uvrstijo v več kategorij ali pa različni uporabniki uvrstijo isto sliko v različne kategorije. To predstavlja pri klasičnih pristopih učenja in evalvacije dodaten izziv. V tem članku bomo obravnavali problem ne-enolično določljivih razredov na primeru industrijske aplikacije klasifikacije desk. Naloga bo klasificirati deske v enega izmed osmih kakovostnih razredov. Kot je razvidno s Slike 1, je izgled desk, ki pripadajo posameznim razredom, dokaj podoben. Razredi se med seboj razlikujejo v podrobnostih, kot sta odtenek in homogenost barve ter velikost in pogostost grč. Meje med razredi je zelo težko enolično postaviti. Zato tudi ni vseeno kakšne napake sistem dela. Včasih je popolnoma sprejemljivo, če je slika klasificirana v enega izmed dveh sosednjih razredov, medtem ko klasifikacija v bolj različen razred prestavlja veliko večjo napako.

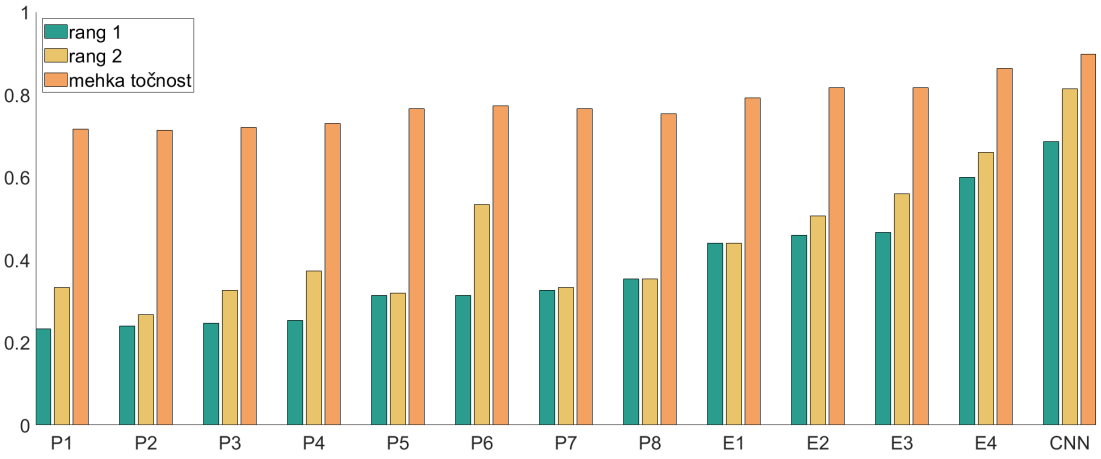


Slika 1: Predstavniki osmih razredov desk.

V tem članku bomo obravnavali ta problem. Analizirali bomo rezultate, ki smo jih dobili s klasifikatorjem, ki temelji na globokem učenju, in rezultate klasifikacij dvanajstih posameznikov, tako poznavalcev dotične aplikativne domene, kot tudi laikov, ter jih analizirali z vidika ne-enolične določljivosti razredov.

1.2 Sorodna dela

Ne-enolična določljivost razredov se v literaturi pogosto naslavlja z uporabo mehkih označb razredov [5] (ang. soft labels) ali z glajenjem označb razredov [6] (ang. label smoothing). V obeh primerih enični vektor (ang. one-hot vector), s katerim označujemo kategorijo, zamenja drugačna označba, in sicer mehke označbe preko več kategorij oz. verjetnostna porazdelitev uvrstitve v posamezne razrede, ki odraža podobnost oz. ne-enoličnost kategorij. V primeru glajenja razredov se običajno uporabi enakomerna porazdelitev med razredi, ki izboljša klasifikacijsko točnost. Kot so pokazali Müller in ostali [4], to povzroči zoženje posamezne gruče razredov v podprostoru značilk, vendar pa to ni povsem zaželeno v prime-



Slika 2: Rezultati klasifikacij.

rih, ko se ne moramo zanašati na točnost podane referenčne označbe oz. zlatega standarda. V takih primerih se za bolj primerne izkažejo mehki razredi, kjer se enični vektor zamenja s porazdelitvijo, ki pravilno odraža nedoločenost med posameznimi razredi. Ta porazdelitev se nato uporabi v razširjeni kriterijiski funkciji prečne entropije [5], ki skuša klasifikator zgraditi na način, da se njegove napovedi čim bolj ujemajo distribucijami pravilnih napovedi posameznih slik. Pristop se pogosto uporablja pri klasifikaciji splošnih objektov, kjer je mogoče nedoločenost med kategorijami definirati na podlagi semantične hierarhije [1].

Ti pristopi se torej osredotočajo na mehkejše napovedovanje razredov, ki ustreza dvoumnosti med razredi v skladu z dvoumnostjo napovedi ljudi. Problem, s katerim se ukvarjamo v tem članku, pa je rahlo drugačen. Osredotočamo se namreč na industrijsko aplikacijo, kjer je pomembna zgolj ena napoved, ki se upošteva v nadaljnjem proizvodnem procesu, in ne celotna distribucija napovedi. Se pa sorodnost med posameznimi razredi izraža pri evalvaciji napovedi. S stališča proizvodnega procesa je namreč veliko manj izrazita napaka uvrstitve slike v soroden razred, torej v razred, ki ima podano dokaj visoko (čeprav ne najvišjo) verjetnost, kot v razred, kjer je ta verjetnost, oz. ustreznost, majhna.

2 Metode

Glavni namen te raziskave je raziskati vpliv dvoumnosti pri določevanju razredov slik na rezultate klasifikacije. Zato smo izvedli študijo z 12 udeleženci in rezultate primerjali z napovedmi avtomatskega klasifikatorja.

Za eksperimentalno evalvacijo smo pridobili osem laikov in štiri eksperte. Vsak od njih je moral za vsako od slik v zbirki določiti razred. Zbirka je obsegala 150 primerov desk, od tega jih je bilo 100 naključno izbranih iz celotne zbirke slik desk, preostalih 50 pa je bilo naključno izbranih iz prvih 100 in naključno premešanih. Celotna zbirka je tako vsebovala eno tretjino ponovljenih slik, oz. 50 parov duplikatov, na kar udeleženci niso bili opozorjeni.

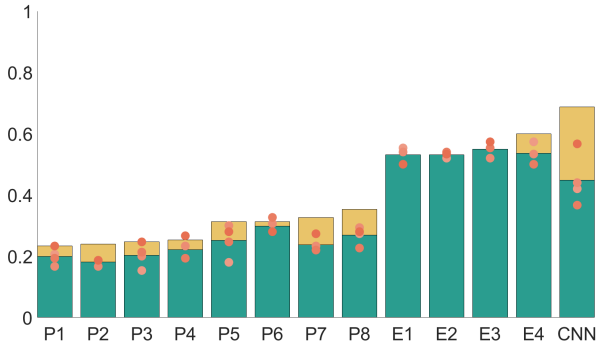
Udeleženci so za pomoč pri klasifikaciji dobili 14 primerov desk iz vsakega razreda. Za vsako od slik v eksperimentalni zbirki so morali udeleženci določiti razred, za katerega so menili, da mu posamezna slika pripada. V primeru negotovosti so lahko določili še drugi razred, za katerega so menili, da bi mu posamezna slika lahko pripadala.

Za avtomatsko klasifikacijo slik smo uporabili arhitekturo ResNet50 [3], prednaučeno na zbirki ImageNet [2]. Mrežo smo učili za klasifikacijo vsakega od osmih razredov desk, za učenje pa smo uporabili našo podatkovno zbirko z več kot 2500 slikami, ki ni vsebovala slik, ki so bili uporabljene v evalvacijski študiji. Za učenje klasifikacije smo uporabili standardno kategorično križno entropijo in optimizator ADAM.

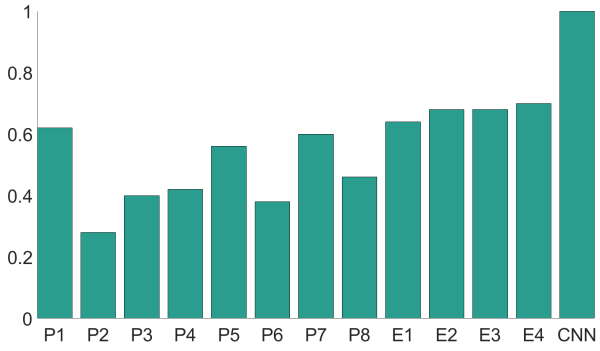
Metrike: Uspešnost rezultatov smo merili s štirimi metrikami:

(i) *rang 1*, ki meri klasifikacijsko točnost s prvo napovedjo,
(ii) *rang 2*, ki meri uspešnost klasifikacije s prvo ali z drugo napovedjo,
(iii) *mehka točnost*, kjer se nepravilni odgovori obtežijo v odvisnosti od kritičnosti napake; pri tem se klasifikacija v sosednje razrede veliko manj kaznuje kot klasifikacija v zelo različne razrede,
(iv) *ponovljivost*, ki izračuna odstotek istih testnih primerov, ki so bili obakrat enako klasificirani.

Poleg standardne klasifikacijske točnosti, ki se običajno uporablja pri evalvaciji klasifikacijskih rešitev, smo torej uporabili tudi mehko klasifikacijsko točnost. Ta bolj realno zajame uspešnost metode, saj ne obravnava vseh napak kot enako problematične. To je v praksi pogosto potrebno, saj vse napake nimajo enake teže, kar predpostavlja klasična klasifikacijska točnost. Težo napake uravnavamo z vpeljano matriko podobnosti, ki napačni klasifikaciji pripiše ne-ničelno ceno, glede na podobnost pravilnega razreda. Mehko klasifikacijsko točnost tako računamo kot $SCA = \frac{1}{N}\sum_{i=1}^{N} w_{c_i,y_i}$, kjer je N število vseh slik, $w_{a,b}$ pa predstavlja utež za primer, ko je napoved klasifikacije slike iz razreda a, razred b. Ta utež je velika (blizu 1), če sta si razreda podobna in 0, če slika nikakor ne bi smela biti klasificirana v ta razred. V primeru standardne klasifikacijske točnosti je matrika W kar



Slika 3: Rezultati z različnimi zlatimi standardi (rang-1).



Slika 4: Delež ponovljivih odgovorov.

diagonalna matrika. V praksi je ta matrika odvisna od lastnosti in podobnosti posameznih razredov, ki jih določijo eksperti.

3 Eksperimentalni rezultati

Klasifikacijska točnost: Eksperimentalni rezultati so prikazani na Sliki 2, kjer so s $P1 - P8$ označeni laični anotatorji, medtem ko so z $E1 - E4$ označeni ekspertni anotatorji. Iz eksperimentalnih rezultatov po rangu 1 in 2 vidimo, da je točnost laičnih anotatorjev precej nizka, saj so si slike med sabo precej podobne, predvsem tiste iz višjih kvalitetnih razredov. Najvišja točnost laičnih anotatorjev je tako 0.35 po rangu 1. Ekspertni anotatorji se po pričakovanjih odrežejo precej bolje, s točnostmi med 0.44 in 0.6 po rangu 1. Še vedno pa se njihova točnost ne približa vrednosti 1, kar priča o težavnosti oz. dvoumnosti klasifikaije. Pri tem je potrebno poudariti tudi, da imajo v praksi eksperti več informacij za odločanje kot zgolj eno sliko, zato bi bila njihova uspešnost v praksi seveda veliko višja. Najvišjo točnost doseže nevronska mreža, učena s standardno kategorično križno entropijo.

Pri evalvaciji z mehko točnostjo opazimo, da se tako merjena uspešnost poveča vsem udeležencem, kar ni presenetljivo, saj so v tem primeru napake manj kaznovane. Že sam naključni klasifikator bi v tem primeru dosegel uspešnost 0.48, kar je veliko višja osnova od 0.125 v primeru klasične klasifikacijske točnosti. Vidimo tudi, da se je predvsem dvignila mehka točnost laikov, kar pomeni, da so ti pogosto zgrešili, a da napake, ki so jih naredili, niso bile tako problematične.

Ponovljivost rezultatov: V eksperimentalno zbirko smo ponovljene slike vključili, da bi preverili konsistentnost človeške klasifikacije. Deske nekaterih razredov so na pogled zelo podobne, kar lahko povzroča težave tudi ekspertom. Na Sliki 4 je prikazan odstotek ponovljenih slik, za katere je posameznik obakrat izbral isti razred.

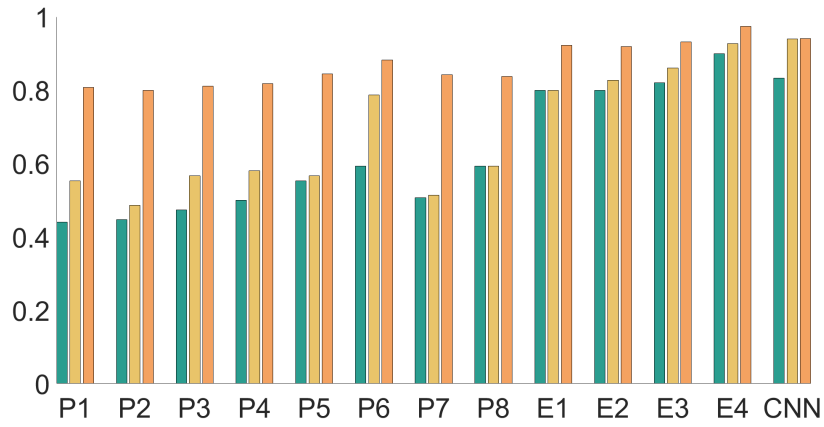
Rezultati prikazujejo precej nizko ponovljivost. Med laiki ponovljivost močno variira, pri čemer sta $P2$ in $P6$ enako klasificirala le 30% oz. 40% primerov, kar dejansko priča o kompleksnosti problema. Po drugi strani so bili eksperti pri klasifikaciji bolj ponovljivi, vendar še vedno ne presegajo 70% ponovljivosti. Model globokega učenja seveda dosega 100% ponovljivost, saj pri detekciji niso prisotni nobeni stohastični procesi.

Različni eksperti kot zlati standard Ker so se pri klasifikaciji pogosto razlikovala tudi mnenja ekspertov, smo naredili še evalvacijo pri kateri smo za zlati standard ("pravilno" vrednost razreda za posamezno sliko) vzeli napovedi posameznih ekspertov. Na Sliki 3 prikazujemo za vsakega udeleženca s štirimi dodanimi pikami klasifikacijsko točnost odgovorov pri zlatih standardih štirih ekspertov. Z zelenim stolpcem prikazujemo povprečje teh štirih vrednosti, medtem ko z rumeno barvo prikazujemo razliko do točnosti pri dejanskem zlatem standardu (kot na Sliki 2). Pri ekspertih vzamemo za zlati standard le preostale tri eksperte.

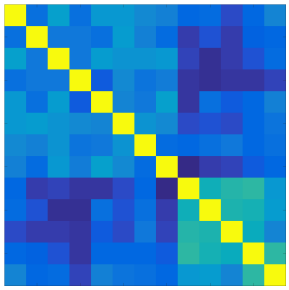
V primeru ko vzamemo druge eksperte za zlati standard se pri večini udeležencev rezultat lahko močno razlikuje med enim in drugim zlatim standardom. Pri večini udeležencev je variacija z različnimi zlatimi standardi lahko tudi po 10%. To implicira veliko razliko med odgovori različnih ekspertov, posledično pa vpliva tudi na učenje globoke metode. Pri slednji lahko opazimo veliko varianco med točnostjo ob različnih zlatih standardih na eni strani, ter izrazito zmanjšanje točnosti, ko uporabimo druge zlate standarde na drugi strani. To nakazuje, da se je globoka metoda naučila specifičnosti podanega zlatega standarda pri učenju, ki pa pogosto ne odraža podobnega zlatega standarda kot ga dojemajo preostali eksperti.

Zmanjšanje povprečne točnosti pri uporabi različnih zlatih standardov je opazno tudi pri laičnih udeležencih ($P1 - P8$), medtem ko tega ni mogoče zaznati pri ekspertih ($E1 - E3$). To nakazuje, da so v povprečju eksperti bolj usklajeni kot laični udeleženci. Razlika je opazna le pri najboljšem ekspertu $E4$, saj je tudi osnovni zlati standard modeliran po istem ekspertu.

Množica ekspertov kot zlati standard Kot zlati standard smo dodatno evalvirali tudi množico ekspertov. Na Sliki 5a prikazujemo rezultate, kjer smo za pravilni odgovor šteli tisti razred, ki ga je navedel vsaj eden izmed ekspertov. Pri $E1 - E4$ smo uporabili le preostale tri eksperte. Ta postopek sloni na predpostavki, da so tisti razredi, ki jih navajajo eksperti, veliko bolj pravilni kot tisti, ki jih niti en ekspert ni navedel. V tem primeru pričakovano rezultati povsod poskočijo. Najbolj očitno je povečanje točnosti po vseh metrikah pri ekspertih, kjer so rezultati poskočili tudi preko 80% celo pri klasifikacij-



(a)



(b)

Slika 5: (a) Rezultati z upoštevanjem mnenja katerega koli izmed ekspertov kot pravilno (top5), ter (b) matrika klasifikacijske točnosti pri različnih zlatih standardih.

ski točnosti ranga 1. To nakazuje, da so napake ki so jih delali eksperti, zelo podobne odgovorom drugih ekspertov.

Ta lastnost se odraža tudi v matriki točnosti na Sliki 5b, ki prikazuje klasifikacijsko točnost posameznega udeleženca, ko je za zlati standard vzeta napoved vsakega drugega udeleženca. Na njej je mogoče videti dva vizualna bloka. Prvi blok zajema vseh osem laikov, medtem ko drugi blok zajema vse štiri eksperte ter CNN. Opazna razlika med blokoma pomeni, da so napake, ki jih delajo laiki, med seboj podobne, in ravno tako, da so napake, ki jih delajo eksperti, med seboj podobne. Model globoke mreže se uvrsti v drugi blok, kar pomeni, da so napake, ki jih dela, tipično zajete v odgovorih drugih ekspertov.

4 Zaključek

V članku smo želeli raziskati problem nezanesljivih oz. težko določljivih označb v specifičnih problemskih domenah, kjer niti eksperti ne morejo vedno enolično in konsistentno določiti pravilnih razredov. Na problemu klasifikacije kakovosti desk smo pokazali, da lahko globoko učenje močno preseže rezultate laikov ter doseže rezultate primerljive ekspertom oz. v določenih primerih celo boljše, hkrati pa smo opozorili na problem neenolične določljivosti razredov.

Pokazali smo, da so napake laikov veliko bolj podobne med seboj, kot so podobne napakam ekspertov, ter obratno. Pri tem tudi klasifikator, ki temelji na globoki nevronski mreži, izkazuje napake, ki so bolj podobne napakam ekspertov. Izkaže pa se tudi, da se v primeru, ko za zlati standard vzamemo odgovor katerega koli izmed odgovorov ekspertov, rezultati nevronske mreže ne izboljšajo tako zelo kot rezultati ostalih udeležencev. Tak rezultat je pričakovan, saj globoka metoda zgradi model ob upoštevanju podanega zlatega standarda, pri čemer ne eksplicitno upošteva sorodnosti med razredi, oz. se tega zaveda samo implicitno preko ugotovljene vizualne podobnosti.

Ravno to dejstvo predstavlja potencial za izboljšanje metode za avtomatsko klasifikacijo. V primeru podane matrike uteži napake bi bilo smiselno to informacijo neposredno upoštevati že pri učenju in tako omogočiti metodi, da se nauči bolj izogibati klasifikaciji v razrede, ki so za podani primer bolj kritični. V ta namen smo preizkusili metodo mehkih označb razredov [5], ki ni dosegla zadovoljivih rezultatov. Bolj obetaven je pristop z uteževanjem funkcije izgube, ki ga nameravamo dodelati v našem prihodnjem delu.

Vsekakor pa se je v aplikativnih raziskavah za reševanje konkretnih problemov v industriji potrebo zavedati, da je uspešnost delovanja končne aplikacije zelo odvisna od podanega zlatega standarda, ki mora biti čim bolj verodostojen. V primerih, ko le-ta ni enolično določljiv, pa je potrebno to dejstvo upoštevati tako pri učenju, kot pri evalvaciji tovrstnih klasifikacijskih sistemov.

Zahvala

To delo je bilo delno financirano s strani ARRS projekta J2-9433 (DIVID) in raziskovalnega programa Računalniški vid (P2-0214). Posebna zahvala gre tudi podjetju Menina d.o.o. za pripravo slikovnih podatkov, ter vsem anotatorjem, tako laičnim kot ekspertnim.

Literatura

[1] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making Better Mistakes: Leveraging Class Hierarchies with Deep Networks. In *CVPR*, pages 12506–12515, 2020.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[4] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When Does Label Smoothing Help? In *NeurIPS*, 2019.

[5] Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. *ICCV*, Oct 2019.

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, volume 2016-Decem, pages 2818–2826, 2016.

Co-segmentation for visual object tracking

Luka Čehovin Zajc

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-pošta: luka.cehovin@fri.uni-lj.si

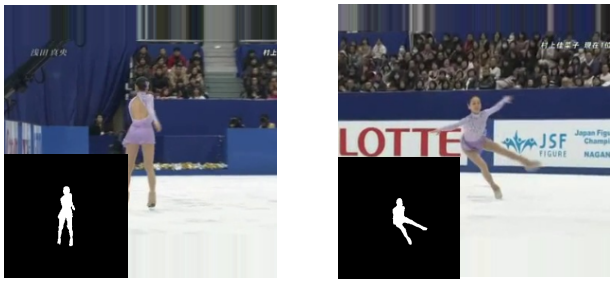
Abstract

Visual trackers have already achieved remarkable robustness due to discriminative template matching paradigm and deep descriptive features that are employed for this purpose. Yet, they are limited in their accuracy due to bounding box constraint and low spatial resolution of deep features. In recent works it has been shown that improving accuracy of discriminative deep trackers with one-shot segmentation module may not only improve the accuracy, but also boost their robustness. In this work we are exploring co-segmentation as another option for determining accurate position and size through the task of segmenting an object. We present our preliminary results on a challenging VOT2020 dataset where the proposed method achieves competitive performance to the state-of-the-art.

1 Introduction

Visual tracking is an important building block in many real-world computer vision applications. Using the most widely accepted definition of the problem, a computer vision algorithm is required to predict the state of the object in a sequence of images using only its initial state to support the prediction. Support data scarcity and the need for robustness of predictions have led most tracking algorithms to only address prediction of low-dimensional state, i.e. target position and scale. According to the recent benchmark [1], the dominant paradigm in the field is correlation bounding box tracking using deep features [2, 3] where the target represented by a multi-channel template is localized by cross-correlation between the template and a search region. These trackers are very robust, yet they operate on fixed-size rectangular patches and cannot deal with deformations that cause aspect changes.

Recently, a proposed combination of a discriminative template matcher and a deep generative segmentation model was proposed in [4]. The template matcher is responsible for coarsely localizing the target, addressing any distractors in the surrounding area. The segmentation network then finely localizes the target and infers its size and aspect ratio by determining its per-pixel mask. This combination has been shown to improve not only the accuracy of the tracker, but also its robustness. In this paper we are investigating an alternative approach to object segmentation using limited support data during the tracking session. Our basis is co-segmentation, where an objective is to jointly segment semantically similar objects in two or more images. We present our adaptations of a co-segmentation method (see Figure 1) to the visual object tracking domain and show some preliminary results that demonstrate the feasibility of the approach.



Anchor image and mask — Query image and mask prediction

Figure 1: Co-segmentation in visual tracking. Our method uses an anchor image and a mask of an object of interest to infer its mask in another frame.

2 Related work

For more than a decade, the most robust approaches to visual tracking have viewed the problem as a machine-learning task of discrimination between the object and the background [5]. In this time the features used to describe an image and the methods that enabled fast learning of a discriminative model have changed dramatically. Most recent methods use deep features, either pre-trained for discriminative localization of an object [2] or train a multi-channel discriminative correlation filter on-line [3, 6].

Different than localization, size estimation was commonly handled using a multi-scale pyramid, which is a greedy but inaccurate approach [7] or not taken into account at all. With the success of deep learning and acceptance of more power-hungry methods, some researchers begun experimenting with region proposal methods to estimate target's size and aspect ratio [8, 9], others employed box refinement methods [10, 6]. While segmentation output was considered in [9] it was not used to improve the overall tracking process. In [4] a segmentation was used in a mutually beneficial interaction with a dis-

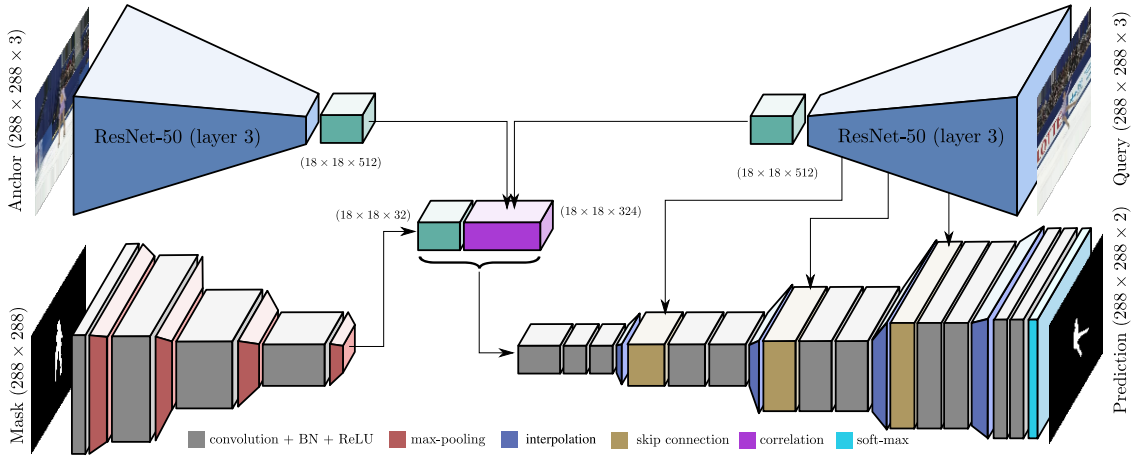


Figure 2: Schematic representation of our co-segmentation model for visual tracking.

criminative tracker for the first time. The discriminative tracker provided initial localization that is then refined using object segmentation. Segmentation is also used to adjust sampling size. All this contributes to increased accuracy and robustness. But the method is still partially handcrafted. In this work we investigate an approach based on co-segmentation method that is fully differentiable and can be trained end-to-end.

The field of co-segmentation is a special branch of segmentation where an objective is to segment one or more objects in more than one image at the same time [11]. There are many types of co-segmentation with assumptions about number of salient objects [11, 12], background constancy [13]. While co-segmentation seems like a good match for visual tracking, the objects of interest in co-segmentation are very loosely defined by salience and the emphasis is on generality. Resulting segmentation lacks specificity to segment only a single object which is the goal in visual tracking which also makes no assumptions about the appearance of the object itself as well as its surrounding area. In the following section we present a modification one of the deep co-segmentation methods [14] that makes it usable in visual tracking scenario.

3 The co-segmentation model

In this work we are evaluating a simple single-shot multi-object co-segmentation approach, inspired by [14]. The main concept behind the original work is a feature correlation layer that originates from FlowNet [15] for deep optical flow estimation. Instead of comparing deep features locally, as it is common in optical flow estimation, [14] compares all features from a reference image with all features from the query image, building an affinity tensor. This is then followed by an up-sampling decoder that is trained to reconstruct a mask of semantically meaningful objects found in both images.

In our co-segmentation network, displayed in Figure 2, the VGG16 [16] backbone is replaced with a pre-trained ResNet50 [17] that we have found to be better suited for segmentation task. We use features from the third layer of the feature extractor to calculate an affinity tensor between the anchor image and the query image. In contrast to the co-segmentation task we have introduced a separate encoder to encode the input mask from the anchor frame to its high-dimensional representation that we use during the prediction mask inference process together with the computed affinity tensor. In contrast to the original work, we also introduce skip connections from the feature encoder for the query image into the decoder to improve the details of the predicted mask.

3.1 Training

The model was trained on YouTube-VOS [18] video segmentation dataset. Each training sample contained two patches selected from the same sequence with corresponding segmentation masks of the target object. Samples are drawn randomly from the dataset and augmented using geometrical transformations (translation, rotation, scale) and pixel alteration (brightness, contrast, noise, blur). Examples of training samples are shown in Figure 3.

The ResNet feature extractor backbone was not updated during training, we have only trained mask encoder and mask decoder to minimize binary cross entropy (BCE) loss. Similarly to [14] we have used each sample multiple times, the final loss is therefore:

$$\begin{aligned} loss = & BCE(P(I_1, M_1, I_2), M_2) \\ & + BCE(P(I_2, M_2, I_1), M_1), \end{aligned} \tag{1}$$

where the tuple (I_1, M_1, I_2, M_2) constitutes a training sample, $P(\cdot, \cdot, \cdot)$ denotes the prediction model.

The model was trained on 40000 unique samples in batches of 16 samples for 33 epochs, each lasting for 400 iterations. We have used Adam optimizer [19] with initial learning rate 0.0001 that was decreased by half every 10 epochs.

3.2 Tracking

Similarly to [4], we have evaluated our segmentation model within a state-of-the-art discriminative tracker, presented in [6] where the IoU bounding box refinement [8] is replaced by our model for region and scale estimation. Both components share the same ResNet50 [17] feature extractor.

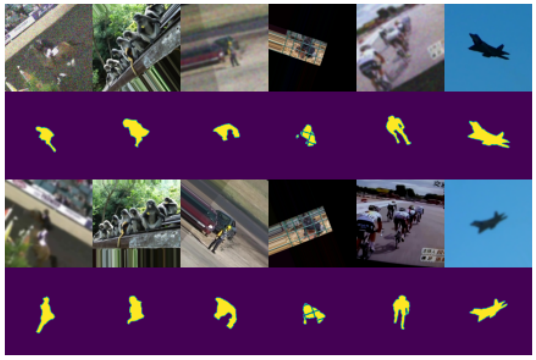


Figure 3: Examples of training samples, each sample column is composed of two augmented samples with corresponding segmentation masks.

Tracker	EAO	A	R	Overlap
ATOM [6]	0.267	0.467	0.708	0.386
SiamMask[9]	0.321	0.624	0.648	0.405
D3S[4]	0.439	0.699	0.769	0.508
Ours	0.416	0.696	0.735	0.578

Table 1: Results for VOT2020 baseline and unsupervised experiments. For all measures higher value is better.

In each frame the target is approximately localized using the discriminative component, then a query patch is extracted and fed into the segmentation model together with the anchor patch (from first frame) and its mask. The resulting mask is used to estimate the position of the target as well as its size. The size is used to adjust trackers internal image sampling scale. We assume that the scale changes according to a random-walk motion model that we estimate using Kalman filter [20]. The uncertainty of a prediction is estimated by using the co-segmentation in reverse: using the predicted mask and query image as reference the model predicts the mask in the anchor frame and compares it to the initialization mask. The overlap between those masks is used to determine prediction uncertainty.

4 Experimental evaluation

We have performed our experiments on a recent VOT2020 dataset [21]. The dataset contains 60 diverse sequences, the target object in each sequence is annotated with a segmentation mask. The overall results for the experiment with multiple restarts and the unsupervised single run experiment are shown in Table 4. Scores in the table show that our method is performing comparably to the D3S [4] and significantly outperforms another recent segmentation tracker SiamMask [9] as well as ATOM [6] from which it was derived. More importantly, the results also show that our method improves results of ATOM in terms of accuracy as well as robustness.

Comparison between our method and D3S [4] is also interesting because the former can be viewed a generalization of the latter. In D3S the mask is directly down-sampled to the feature size, the generative model and its application are tuned by hand. Our method performs down-sampling and matching in a fully differentiable manner, allowing the network to take care of the exact details of which data is important. This requires a larger network, at the moment we have been very liberal with its size, but we believe that the number of weights can be reduced with further experiments. Still, it is worth noting that our model was trained with around 2% of data used by D3S and in 20% of the time (around 4 hours).

Even though the overall performance of D3S and our method are similar, the results on per-sequence basis show some conceptual differences. Figure 4 shows frames of selected sequences. In general D3S better estimates regions with highly deformable objects on a clearly different background (Figure 4, sequences gymnastics1 and iceskater1). Our method is more conservative and does not segment every limb in such cases. On the other hand it can handle occlusions with similar objects better because of the more verbose correspondence description (Figure 4, sequences girl, flamingo1, and nature).

The main weaknesses of the proposed method is the static anchor template and the fragile scale estimation approach. If the scene will change too much from the first frame, the match will become unreliable. While our method does not tend to explode the segmentation as frequently as D3S, it is much more prone to collapse: because the difference between the anchor and query are too great, the segmentation only captures part of the object (hence wrong scale) or confuses it with with a similar object with better matching appearance (Figure 4, sequence fish2). This shows the need for a scale estimation that is decoupled from segmentation as well as some kind of anchor update mechanism.

5 Conclusion

We have presented our preliminary work with integration of a co-segmentation model into a visual tracker. We have modified a co-segmentation model [14] and integrated into a state-of-the-art discriminative tracker. The results on a challenging VOT2020 dataset show that our proposed tracker outperforms a tracker that it was derived from [6] and performs comparably to a very recent method that integrates segmentation into tracking in a conceptually similar way [4].

We believe that our findings open up new opportunities for more accurate and robust tracking using segmentation. In the future we plan to investigate how to reliably update the anchor image to keep it relevant in the long run or to even use multiple anchor images. We will also investigate how to estimate object's scale directly in the model and how to resolve ambiguities and distractions by observing the context from the past frames.

Acknowledgement: This research was supported by the ARRS grant Z2-1866.

References

[1] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J. Kämäräinen, L. Cehovin Zajc, O. Drbohlav, A. Lukezic, A. Berg, A. Eldesokey, J. Käpylä, G. Fernández, and et al. The seventh visual object tracking

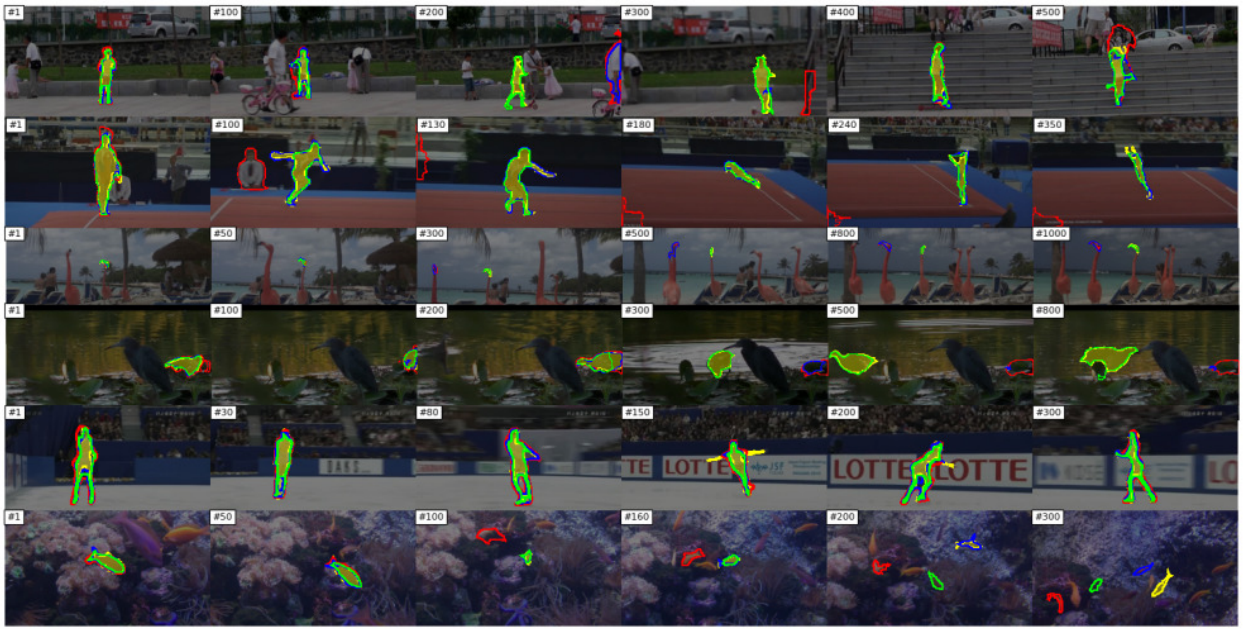


Figure 4: Qualitative evaluation on selected sequences from VOT2020: girl, gymnastics1, flamingo1, nature, iceskater1, and fish2. Groundtruth is shown with yellow, tracker results with red (SiamMask [9]), blue (D3S [4]) and green (ours). All images were dimmed for better region visibility.

vot2019 challenge results. In *ICCVW*, pages 2206–2241, 2019.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, pages 850–865. Springer, 2016.

[3] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, pages 6638–6646, 2017.

[4] Alan Lukezic, Jiri Matas, and Matej Kristan. D3s-a discriminative single shot segmentation tracker. In *CVPR*, pages 7133–7142, 2020.

[5] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *BMVC*, 2006.

[6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019.

[7] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015.

[8] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[9] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.

[10] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018.

[11] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *CVPR*, volume 1, pages 993–1000. IEEE, 2006.

[12] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[13] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Cosegmentation revisited: Models and optimization. In *ECCV*, pages 465–479. Springer, 2010.

[14] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In *ACCV*, pages 638–653. Springer, 2018.

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[20] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[21] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J. Kämäräinen, M. Danelljan, L. Cehovin Zajc, A. Lukezic, O. Drbohlav, L. He, Y. Zhang, S. Yan, J. Yang, G. Fernández, and et al. The eight visual object tracking vot2020 challenge results. In *ECCVW*, 2020.

Using Google Colaboratory for Image Classification

Martina Đulabić, Damir Demirović, Amira Šerifović-Trbalić

Faculty of Electrical Engineering, University of Tuzla, Tuzla, Bosnia and Herzegovina
E-mail: damir.demirovic@untz.ba

Abstract. *Deep learning is a part of machine learning that uses multiple layers of neural networks to extract a higher level of features from the image. Convolutional neural networks (CNN) are particularly suited for processing images and classification. The aim of this paper is the training and evaluation of the well known CIFAR-10 dataset in the free Google Cloud environment on the Graphics Processing Unit (GPU) and Central Processing Unit (CPU). For the given model, performances are evaluated for both platforms. The obtained mean precision of this data set was 86.6% and the GPU speedup against CPU was about 10 times. Differences between precision on GPU and CPU are also discussed. Additionally, classification performance was also evaluated on the randomly collected dataset images from the Internet, and in this case, the mean precision is 94%.*

1 Introduction

Deep learning [1-3] represents a branch of machine learning, especially suited for problems of artificial intelligence.

Convolutional neural network (CNN) is a class of deep neural networks which are inspired by a biological process so that the organization of neurons resembles an animal visual cortex.

In [4] authors give a review of Convolutional neural networks for image classification. Performance of several Convolutional neural networks are evaluated with a several known datasets and as well with CIFAR-10 [5].

In [6] the author shows how to train a multi-layer generative model that learns to extract meaningful features that resemble those found in the human visual cortex.

The performance of different classifiers on the CIFAR-10 was presented in [7]. The ensemble of the classifier was built to reach a better performance. The conclusions are that k-Nearest Neighbour (kNN) and CNN can yield a better result when combined.

In [8] authors introduce GPipe, a pipeline parallelism library that allows scaling any network that can be expressed as a sequence of layers. The GPipe allows scaling a variety of different networks to gigantic sizes efficiently.

The data augmentation as an effective technique for improving the accuracy of modern image classifiers was introduced in [9]. The proposed simple procedure AutoAugment automatically searches for improved data augmentation policies. A search algorithm has been used to find the best policy, such that the neural network yields the highest validation accuracy on a target dataset.

In [10] authors provide formalism which unifies and categorizes the landscape of existing methods along with detailed analysis that compares and contrasts different approaches.

This paper uses the convolutional neural network for image classification. The image classification is well suited for CNN because of their convolutional nature. For classification, the well known CIFAR-10 dataset was used which groups images in 10 classes.

2 Methods

For training neural network CIFAR-10 dataset was used. CIFAR-10 dataset is collection of natural images that are commonly used to train machine learning and computer vision algorithms. It is one of the most widely used datasets for machine learning research. CIFAR-10 represents the dataset which contains 60000 color images of dimensions 32x32 pixels obtained from the work of Alex Krizhevsky [6]. The images were taken under varying conditions in pose, orientation, illumination, and even partial occlusion.

The images were preprocessed in the following way. Images were cropped to 24x24 pixels, then standardization on image pixels was used. In the training phase, the random transformations were used: random left flip or right flip, random brightness and random contrast. The image transformations can slow the training phase, so the transformations were performed in parallel. Classes are denoted as: I – airplane, II- automobile, III- bird, IV- cat, V- deer, VI- dog, VII- frog, VIII- horse, IX- ship and X- truck.

The model is trained and evaluated using TensorFlow on Google Colaboratory. TensorFlow [11] is an open-source software library. Compute Unified Device Architecture (CUDA) [12] is a parallel programming platform and programming model developed by NVIDIA for general computing on graphical processing units (GPGPU). Google Colaboratory [13] enables the simple way of using the TensorFlow. This is the free framework based on Jupyther notebook [14], and it is executed directly in the Google Cloud. The code can be executed on the CPU and GPU and in the last time the most popular option is used with Tensor Processing Unit (TPU) [15].

The hardware specification of the used machine for the training and testing was as following. The desktop 64-bit notebook with Linux was used, with Intel i3-5005U

processor with 2.00 GHz and 4 cores with 4 GB main memory. The machine on Google Cloud on Intel(R) Xeon(R) processor at 2.2 GHz with 2 cores, and NVIDIA Tesla K80 GPU with 12 GB RAM.

The model has multilayer architecture with convolution and nonlinearity, and the executed operations are shown (Figure 1). The model follows the model architecture described in the Alex Krizhevsky, with the couple differences in the several higher layers. The training model has softmax regression, and for the regularization L2, a loss is used for the learned variables. Table I gives the class and mean performance for two datasets. Table II gives an overview of the performances of the model using GPU and CPU. The training time on the Google Cloud for the GPU was 1 hour and 15 minutes, the CIFAR-10 model has 10112 images and the proposed model correctly predicted the 8731.It can be seen that the model achieved a mean precision of 86.3%. When CPU was used on Google Colaboratory, the training was much slower and the total training time was 19 hours and 50 minutes. The mean precision, in this case, was 86.6%. With CPU the speed of processing does not increase. The difference in precision, about 0.3%, is due to hardware differences in processing units, which is negligible for this model (Table 2). Speedup and total training time for two platforms are shown in Figure 2. Calculated precision and recall are given in (Table I and Table II). In Figure 3 confusion matrix obtained with the proposed method for original CIFAR-10 is shown.

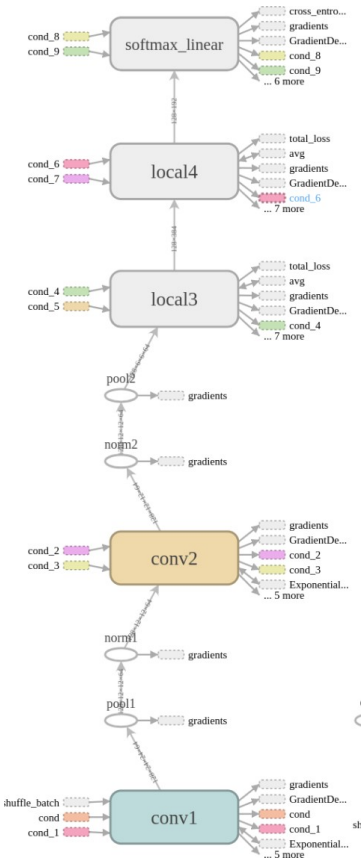


Figure 1: Model prediction diagram

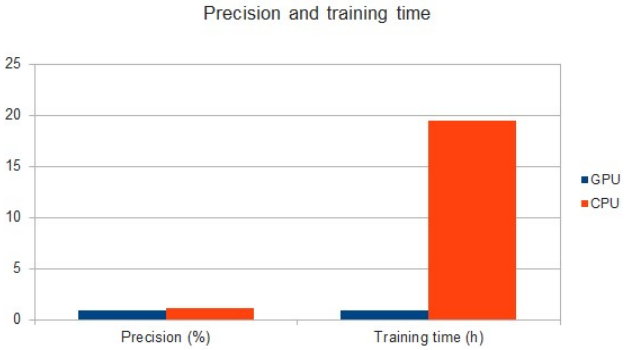


Figure 2: Training time and precision for GPU and CPU

It can be seen that the model correctly predicts 8731 from the 10112 input images.

Most images are correctly predicted. The reason for this is the obvious similarities in shape and texture between these groups of images. Further, the performance of the classifier on new images collected from the Internet was evaluated. For this evaluation, 100 color images, were randomly selected, 10 images from each class. These images were cropped to 24x24 pixels and converted into binary format and their labels are saved. These data are used for the evaluation of the already trained CIFAR-10 model. From 100 input images, the model correctly classifies 94 images, so for this data set, precision is 94%.

I	851	9	25	11	10	9	15	9	33	13
II	9	865	10	12	14	10	14	10	13	12
III	69	11	830	15	11	10	9	11	9	15
IV	15	9	13	855	12	85	15	13	14	9
V	13	12	11	13	880	14	11	55	8	13
VI	14	8	15	75	10	858	15	10	12	11
VII	10	12	9	14	15	15	901	15	9	10
VII	15	11	11	10	11	9	15	898	11	10
IX	9	10	10	9	12	15	11	10	907	14
X	15	72	11	10	11	10	15	9	13	886
	I	II	III	IV	V	VI	VII	VII	IX	X

Figure 3: Confusion matrix for classification for CIFAR - 10 dataset for the proposed model

The confusion matrix for this data set is given in Figure 4. From 100 input images the model correctly classifies 94 images, so for this data set precision is 94%. It can be seen that the model give a better performance when additional collected images were used. Zero values indicate there are no misclassifications.

2.2 Comparison with other models

In this section, the collected Internet dataset is additionally evaluated with the following three well-known models ResNet V20 [16], LeNET [17] and

	I	II	III	IV	V	VI	VII	VII	IX	X
I	9	0	0	0	0	0	0	0	0	0
II	0	8	0	0	0	0	0	0	0	0
III	1	0	10	0	0	0	0	0	0	0
IV	0	0	0	9	0	1	0	0	0	0
V	0	0	0	0	10	0	0	0	0	0
VI	0	0	0	1	0	9	0	0	0	0
VII	0	0	0	0	0	0	10	0	0	0
VII	0	0	0	0	0	0	0	10	0	0
IX	0	0	0	0	0	0	0	0	10	0
X	0	2	0	0	0	0	0	0	0	10

Figure 4: Confusion matrix for classification for images collected from the Internet for the proposed model

Network in Network [18] which can be trained with CIFAR-10 dataset.

In [16], authors present residual learning framework (ResNet) which is meant to train substantial deeper models. For the used model the number of trainable parameters was 270.066. Precision on CIFAR-10 set is 91.25% and with Internet dataset is 91% (Figure 5). LeNet [17] is a classical architecture that is straightforward and simple to understand. The number of trainable parameters was 62.006 (Figure 6).

	I	II	III	IV	V	VI	VII	VII	IX	X
I	9	0	0	0	0	0	0	0	1	0
II	0	10	0	0	0	0	0	0	0	0
III	2	0	7	1	0	0	0	0	0	0
IV	0	0	0	8	1	1	0	0	0	0
V	0	0	0	0	10	0	0	0	0	0
VI	0	0	0	0	0	9	0	0	0	0
VII	1	0	1	0	0	0	8	0	0	0
VII	0	0	0	0	0	0	0	10	0	0
IX	0	0	0	0	0	0	0	0	10	0
X	0	0	0	0	0	0	0	0	0	10

Figure 5: Confusion matrix for classification for images collected from the Internet for ResNET V20

Network in Network (NiN) [18] is a micro neural network with a more complex structure to abstract data with a receptive field, which is in contrast with the conventional convolutional neural network layer which uses linear filters followed by a nonlinear activation function. The number of trainable parameters is 966.986. On CIFAR-10 precision is 76.23%, and precision on the Internet dataset is 62% (Figure 7).

3 CONCLUSIONS AND FUTURE WORK

In this work, the convolutional neural network for classifying the well known CIFAR-10 model is evaluated.

The model is trained and evaluated on the CPU and GPU ~~on~~ using the Google Colaboratory and with Google TensorFlow on the desktop CPU. The performances of the trained models on the CPU and GPU were evaluated in the Google Cloud environment. The obtained mean precision is 86.6% for the CPU. The

	I	II	III	IV	V	VI	VII	VII	IX	X
I	8	0	1	0	0	0	0	0	1	0
II	0	6	0	0	0	0	0	0	2	2
III	2	0	7	0	0	0	0	0	0	1
IV	1	0	0	2	0	5	2	0	0	0
V	1	0	1	0	7	0	0	1	0	0
VI	0	1	2	1	1	2	2	1	0	0
VII	2	0	2	0	1	0	5	0	0	0
VII	0	0	0	0	0	1	0	9	0	0
IX	0	1	0	0	0	0	0	0	9	0
X	0	3	0	0	0	0	0	0	0	7

Figure 6: Confusion matrix for classification for images collected from the Internet for LeNet

	I	II	III	IV	V	VI	VII	VII	IX	X
I	9	0	0	0	0	0	0	0	1	0
II	4	0	2	0	0	0	0	0	3	1
III	3	0	7	0	0	0	0	0	0	0
IV	0	0	2	4	4	0	0	0	0	0
V	0	0	0	0	10	0	0	0	0	0
VI	0	0	1	1	3	5	0	0	0	0
VII	0	0	3	0	1	0	6	0	0	0
VII	0	0	1	0	4	0	0	5	0	0
IX	0	0	0	0	0	0	0	0	10	0
X	2	0	1	0	0	0	0	0	2	5

Figure 7: Confusion matrix for classification for images collected from the Internet for NiN

speedup for the GPU is about 10 times than the CPU. The differences in precision between the CPU and GPU results were shown and in this case, doesn't have a significant impact on classification results.

The proposed approach gives a relatively good performance if the training time is taken into account. For the evaluation purposes, the network is tested on another set of images collected from the Internet and the obtained mean precision for this case is 94%. The performance of results should increase when the data set grows. Additionally, the performances were evaluated

Table I: Precision and recall on the GPU for the CIFAR-10 and Internet dataset

		Class										
Dataset	**Measure**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**Average mean**
CIFAR-10	Precision	0.83	0.85	0.88	0.83	0.89	0.83	0.88	0.86	0.88	0.89	0.86
	Recall	0.86	0.89	0.84	0.82	0.85	0.83	0.89	0.89	0.90	0.84	0.86
Internet	Precision	0.90	0.80	1.0	0.90	1.0	0.90	1.00	1.00	1.00	1.00	0.94
	Recall	1.00	1.00	0.91	0.90	1.0	0.90	1.00	1.00	1.00	0.83	0.94

Table II: Performances for training on GPU and CPU

		GPU		CPU	
Processing speed	per batch (s)	0.962	0.045	0.658	0.772
	examples (s)	133.1	2841.1	194.6	165.8
Total loss		4.67	0.73	4.67	0.74
Training time	H	1.15		19.50	
Precision	%	86.3		86.6	

on the three well-known models ResNet, LeNet and NiN. When training with these three models only ResNet gives a comparable performance using the Internet dataset, with 91% precision, with the lowest performance on the cat data, whereas the LeNet and NiN give about 60%. The LeNet model is to simple to handle this task and get the lowest performance on the dog data and the NiN model needs additional tuning to get better performance and show the lowest performance on the bird data. Future work can be taken in increasing the classification precision, and additional data augmentation techniques, tweaking the network layers. Also exploiting the multiple GPU for training can be used.

BIBLIOGRAPHY

[1] I. Goodfellow, Y. Bengio, A. Courville: Deep Learning, MIT Press, 2016.

[2] J.Schmidhuber: Deep Learning in Neural Networks: An Overview. CoRR abs/1404.782, 2014

[3] N. Buduma, N. Locascio, Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms, O'Reilly Media, 2017.

[4] W. Rawat, Z. Wang: Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review, Neural Computation 29:9, 2017

[5] N. Sharma, V. Jain, A. Mishra: An Analysis Of Convolutional Neural Networks For Image Classification, Procedia Computer Science,132, 2018

[6] A. Krizhevsky: Learning multiple layers of features from tiny images, 2009.

[7] Y. Abouelnaga, O. S. Ali, H. Rady, M. Moustafa: CIFAR-10: KNN-Based Ensemble of Classifiers, 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 12 2016.

[8] Y. Huang, Y. Cheng, A. Bapna, O. Firat, M. X. Chen, D. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu i Z. Chen: GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism, 2018.

[9] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le: AutoAugment: Learning Augmentation Policies from Data, 2018.

[10] M. Wistuba, A. Rawat, T. Pedapati: A Survey on Neural Architecture Search, 2019.

[11] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu i X. Zheng: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

[12] https://developer.nvidia.com/cuda-zone (accessed 18.09.2019)

[13] https://colab.research.google.com (accessed 18.06.2020)

[14] https://jupyter.org/ (accessed 18.06.2020)

[15] https://cloud.google.com/tpu/docs/tpus (accessed 18.06.2020)

[16] He, K. Zhang, X. Ren, S. Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 770-778), 2016.

[17] LeCun, Yann. "LeNet-5, convolutional neural networks." URL: http://yann. lecun. com/exdb/lenet 20, 2015.

[18] Lin, Min. Qiang, Chen. Shuicheng, Yan. Network in network. arXiv preprint arXiv:1312.4400, 2013.

A review and comparison of time series similarity measures

Maša Kljun[1], **Matija Teršek**[1], **Erik Štrumbelj**[1]

[1]*Faculty of Computer and Information science, University of Ljubljana*
E-mail: mk2700@student.uni-lj.si, mt2421@student.uni-lj.si, erik.strumbelj@fri.uni-lj.si

Abstract

We review 12 time series similarity measures and investigate their time complexity, normalization, invariance with respect to warping and scaling, support of time series of different lengths, and other properties. We show on simulated data that several similarity measures perform well on average, but none perform well in all cases and in some cases measures that typically perform poorly, such as compression-based similarity, are a better alternative.

1 Introduction

Measuring similarity between time series is an important component in time series data analysis, especially unsupervised learning. Many different measures exist and it is often not clear which measure is the best choice for the test at hand or how different measures compare with respect to relevant properties such as invariance to scaling/warping and time complexity.

The few related works are Wang et al. [13] who compare 9 measures but omit those based on correlation coefficients or compression. Serra and Arcos [10] and Górecki and Piasecki [6] compare 7 and 30 measures, respectively, but focus on 1-NN classification performance and not clustering performance and other properties as we do. Esling and Agon [4] do focus on other properties, but not on clustering performance.

We aim to provide a compact review and classification of the most commonly used similarity measures and relevant properties which are often excluded in related work. Furthermore, we use several simulated data sets to empirically evaluate how different measures compare to each other and how well they perform in clustering.

2 Distance measure features

In this paper we will view time series similarity measures in terms of these properties, which are relevant to choosing the best similarity measure for the task at hand:

- **Time complexity.**
- **Can compare time series of different lengths.**
- **Normalization.** Does increasing the length or sampling frequency of the time series, without changing any other properties, change the value? If so, we provide a factor that normalizes the measure and facilitates comparison across time series of different lengths.
- **Invariance/robustness with respect to warping and scaling.** Warping is a change of the time series' times that preserves the ordering. Scaling is multiplication of the time series' values with a constant. Related work is inconsistent about warping, so we additionally define *weak invariance* (the same change is applied to both compared time series) and *strong invariance* (the change is applied to only one of the time series). Strong invariance to warping implies weak invariance to warping.

A summary of similarity measures is in Table 1.

3 Distance measures

Let $X = x_1, ..., x_n$ and $Y = y_1, ..., y_m$ be the two time series whose similarity we are interested in. We also use X_{-n} and X_{-1} to represent X without the last and first point, respectively.

3.1 L_p norms/distances

Depending on the value of p, we have:

- **Manhattan** ($p = 1$): $\sum_{i=1}^{n} |x_i - y_i|$.
- **Minkowski** ($1 < p < \infty$): $\sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p}$.
- **Euclidean** ($p = 2$): $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
- **Infinite norm** ($p = \infty$): $\max_{i=1,...,n} |x_i - y_i|$.

In empirical evaluation, we use Euclidean distance.

3.2 DISSIM

DISSIM is designed to work with time series with different sampling rates. It is defined as the integral of the Euclidean distance between the two series, which are assumed to be linear between sampling points. It can be normalized by dividing by $(n - 1)$ when time series are extended, but no normalization is required if we increase the sampling frequency on the same time interval.

3.3 Dynamic Time Warping

Dynamic time warping (DTW) was introduced in order to overcome some of the restrictions of simpler similarity measures such as Euclidean distance. It is to this day one of the most popular measures.

The DTW costructs a mn matrix of squared distances between points of both time series, which is then used as a cost matrix when searching for the cheapest path between $(1,1)$ and (m,n). Path cost determines the similarity.

Normalization depends on the step pattern - allowed transitions and weights between matched pairs, when searching for an optimal path. Normalization is then made by dividing the distance by n, m or $n+m$ depending on the step pattern and slope weighting [5].

3.4 Edit distance on Real Sequences

The Edit distance algorithm is based on counting the number of insert, delete, and replace operations required to transform one string into another. It can be applied to time series where points from X and Y are considered a match if their absolute distance is less than some ϵ.

EDR is defined recursively as: $E(X,Y) =$

$$\min\{E(X_{-1},Y_{-1})+s, E(X_{-1},Y)+1, E(X,Y_{-1})+1)\},$$

where subcost s is 0 if elements match and 1 otherwise [3]. $E(X,Y)$ is 0 if either of the time series are empty. According to Chen et al. [3] we get the best clustering results, when ϵ is set to a quarter of the maximum standard deviation of time series.

3.5 Edit distance with Real Penalty

Edit distance with Real Penalty method (ERP) is also based on the Edit distance algorithm. When addition or deletion happens in Y, ERP treats this as a gap. The ERP distance is defined as $E(X,Y)$

$$= \begin{cases} \sum_{i=1}^{m} |x_i - g| & ; n = 0 \\ \sum_{i=1}^{n} |y_i - g| & ; m = 0 \\ \min\{E(X_{-1},Y_{-1}) + d(x_1,y_1), & \\ E(X_{-1},Y) + d(x_1,g), & \\ E(X,Y_{-1}) + d(y_1,g)\} & ; else \end{cases}, \quad (1)$$

with $d(x,y)$ being defined as a L_1 norm between the points x and y [2]. If one of them is a gap point, it is equal to a user defined constant g.

3.6 Longest common subsequence

We use the Longest common subsequence (LCSS) model to cope with various problems such as different sampling rates, different lengths, outliers, and efficiency [12]. It allows for unmatched elements and efficient approximate calculation. It is defined as

$$D(\delta,\epsilon,X,Y) = 1 - \frac{L_{\delta,\epsilon}(X,Y)}{\min(n,m)}, \quad (2)$$

where L is a function defined as $L_{\delta,\epsilon}(X,Y)$

$$= \begin{cases} 0, & A \\ 1 + L_{\delta,\epsilon}(X_{-n},Y_{-m}), & B \\ \max\{\, L_{\delta,\epsilon}(X_{-n},Y), L_{\delta,\epsilon}(X,Y_{-m})\}, & \text{else} \end{cases}, \quad (3)$$

where $A = \{n = 0 \vee m = 0\}$ and $B = \{|x_{xn} - y_{xm}| < \epsilon, |x_{yn} - y_{ym}| < \epsilon, |n-m| < \delta\}$.

3.7 TQuEST

This similarity measure starts by transforming each time series into disjoint intervals such that within every interval all the time series' values are above a threshold ϵ. Let $S(X,\epsilon)$ be the unique such transformation of X where the intervals are the largest.

The TQuEST measure is defined as:

$$T(X,Y) = \frac{1}{|S(X,\varepsilon))|} \sum_{s\in(X,\varepsilon))} \min_{s'\in S(Y,\varepsilon))} d(s,s') + \frac{1}{|S(Y,\varepsilon)|} \sum_{s'\in S(Y,\varepsilon)} \min_{s\in S(X,\varepsilon))} d(s',s) \quad (4)$$

where $d(a,b)$ is the Euclidean distance between two time intervals [1].

3.8 Cross-corelation

Cross-correlation (CCor) is based on the Pearson correlation coefficient [7]. It is defined as $d_{CC}(X,Y) = \sqrt{\frac{1-CC_0(X,Y)}{\sum_{k=0}^{ml} CC_k(X,Y)}}$, where CC_k is the lag-k covariance and ml is the maximum allowed lag between X and Y and should not exceed the length of the series. By default, it is $\min\{n,m\} - 1$ [7].

3.9 Compression-based similarity measure

The compression-based similarity measure (CDM) is a class of measures defined as $\text{CDM}(X,Y) = \frac{C(XY)}{C(X)+C(Y)}$, where $C(X)$ is the size in bytes of the compressed time series X and XY stands for concatenated time series X and Y. Any compression algorithm can be used for $C(\cdot)$. In the empirical evaluation, we use gzip.

3.10 Piccolo distance

This similarity measure was introduced by Piccolo [9] as a measure of similarity between two ARIMA processes. It is defined as the Euclidean distance of the coefficients of the series' $AR(\infty)$ formulations. The coefficients of the lower order series are padded with zeros to the length of the larger order. The Piccolo distance exist for any invertible ARIMA process [8].

3.11 Prediction-based distance

This is a class of similarity measures based on the idea that two time series are similar if they are close at a specific time point in the future. Vilar et al. [11] is an implementation of this idea where forecast densities at a specific point in the future $T + h$ are compared. The distance is then calculated as an indefinite integral of absolute difference between estimates of the forecast densities for time series X and Y at time $T + h$ [8]. We set the forecast horizon h to 1 in our empirical evaluation.

3.12 Embedding-based similarity

This class of measures is based on learning a vector representation of time series and then computing their similarity using a vector similarity measure, such as Euclidean distance. We implemented this idea using Euclidean distance and Random Warping Series (RWS) to create a vector representation of time series [14]. This method uses

the DTW between the given time series and the random time series distribution. Then a family of positive definite kernels can be constructed from a map given by DTW. Optimal are found using cross-validation.

The time complexity depends on the model used but is typically $O(n)$, excluding the time we require to learn the representation. It is normalized, weakly invariant to warping, but it is not invariant to scaling. In the empirical evaluation, we use an 8-dimensional representation (16 and 32-dimensions do not lead to better results).

4 Classification of similarity measures

Wang et al. [13] propose the following categories: lock-step, elastic, threshold-based, and pattern-based measures. Montero et al. [8] propose: complexity-based, prediction-based, model-free, and model-based measures. Esling and Agon [4] propose: shape-based, edit-based, feature-based, and structure-based measures. Each of these classifications provides a different but incomplete view of similarity measures. A reconciliation is beyond the scope of this paper but for the sake of completeness, we classify the measures used in this paper according to each of the three classifications (see Table 1).

5 Empirical evaluation and comparison

We generated 9 groups of 100 time series of length 100 (see Table 2). For each pair of groups we computed the similarity for each pair of time series. We then clustered them into 2 clusters with k-medoids clustering. We evaluated the clustering using the adjusted Rand index. We repeated this process for each similarity measure.

The purpose of this experiment was twofold. First, to identify which measures give similar values (see Figure 1). And second, to highlight scenarios where a similarity measure might fail and which alternative could be used (see Table 3 for a summary).

6 Discussion

While we did not cover all, we did cover the most popular similarity measures and at least one representative from each class of similarity measures, except Spatial Assembling Distance (SpADe). Several similarity measures perform well on average, but none perform well in all cases. In some cases less known measures like compression-based similarity are better, even though they typically perform poorly. Therefore, choosing the best similarity measure for the task at hand is not a trivial task and there is value in our review of their properties. Piccolo distance stands out as the only linear complexity similarity measure with good overall performance. Our embedding-based approach achieves similar performance and also has linear time complexity, excluding the time we require to learn the embedding. However, it was learned on labelled data and is not generally applicable. The results are consistent with the results from Wang et al. [13], Serra and Arcos [10], and Górecki and Piasecki [6], where DTW and Edit distance measures performed best although Serra and Arcos [10] and Górecki and Piasecki [6] report that advanced modifications of DTW outperformed other measures.

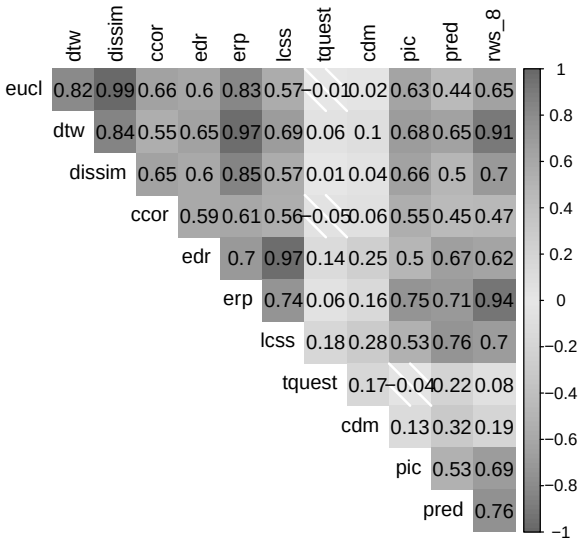


Figure 1: Pearson correlation between similarity measures across all pairs of time series from all nine groups. DTW, Euclidean distance, DISSIM, EDR, ERP, LCSS, and RWS all give numerically similar results. The following are particularly similar: Euclidean distance and DISSIM, DTW and ERP and RWS, and EDR and LCSS. CCor, PIC, and PRED are somewhat similar to each other and other methods. TQuEST and CDM strongly differ from the other measures and each other.

Further investigation of embeddings-based approaches is our main direction for further work. In particular, embeddings that are based on unlabelled data. While these approaches have been tremendously successful in image and video analysis and there have been notable applications in time series, there is no systematic treatment in the context of time series similarity and clustering. Finally, classification of time series similarities requires further work in order to reconcile the differences and inconsistencies between existing classifications and produce a more generally applicable classification.

References

[1] Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., and Renz, M. (2006). Tquest: threshold query execution for large sets of time series. In *International Conference on Extending Database Technology*, pages 1147–1150. Springer.

[2] Chen, L. and Ng, R. (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 792–803.

[3] Chen, L., Özsu, M. T., and Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 491–502.

	warping	scaling	Comp.	normalization	Wang	Esling	Montero	Diff. lengths
L_p	strong	no	$O(n)$	$\sqrt[p]{n}$	lock-step	shape	model-free	no
DISSIM	no	no	$O(n+m)$	$n-1$	lock-step	shape	*model-free*	no
DTW	strong	no	$O(nm)$	$n, m, n+m$	elastic	shape	model-free	yes
ERP	weak†	no	$O(nm)$	$max(n,m)$	elastic	edit	*model-free*	yes
EDR	weak†	no	$O(nm)$	$max(n,m)$	elastic	edit	*model-free*	yes
LCSS	weak†	no	$O(nm)$	normalized	elastic	edit	*model-free*	yes
Tquest	weak†	no	$O(nm \log nm)$	normalized	threshold	feature	*model-free*	yes
CCor	strong	yes	$O(nm)$	$(\sqrt{ml})^{-1}$	*elastic*	*feature*	model-free	yes
PIC	strong	yes	$O(n+m)$	$max(k1, k2)$	/	structure	model-based	yes
CDM	weak†	no	$O(n+m)^*$	normalized	/	structure	complexity	yes
PRED	no	no	$O(n+m)^*$	normalized	/	*structure*	prediction	yes*
RWS	weak	no	$O(n)^*$	normalized	/	*structure*	*model-based*	no

Table 1: Our classifications are in *italics*, / denotes that Wang et al. [13] treated different representations separately from similarity measures, $*$ denotes the most typical implementation but may vary with choice of underlying models, † denotes invariance based on the original definitions, where time is considered, but this vary with implementation.

group	type	noise
G1	-	N
G2	-	G
G3	linear	N
G4	linear	G
G5	linear + varying slope	N
G6	sine	N
G7	sine	G
G8	sine + varying phase	N
G9	sine + varying amplitude	N

Table 2: A summary of the 9 groups of simulated time series. N is normally distributed noise ($\mu = 0, \sigma = 0.5$), G is gamma-distributed noise ($\alpha = 0.5, \beta = 3$).

	mean	median	min
eucl	0.61	0.61	0.00
dissim	0.65	0.97	0.00
dtw	0.80	1.00	-0.00
edr	0.80	1.00	-0.00
erp	0.85	1.00	-0.00
lcss	0.83	1.00	-0.00
ccor	0.64	0.70	-0.00
cdm	0.21	0.04	-0.00
pic	0.85	1.00	0.01
pred	0.55	0.54	-0.00
tquest	0.36	0.19	-0.00
rws	0.87	1.00	0.25

Table 3: Aggregated adjusted Rand index. DTW, EDR, ERP, LCSS, CCor, PIC, and RWS are the best, on average, and CDM and TQuEST perform poorly. However, all similarity measures have poor worst-case performance - for each there is at least one pair of groups where it performs no better than random. CDM is the best at distinguishing between linear time series and linear time series with varying slope, where all other similarity measures perform poorly. Most measures also poorly distinguish between G1-G2 (normal noise vs gamma noise), G1-G5, and G2-G5 (normal or gamma noise vs linear time series with varying slope). Complete results can be found at `https://github.com/tersekmatija/ERK-2020`.

[4] Esling, P. and Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):1–34.

[5] Giorgino, T. et al. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, 31(7):1–24.

[6] Górecki, T. and Piasecki, P. (2019). A Comprehensive Comparison of Distance Measures for Time Series Classification. In *Workshop on Stochastic Models, Statistics and their Application*, pages 409–428. Springer.

[7] Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874.

[8] Montero, P., Vilar, J. A., et al. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43.

[9] Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 11(2):153–164.

[10] Serra, J. and Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, 67:305–314.

[11] Vilar, J. A., Alonso, A. M., and Vilar, J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis*, 54(11):2850–2865.

[12] Vlachos, M., Kollios, G., and Gunopulos, D. (2002). Discovering similar multidimensional trajectories. In *Proceedings 18th international conference on data engineering*, pages 673–684. IEEE.

[13] Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309.

[14] Wu, L., Yen, I. E.-H., Yi, J., Xu, F., Lei, Q., and Witbrock, M. (2018). Random warping series: A random features method for time-series embedding. *arXiv preprint arXiv:1809.05259*.

Biomedicinska tehnika

Biomedical Engineering

Pasti zajema in obdelave znotrajsrčnih elektrogramov s primeri iz klinične prakse

Jernej Štublar[1,2], Matevž Jan[2], David Žižek[2], Damijan Miklavčič[1], Tomaž Jarm[1]

[1] Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
[2]Univerzitetni Klinični center Ljubljana, Zaloška cesta 7, 1000 Ljubljana
jernej.stublar@gmail.com

Pitfalls of intracardiac electrograms recoding and processing with examples from clinical cases

Abstract. *Proper recoding, processing and clinical interpretation of intracardiac electrograms (iEGMs) is the fundamental tool in electrophysiological studies. Electrophysiological laboratory is a very challenging environment with numerous electronic devices that could introduce noise into iEGMs. Thus, considerable effort should be made to identify the source and minimize its impact as much as possible.*

With the presented clinical cases we would like to stress, that notch filter is an effective tool in minimizing 50 Hz power line noise, however, it can also significantly alter iEGM morphology. Altered iEGM morphology can lead to inaccurate clinical interpretation and consecutively in reduced clinical efficacy.

Understanding the fundamentals of iEGM recording and processing is mandatory while developing algorithms for automatic iEGM morphology analysis, that could enable electrophysiologist to assess and effectively treat tachycardias.

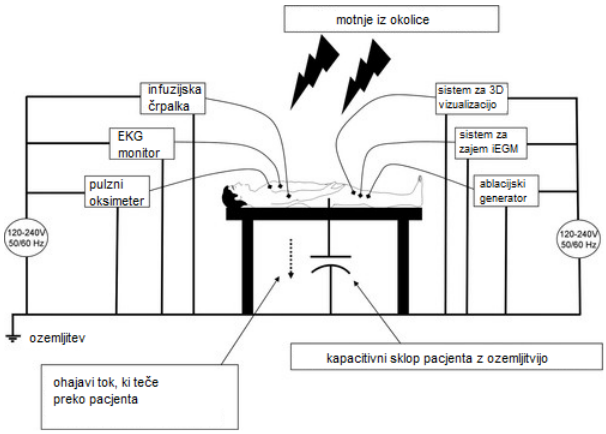
1 Uvod

Za zdravljenje prehitrega srčnega ritma (tahikardij) je danes uveljavljena minimalno invazivna srčna elektrofiziološka (EF) preiskava. Pri EF preiskavi kardiolog elektrofiziolog z namenom diagnostike ali terapije s posebnimi EF katetri preko velikih žil dostopa do notranjosti srca. Te posege omogočajo zmogljivi računalniški sistemi za tridimenzionalno (3D) vizualizacijo EF katetrov znotraj srčnih votlin [1].

Kljub petim desetletjem tehnološkega razvoja na tem področju, ostaja teorija zajema, obdelave in interpretacije znotrajsrčnih elektrogramov (iEGM) nespremenjena in še vedno predstavlja glavno orodje za razumevanje tahikardij ter posledično za klinično uspešnost vsake EF preiskave. Preko predstavljenih kliničnih primerov bomo tako poskušali opozoriti na težave, ki lahko sledijo nepravilni obdelavi ter posledično nepravilni interpretaciji signalov iEGM.

2 Elektrofiziološki laboratorij

EF laboratorij je zelo zahtevno območje za kvaliteten zajem in obdelavo signalov iEGM, saj je bolnik med preiskavo priključen na različne elektronske naprave (slika 1), ki lahko povzročajo neželene uhajave električne tokove (angl. leakage currents) [2]. Uhajavi tokovi, ki tečejo skozi bolnika v zemljo, ustvarijo na svoji poti motilne padce napetosti, ki se na zajetem signalu kljub najkvalitetnejšim diferenčnim ojačevalnikom z velikim rejekcijskim faktorjem CMRR (angl. Common-Mode-Rejection-Ratio) odrazijo kot nezanemarljiv šum (reda vsaj nekaj 10 µV). Tako ustvarjen šum ima dominantno frekvenco pri omrežni frekvenci 50 Hz, kar je znotraj uporabnega frekvenčnega območja za signale iEGM. Klinično uporabni signali iEGM imajo na primer v področju brazgotine amplitudo reda 25 µV, kar ob prisotnosti šuma predstavlja velik izziv za kvaliteten zajem in nadaljnjo obdelavo signalov iEGM.



Slika 1. Ponazoritev običajnih zunanjih izvorov šuma v signalu iEGM. Poleg elektronskih naprav priključenih na bolnika so tu še vplivi iz okolice [2].

Za dosego ustreznega razmerja med uporabnim signalom in šumom (angl. Signal-Noise-Ratio SNR), moramo v EF laboratoriju zagotoviti ustrezno zaščito (angl. shielding) ter ozemljitev vseh elektronskih naprav. Zelo pomembno je tudi, da so napajalni ali komunikacijski kabli čimbolj oddaljeni od kablov za zajem signalov. V primeru, da kabel za zajem signalov ne more biti ločen od vira šuma (na primer kabel za dovajanje RF energije med ablacijo), moramo zagotoviti, da ne potekata vzporedno. V sodobnih sistemih za zajem signalov iEGM se uporabljajo instrumentacijski ojačevalniki z velikim CMRR, ki skupni motilni signal prisoten na obeh vhodih oslabijo za vsaj 100 dB v primerjavi s koristnim signalom

iEGM, ki ga predstavlja razlika med obema vhodoma [2].

3 Zajem znotrajsrčnih signalov

Signal srčne električne aktivnosti je zapis spreminjanja potencialne razlike med dvema elektrodama. Medtem ko signal površinskega elektrokardiograma (EKG), ki ga zajemamo s površine telesa, vsebuje prispevke električne aktivnosti celotne srčne mišice, je signal iEGM zapis lokalizirane električne aktivnosti srčnih mišičnih celic v neposredni bližini zajemalnih elektrod. Iz signala iEGM lahko razberemo 3 ključne klinično pomembne podatke za razumevanje prehajanja akcijskega potenciala iz ene vzdražne srčno-mišične celice na drugo, kar imenujemo tudi depolarizacijski val. Iz signala iEGM lahko tako določimo trenutek lokalne aktivacije (angl. Local Activation Time - LAT), kar predstavlja prehod depolarizacijskega vala tik pod elektrodo, smer potovanja depolarizacijskega vala ter kompleksnost oziroma homogenost širjenja depolarizacijskega vala [3].

Pri unipolarni konfiguraciji zajema signala iEGM je ena od elektrod v stiku s preiskovanim delom srčne mišice (aktivna ali merilna elektroda, vhod +), druga elektroda pa mora biti oddaljena od srca (referenčna elektroda, vhod -). Referenčna elektroda je lahko fizično ena od EKG elektrod, elektroda za zajem signala iEGM v veliki votli veni ali pa je to Wilsonov centralni potencial (angl. Wilson central terminal - WCT), ki je iz ekstremitetnih odvodov signala EKG (standardni odvodi I, II in III) izračunano električno središče srca. Ko depolarizacijski val potuje proti aktivni pozitivni elektrodi, vidimo na signalu pozitiven odklon. Ko depolarizacijski val potuje stran od elektrode, vidimo na signalu negativni odklon. Pri potovanju depolarizacijskega vala mimo aktivne elektrode ima tako unipolarni signal iEGM značilno obliko s pozitivnim in negativnim vrhom, kjer točka največje negativne strmine (največji -dV/dt) predstavlja LAT tkiva pod aktivno elektrodo [3]. Ker unipolarni signal iEGM zajema potencialno razliko med zelo oddaljenima elektrodama, je njegova največja pomanjkljivost velik prispevek električne aktivnosti oddaljenih področij (angl. far-field signal iEGM) ter večja prisotnost šuma.

O bipolarnem signalu iEGM govorimo takrat, ko sta pozitivna in negativna elektroda blizu skupaj in je tako zajeti signal iEGM razlika električne aktivnosti tkiva med obema elektrodama. Ključna prednost tako zajetega signala je, da je šum bolje odstranjen. Bližje kot sta obe elektrodi (v klinični praksi se uporablja razdalja 2 mm ali manj), bolj je v takem bipolarnem signalu iEGM izražena lokalna komponenta (angl. near field signal iEGM), ki je ključna za pravilno klinično interpretacijo signalov iEGM in razumevanje različnih tahikardij.

Amplituda bipolarnega signala iEGM je zelo odvisna od smeri aktivacije glede na smer zveznice med obema elektrodama in je največja v primeru širjenja aktivacije vzdolž te zveznice (od ene elektrode proti drugi) ter je enaka nič v primeru širjenja aktivacije pravokotno na smer zveznice med elektrodama. Polega tega vplivajo na amplitudo bipolarnega iEGM tudi hitrost širjenja aktivacije, masa aktiviranega mišičnega tkiva ter razdalja med elektrodama. Kljub temu, da na amplitudo bipolarnega iEGM vpliva toliko faktorjev, je prav razlika amplitude med najnižjim ter najvišjim vrhom (angl. peak-to-peak voltage) najpogosteje uporabljen objektivni parameter v klinični praksi. Na primer, če je v nekem področju prekata napetost med obema vrhoma bipolarnega signala iEGM večja kot 1,5 mV, je to najverjetneje področje normalne električne aktivnosti in s tem zdrave srčne mišice. To je splošno sprejeto pravilo v trenutni klinični in raziskovalni praksi [3]. V primeru bipolarnih signalov iEGM zajetih v preddvorih pa je področje brazgotine definirano s peak-to-peak amplitudo manjšo kot 0,1 mV [4].

4 Obdelava znotrajsrčnih signalov

Vsi trenutno dostopni sistemi za zajem signalov iEGM omogočajo sočasen zajem vsaj 40 signalov iEGM, njihovo analogno-digitalno pretvorbo z vzorčno frekvenco vsaj 1.000 Hz. Naprave imajo vgrajene različne filtre za izločitev motenj ter omejitev pasovne širine signala samo na območje s koristno elektrofiziološko vsebino. Velja namreč, da prispevek različnih vrst šuma v zajetem signalu narašča s širjenjem pasovne širine [5]. V trenutni klinični praksi velja, da je glavnina uporabne informacije signala iEGM v območju od 0 Hz do 300 Hz (tabela 1) [3].

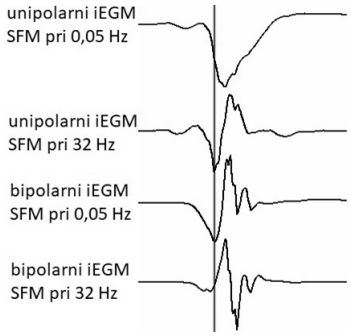
4.1 Vpliv visokoprepustnega filtra

V nizkofrekvenčnem delu spektra signala iEGM se poleg uporabne informacije nahaja tudi šum zaradi premikanja srčnega katetra (mehanska motnja zaradi dihanja in utripanja srca), kar povzroči nizkofrekvenčno premikanje osnovnega nivoja signalov (angl. baseline drift) [2]. Postavitev spodnje frekvenčne meje pasovno prepustnega filtra za unipolarni signal iEGM je v klinični praksi tako vedno kompromis med iskano uporabno informacijo (slika 2) in stabilnostjo prikaza signala iEGM. Omenili smo že, da bipolarni signal iEGM nima toliko uporabne informacije pri nizkih frekvencah, saj ga ne uporabljamo za določitev smeri širjenja depolarizacijskega vala, zato se je za spodnjo mejo uporabnega frekvenčnega območja bipolarnih signalov iEGM uveljavila meja pri 30 Hz (tabela 1).

4.2 Vpliv nizkoprepustnega filtra

V trenutni klinični praksi lahko zajemamo signale iEGM z vzorčno frekvenco 4.000 Hz ter zgornjo mejo pasovno prepustnega filtra pri 2.000 Hz. S frekvenčno analizo zajetega iEGM signala smo podobno kot sledi iz iz literature [3] potrdili, da je večina uporabnega signal iEGM pri frekvencah nižjih od 300 Hz. V klinični

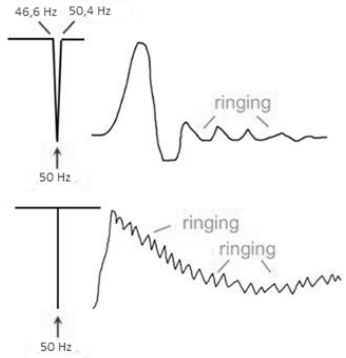
praksi smo tako obdržali splošno uveljavljeno vzorčno frekvenco pri 1.000 Hz ter zgornjo mejo pasovno prepustnega filtera pri 500 Hz (tabela 1) kar izpolnjuje tudi zahteve Shannon-Nyquistovega teorema, kjer mora biti vzorčna frekvenca dvakrat višja od željene frekvenčne vsebine zajetega signala.



Slika 2. Ponazoritev oblike unipolarnega in bipolarnega signala ter vpliva spodnje frekvenčne meje (SFM) pasovno prepustnega filtra. Črta predstavlja točko LAT. (Prirejeno po: slideshare.net, 2019)

4.3 Vpliv pasovno zapornega filtra

Za izločitev šuma omrežne frekvence 50 Hz ima vsak sistem za zajem signalov iEGM možnost uporabe dodatnega pasovno zapornega filtra (angl. notch filtra), ki iz signala odstrani vsebino v ozki okolici nastavljive centralne frekvence (pri nas 50 Hz). Z vklopom notch filtra pa žal neizogibno odstranimo tudi del koristnega signala iEGM ter vnesemo popačitve v signal zaradi faznega zamika ter tako imenovanega ringing efekta (slika 3) [6].



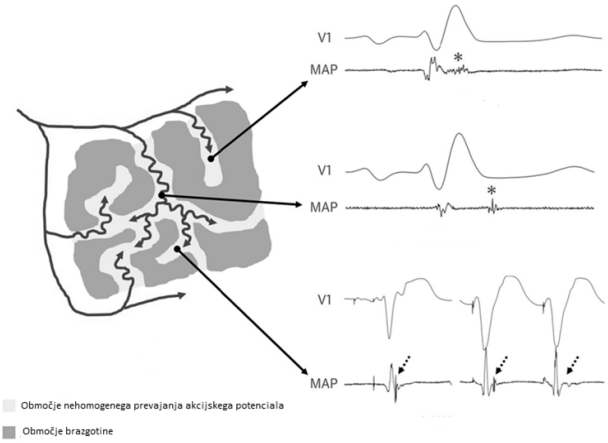
Slika 3. Prikaz vpliva realnega notch filtra na zvonjenje (zgoraj). Teoretični prikaz vpliva idealnega notch filtra, ki lahko zvoni v neskončnost (spodaj) (Prirejeno po: [6]).

Tabela 1. Pregled uporabljenih spodnjih in zgornjih mej pasovno prepustnih filtrov pri zajemu elektrofiziološke aktivnosti srca v trenutni klinični praksi.

	Zgornja meja	**Spodnja meja**
Signal EKG	0,5 Hz	100 Hz
Bipolarni signal iEGM	30 Hz	500 Hz
Unipolarni signal iEGM	0,5 – 2 Hz	500 Hz
Nefiltrirani unipolarni signal iEGM	0,05 Hz	500 Hz

5 Primeri iz klinične prakse

V okolici frekvence 50 Hz imajo največ uporabne informacije signali iEGM pljučnih ven ter tudi nizko-amplitudni kompleksni fragmentirani signali iEGM (signali z več kot dvema vrhovoma), ki so posledica nehomogenega širjenja depolarizacijskega vala v področju brazgotine (slika 4) in povzročajo t.i. krožečo tahikardijo [2, 7, 8]. V klinični praksi je odkritje in odprava kompleksnih fragmentiranih signalov iEGM ključna za klinično uspešnost vsake EF preiskave, vendar je kompleksen fragmentiran signal iEGM lahko tudi posledica popačitev ob nepravilnem zajemu teh signalov [8].

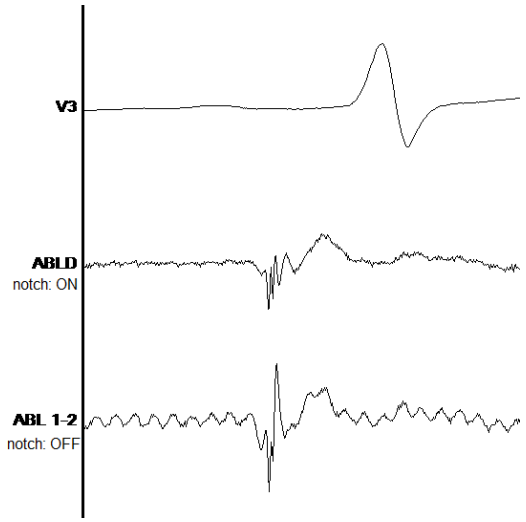


Slika 4. Prikaz zajetega signala iEGM na različnih področjih brazgotine. Z zvezdico je označen kompleksen fragmentiran signal IEGM. S puščico je prikazano razkritje (zakasnitev) lokalne komponente signala iEGM ob stimulaciji. (Prirejeno po: [6]).

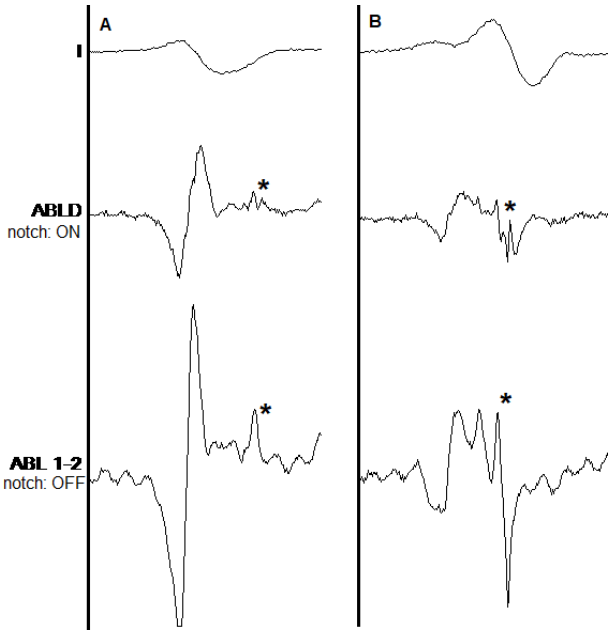
Popačitve so lahko neusklajene z utripanjem srca, na primer kot posledica premikanja (dihanja), medsebojnega dotikanja dveh merilnih elektrod EF katetra ter električne aktivnosti okoliških mišic (ne srčne mišice) [8]. Te popačitve lahko v klinični praksi dokaj enostavno prepoznamo, saj mora biti opazovani signal iEGM ponovljiv preko več utripov ob stabilni legi EF katetra. Za klinično interpretacijo so tako bolj problematične popačitve, ki so usklajene z utripanjem srca in se pojavljajo ob vsakem utripu.

Uporaba notch filtra nam v rutinski klinični praksi izboljša razmerje med uporabnim signalom in šumom v signalu iEGM, vendar se moramo hkrati zavedati pasti take obdelave. Zaradi faznega zamika, ki ga vnaša uporabljeni filter, se lahko zelo spremeni oblika signala iEGM, kjer se lahko amplituda zmanjša tudi za polovico in se pojavijo dodatni vrhovi, ki spominjajo na prej opisane patološke kompleksne fragmentirane signale iEGM (slika 5) [8, 9].

Oblika preučevanega signala iEGM se lahko spremeni tudi zaradi že omenjenega ringing efekta, ki ob uporabi notch filtra vnaša popačitve levo in desno od maksimalnega vrha preučevanega signala iEGM, kar nas lahko zavede, da normalen signal iEGM zamenjamo za patološki (slika 6) [9], kar ima za posledico neustrezno klinično interpretacijo.



Slika 5. Prikaz zajetega signala iEGM z EF ablacijskim katetrom v stiku z zdravo srčno mišico. ABLD z vključenim notch filtrom, ABL 1-2 brez vključenega notch filtra pri istih nastavitvah pasovno prepustnega filtra (30 Hz - 500 Hz).



Slika 6. Prikaz zajetega signala iEGM z EF ablacijskim katetrom v področju brazgotine, kjer je z zvezdico označena pozna komponenta signala iEGM. A: oddaljena komponenta signala iEGM se zaradi notch filtra spremeni v patološki signal iEGM. B: lokalna komponenta signala iEGM normalnih amplitud se zaradi notch filtra spremni v patološki signal iEGM.

6 Zaključek

Pri klinični interpretaciji se zanašamo na specifične vzorce in obliko signala iEGM, ki se lahko povsem spremeni ob neustrezni obdelavi signala iEGM. V tem pogledu je najbolj problematična uporaba notch filtra, kar smo pokazali na treh kliničnih primerih, ki so po našem mnenju dovolj nazorni, da se notch filter preneha uporabljati v klinični praksi. V vsakem EF laboratoriju moramo tako storiti vse, da zmanjšamo 50 Hz šum iz omrežja in tako odpravimo potrebo po uporabi notch filtra.

V zadnjem obdobju se razvijajo avtomatski algoritmi za analizo oblike signalov iEGM s pomočjo katerih bi lahko razlikovali med lokalno in oddaljeno komponento signala iEGM, kar bi bistveno pripomoglo k razumevanju in klinični uspešnosti zdravljenja tahikardij. Pri razvoju teh avtomatskih algoritmov je ključno poznavanje osnovnih zakonitosti zajema in obdelave signalov iEGM, pri čemer mora biti analiza osnovana na realnih surovih signalih iEGM brez vnesenih umetnih popačitev.

Literatura

[1] J. Štublar, T. Jarm: Zdravljenje motenj srčnega ritma, Življenje in tehnika 11/19, november 2019.

[2] KL. Venkatachalam, JE. Herbrandson, SJ. Asirvantham: Signals and Signal Processing for the Electrophysiologist Part I: Electrogram Acquisition. Circ Arrhythm Electrophysiol. 2011;4:965-973.

[3] ZF. Issa, JM. Miller, DP. Zipes: Clinical Arrhythmology and Electrophysiology 2nd edition, 2012.

[4] NMS. De Groot, MJ. Schalij, K. Zeppenfeld, et al.: Voltage and Activation Mapping How the Recording Technique Affects the Outcome of Catheter Ablation Procedures in Patients With Congenital Heart Disease. Circulation 2003; 2099-2106.

[5] Northrop, R. (2018). Non-Invasive Instrumentation and Measurement in Medical Diagnosis. Boca Raton: CRC Press, https://doi.org/10.1201/9781315117690

[6] SS. Po: Warren Jakman's Art of War: A Sniper's Approach to Catheter Ablation, 2019.

[7] SR. Dukkipati, JS. Koruth, S. Choudry, et al.: Catheter Ablation of Ventricular Tachycardia in Structural Heart Disease Indications, Strategies, and Outcomes—Part II. Journal of the American College of Cardiology. 2017, 70(23), 2924–2941. doi:10.1016/j.jacc.2017.10.030.

[8] JMT. De Bakker, FHM. Wittkampf: The Pathophysiologic Basis of Fractionated and Complex Electrograms and the Impact of Recording Techniques on Their Detection and Interpretation. Circulation : Arrhythmia and Electrophysiology. 2010, 3(2), 204–213. doi:10.1161/circep.109.904763.

[9] M. Martínez-Iniesta, J. Rodenas, R. Alcaraz, JJ. Rieta: "Application of the stationary wavelet transform to reduce power-line interference in atrial electrograms," 2017 Computing in Cardiology (CinC), Rennes, 2017, pp. 1-4, doi: 10.22489/CinC.2017.112-049.

Izdelava fantoma za ultrazvočno medicinsko diagnostiko s postopkom 3D tiskanja

Nejc Klanjšček[1], Sašo Arnuga[2], Miha Fošnarič[2], Samo Penič[1]

[1]*Fakuteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, SI-1000 Ljubljana, Slovenija*
[2]*Zdravstvena fakulteta, Univerza v Ljubljani, Zdravstvena pot 5, SI-1000 Ljubljana, Slovenija*
E-pošta: samo.penic@fe.uni-lj.si

3D-printed Medical Ultrasound Phantom

Medical ultrasound is today one of the most common and accessible diagnostic imaging method in medicine. No ionizing radiation and low heat exposure o the tissue makes modern ultrasonography a diagnostic technique with almost no health threat. However, quality control of the ultrasound equipment must still be adequate, otherwise safety of the patient can be endangered, especially from misdiagnosis.

The 3D printing technology is becoming widely available and therefore provides efficient means for producing complicated structures. In the presented research, the method of 3D printing was applied to develop and build a low-cost tissue mimicking body (phantom). Such phantoms are used to test ultrasonic probes and for scientific and educational purposes. In the process of developing a phantom for ultrasonic imaging, several challenges had to be overcome, such as choosing appropriate materials that mimicked human tissues or keeping the phantom watertight.

1 Uvod

Ultrazvočne preiskave so najpogostejša in najdostopnejša slikovna diagnostična metoda v medicini, saj je relativno varna, cenovno ugodna in sorazmerno enostavna za uporabo [1, 2, 3]. Zaradi odsotnosti ionizirajočega sevanja je prva izbira v predrojstveni diagnostiki (Slika 1) in drugih medicinskih preiskavah, kjer je varnost prioriteta.

Akustična impedanca, z, je lastnost snovi, ki nam pove, kako se snov odziva na tlak, ki ga povzroča njeno mehansko valovanje. V preprostem primeru ravnega vala brez faznega zamika med tlakom, p, in hitrost delca snovi, v, velja preprosta zveza $z = p/v$. Izkaže se [1], da lahko takšno akustično impedanco izrazimo kot produkt

Fizikalna osnova ultrazvočnih diagnostičnih naprav je mehansko valovanje nad človeško slušno mejo (frekvence nad 20 kHz), ki ga sonda naprave v obliki pulzov oddaja in zaznava njihov odboj od različnih delov znotraj telesa. Pri tem izkorišča dejstvo, da je na meji dveh snovi, skozi kateri se ultrazvok širi, njegova odbojnost sorazmerna z razliko v akustičnih impedancah obeh sredstev [1]. Akustična impedanca, z, je lastnost snovi, ki nam pove, kako se snov odziva na tlak, ki ga povzroča njeno mehansko valovanje. V preprostem primeru ravnega vala brez faznega zamika med tlakom, p, in hitrost



Slika 1: Ultrazvočni posnetek človeškega plodu v 20. tednu nosečnosti [4].

delca snovi, v, velja preprosta zveza $z = p/v$. Izkaže se [1], da lahko takšno akustično impedanco izrazimo kot produkt

$$z = \rho c, \tag{1}$$

kjer je ρ gostota snovi in c hitrost širjenja valovanja. V tabeli 1 je nekaj akustičnih impedanc v človeškem telesu. Odbojnost, R, na meji dveh snovi lahko zapišemo kot [1]

Organ	**Impedanca [$\mathrm{kg/(m^2 s)}$]**
Pljuča	$0,18 \cdot 10^6$
Ledvice	$1,63 \cdot 10^6$
Jetra	$1,65 \cdot 10^6$
Mišice	$1,71 \cdot 10^6$
Kosti	$7,8 \cdot 10^6$
Voda	$1,48 \cdot 10^6$
Zrak	$0,0004 \cdot 10^6$

Tabela 1: Akustične impedance v človeškem telesu

$$R = \frac{I_r}{I_i} = \left(\frac{z_2 - z_1}{z_2 + z_1}\right)^2, \tag{2}$$

kjer sta I_r in I_i jakosti (povprečni gostoti energijskih tokov) odbitega in vpadnega valovanja, z_1 in z_2 pa aku-

stični impedanci snovi po kateri valovanje pride do mejne mejne ploskve in od katere se odbije.

Na podlagi jakosti odboja in pretečenega časa med oddajo in sprejemom signala nato naprava rekonstruira sliko in jo v realnem času izriše na zaslon.

Za kontrolo kakovosti ultrazvočnih naprav za medicinsko diagnostiko, pa tudi za pedagoško in raziskovalno delo, se pogosto uporablja testne naprave, *t.i.* utrazvočne fantome, ki posnemajo fizikalni odziv človeškega tkiva na ultrazvok, ki ga oddajajo diagnostične sonde [5, 6, 7, 8].

Tržišče ponuja ultrazvočne fantome za različne namene, vendar pa so ti zaradi cene nedostopni in/ali zaradi poslovnih skrivnosti izdelave neprimerni za izobraževanje in raziskovalno delo. Po drugi strani pa sodobna tehnologija 3D tiskanja omogoča izdelavo skoraj poljubnih trirazsežnih struktur iz digitalnega modela. Tako lahko ob simuliranju primernih fizikalnih lastnosti sami izdelamo cenovno ugodne ultrazvočne fantome [9].

V prispevku predstavljamo poskus lastne izdelave ultrazvočnega fantoma s pomočjo aditivne proizvodnje oziroma 3D tiskanja. V nadaljevanju najprej natančno opišemo uporabljeno metodo, nato predstavimo rezultate testiranja izdelanega prototipa z ultrazvočno diagnostično napravo in v diskusiji med drugim predstavimo naslednje korake za izboljšavo prototipa.

2 Metoda 3D tiskanja

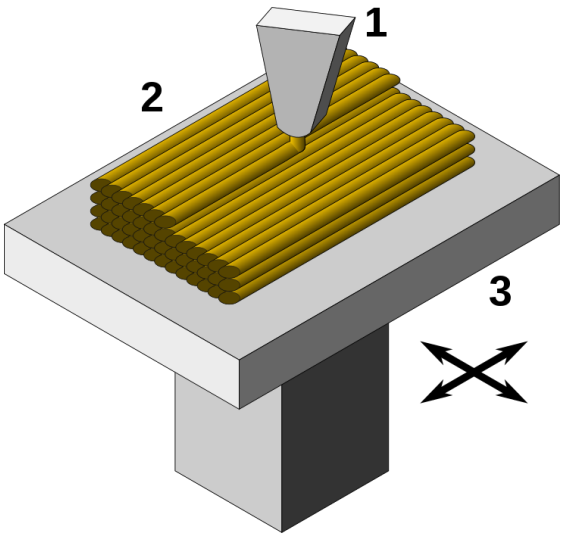
Metoda 3D tiskanja spada med t.i. aditivne metode, katerih skupna lastnost je, da material h končnemu modelu (izdelku) dodajamo. V tem pogledu se razlikuje od tradicionalnih metod, kjer navadno surovemu obdelovancu material odvzemamo (subtraktivne metode). Pod krovnim izrazom 3D tiskanja poznamo množico različnih tehnologij kot so "Fused deposition Modelling" ali FDM, "Powder bed fusion", "Inkjet printing", Stereolitografija ali SLA in druge [10]. Širši javnosti sta najdostopnejši metodi sta SLA in FDM.

Najpogosteje uporabljena metoda 3D tiskanja je metoda FDM, kjer tiskalnik nalaga plasti materiala tako, da stali surovino (običajno v obliki plastičnega filamenta) in iz taline gradi model. Principa delovanja prikazuje slika 2. Na voljo je več vrst plastičnih materialov, primernih za uporabo v tiskalnikih tega tipa. Velikost modela je pogojena z velikostjo mize na kateri gradimo model, ta pa je omejena z dolžino vodil, po katerih koračni motorji premikajo mobilno mizo in/ali glavo ektruderja.

Večino delov fantoma smo izdelali na FDM 3D tiskalniku Anycubic i3 Mega, poizvajalca Anycubic. Po navedbah proizvajalca je ločljivost tiskanja $12,5$ µm v ravnini in 2 µm v višino, velikost delovne površine pa je 210×210 mm. Največja dovoljena višina tiskanega modela je 205 mm. Priporočene hitrosti tiskanja so 20–100 mm/s. Uporabili smo glavo s šobo premera 0,4 mm.

3 Načrtovanje in izdelovanje fantoma

Za načrt in izdelavo 3D modela fantoma smo uporabili programsko opremo Solidworks 2017 podjetja Dassault Systèmes. Na sliki 3 sta prikazana dva glavna sestavna



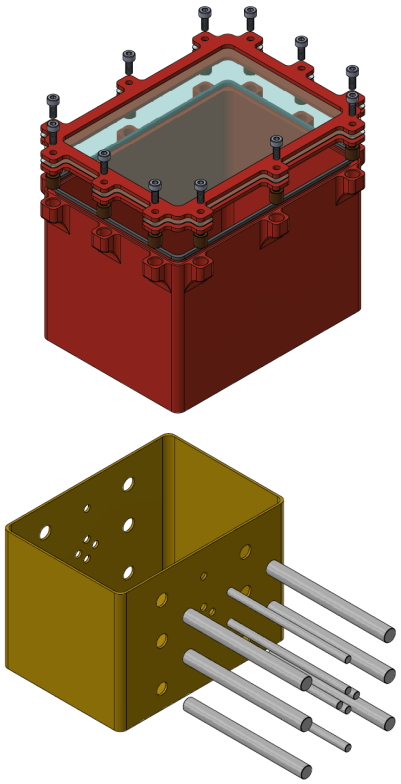
Slika 2: FDM metoda 3D tiskanja: segreta kovinska glava ekstruderja stali plastiko v filamentu (1) in jo naloži v plasteh tankih kolobarjev na model (2). Mobilna miza se premika v dveh ali treh oseh in s tem omogoča oblikovanje strukture (3) [11].

dela fantoma. Zgornji del slike 3 je tri-dimenzionalni pogled na ohišje, škatlo, v katero vstavimo notranje ogrodje s prečnimi palicami. Na isti sliki spodaj pa je prikazano ogrodje in prečne palice, ki služijo kot ovire za ultrazvočno valovanje – fantomski objekti, ki simulirajo različna tkiva v telesu. Za takšno konstrukcijo smo se odločili predvsem zato, ker smo želeli spreminjati geometrijo samih prečnih palic in s tem primerjati rezultate. Palice so iz materialov različnih akustičnih impedanc, katere posnemajo obnašanje različnih tkiv v človeškem telesu.

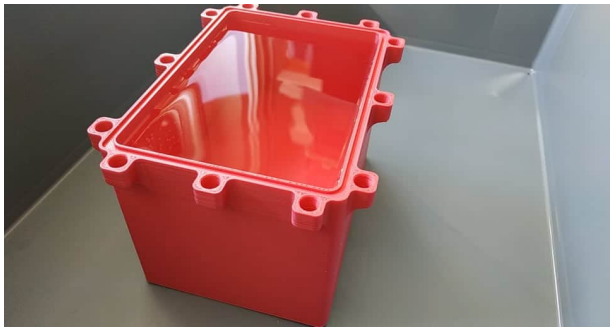
Notranjost fantoma je zapolnjena s tekočino, ki naj bi bila kar najbolj akustično podobna različnim tkivom v človeškem telesu. Pregled snovi, primernih za simuliranje različnih tkiv v človeškem telesu smo našli v obstoječi literaturi [12].

Pri samem načrtovanju smo naleteli na kar nekaj izzivov. Eden večjih je bila vodotestnost fantoma (slika 4). Pri prvi izvedbi smo tiskali fantoma pri preveliki hitrosti ter tako naredili stene porozne za vodo. To smo poskušali odpraviti s slojem iz bitumenske mase, s katero smo premazali notranjost fantoma. Ta tehnika se ni obnesla in je bilo potrebno natisniti še eno ohišje fantoma. To ohišje smo prilagodili tako, da smo zamnjšali debelino sten fantoma in izbrali manjšo hitrost tiskanja.

Čas tiskanja ohišja fantoma je bil 34 ur, ogrodja za notranjo konstrukcijo plastičnih ovir iz palic pa 12 ur. Nizka hitrost tiskanja in povečana količina izločene plastike (105%) sta bila nujno potrebna parametra, da smo dobili plasti ohišja fantoma, ki so bile trdno povezane in so dobro zatesnile stene. STL datoteko smo pripravili v rezalniku (angl. "slicer") Cura 4.2, podjetja Ultimaker, pri čemer smo uporabili naslednje nastavitve, kateri so bili ključni za kvalitetno izdelano ogrodje:



Slika 3: Načrt fantoma. Zgoraj: ohišje, škatla, v katero vstavimo notranje ogrodje s prečnimi palicami. Spodaj: Nosilna konstrukcija s prečnimi ovirami okroglega premera.

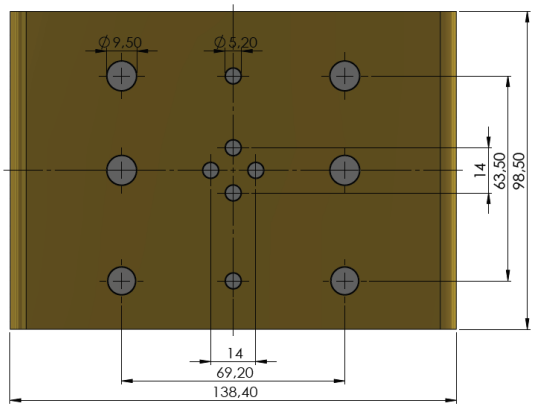


Slika 4: Testiranje vodotesnosti natisnjenega ohišja fantoma.

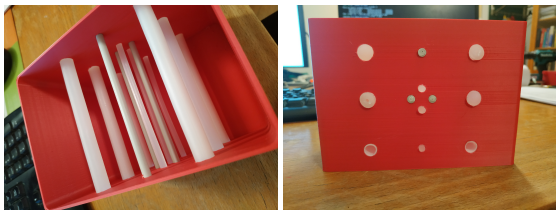
- Layer Height: $0,15$ mm
- Material bed temperature: 55°C
- Nozzle temperature: 210°C
- Print speed: 60 mm/s
- Infill density: 70%
- Flow rate extrusion offset factor: 105%
- Cool fan speed: 70%

Za ohišje in notranje ogrodje fantoma smo uporabili plastiko PLA (poliaktična kislina), zaradi enostavnosti tiskanja, biološke razgradljivosti in dobre mehanske trdnosti. Plastični filament je mat rdeče barve, premera 1,75 mm, gostote $1,24\ \mathrm{g/cm^3}$, priporočene temperature tiskanja $180 - 220$°C, proizvajalca Plastika Trček.

Ko smo imeli vodotesno ohišje, smo izdelali še nosilec fantomskih elementov – v našem primeru palic z različnimi akustičnimi impedancami, predstavljenimi v tabeli 2. Notranjost fantoma je zasnovana modularno, saj lahko palice menjavamo in spreminjamo njihov položaj v fantomu (sliki 5 in 6):



Slika 5: Postavitev prečnih ovir v nosilnem ogrodju



Slika 6: Struktura nosilne konstrukcije in prečnih palic iz različnih plastičnih materialov.

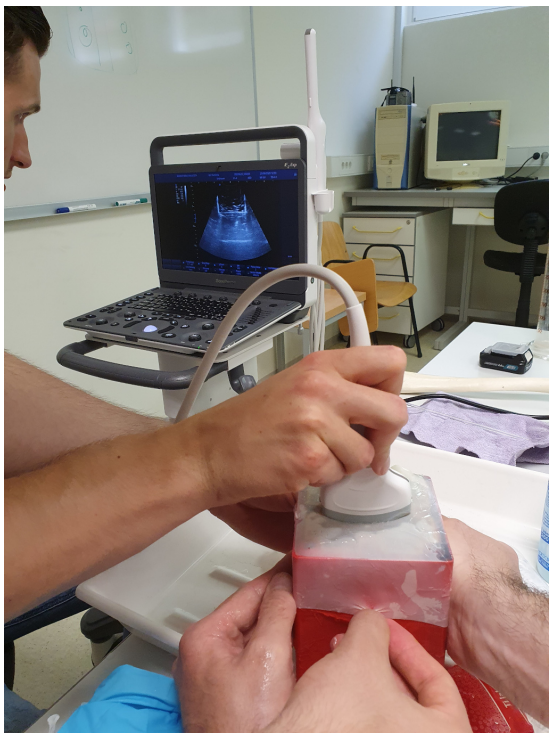
Tudi pri izdelavi pokrova fantoma smo naleteli na nekaj vprašanj, namreč kako zatesniti pokrov in iz kakšnega materiala ga izdelati. Zatesnili smo ga s pomočjo posebej natisnjenega tesnila iz elastičnega plastičnega materiala TPU (Plastika Trček). Nato smo izdelali kar nekaj pokrovov iz različnih materialov in različnih debelin. Pri tem smo ugotovili, da pokrov ne sme biti predebel in za najprimernejše se je izkazalo, da je izdelan iz tanke gume oz. lateksa (slika 7).

4 Razprava in zaključek

Z uporabo metode 3D tiskanja smo z izjemo palic različnih parametrov za ultrazvočno valovanje in membrane na pokrovu, uspeli natisniti prototip fantoma. Pri prvih meritvah smo nas slikah uspešno izmerili razdalje med sondo in palicami, ki so ustrezale razdaljam modela. Soočili smo se z nekaterimi problemi, kot so izbira primernega materiala in parametrov tiskanja, da smo dosegli tesnenje fantoma.

Tabela 2: Lastnosti materialov za prečne ovire (filamente) v fantomu.

Material	**Hitrost zvoka (m/s)**	**Gostota (g/cm^3)**	**Akustična impedanca [$\mathrm{kg}/(\mathrm{m}^2\mathrm{s})$]**
ABS (siva)	2170	1,07	$2,30 \cdot 10^6$
Akril	2750	1,18	$3,25 \cdot 10^6$
Najlon	2600	1,12	$2,91 \cdot 10^6$
Polikarbonat (prosojen)	2270	1,18	$2,68 \cdot 10^6$
Polietilen	1950	0,90	$1,56 \cdot 10^6$
Polietilen (HD)	2430	0,96	$2,33 \cdot 10^6$
Polietilen (LD)	1950	0,92	$1,79 \cdot 10^6$
PVC	2380	1,38	$3,28 \cdot 10^6$



Slika 7: Merjenje razdalj med palicami z ultrazvokom skozi membrano iz kirurške rokavice iz lateksa.

Pri nadaljnih raziskavah bomo poskusili izdelati fantoma z dolgotrajno stabilno zapolnitvijo s primerno tekočino ter z dodatki za merjenje pretokov tekočin (z t.i. Doppler UZ).

Zahvala. Ideja za pričujoče delo se je razvila iz projekta *Odprtokodni sistem za zajem in obdelavo medicinskih slik iz ultrazvočnih diagnostičnih naprav po DICOM standardu*, ki smo ga med marcem in junijem 2020 izvedli v okviru programa *Po kreativni poti do znanja*. Zato se avtorji zahvaljujemo *Javnemu štipendijskemu, razvojnemu, invalidskemu in prežívninskemu skladu RS*, *Ministrstvu za izobraževanje, znanost in šport RS*, *Evropskemu socialnemu skladu EU*, *Zdravstveni fakulteti UL* ter *Univerzi v Ljubljani* za finančno in logistično podporo. In nenazadnje seveda tudi vsem sodelujočim na projektu: podjetju *ECHOSON d.o.o.*, mentorju prof. dr. Janezu Žibertu in študentom Manci Pišek, Pii Pikelj, Sari Kidrič, Ninu Kleindienstu in Žanu Regoršku.

Literatura

[1] A. Thrush, P. R. Hoskins, and K. Martin, *Diagnostic Ultrasound: Physics and Equipment*. CRC Press, 2019.

[2] V. Jevtič, J. Matela, and Šurlan M., *Diagnostična in intervencijska radiologija. Splošni del.* Založba Pivec, 2014.

[3] S. Arnuga, *Izobraževanje s področja ultrazvoka za radiološke inženirje: primerjava evropskih izobraževalnih programov in predlog izobraževanja v Sloveniji.* Ljubljana: Zdravstvena fakulteta, 2019.

[4] Wikipedia contributors, "Ultrasound — Wikipedia, the free encyclopedia," 2020. [Na spletu; dostopano 20. junij 2020].

[5] Y. H. Kim, "Ultrasound phantoms to protect patients from novices," *The Korean journal of pain*, vol. 29, no. 2, p. 73, 2016.

[6] K. Nattagh, T. Siauw, J. Pouliot, I.-C. Hsu, and J. A. Cunha, "A training phantom for ultrasound-guided needle insertion and suturing," *Brachytherapy*, vol. 13, no. 4, pp. 413–419, 2014.

[7] G. Carbajal, A. Lasso, Á. Gómez, and G. Fichtinger, "Improving n-wire phantom-based freehand ultrasound calibration," *International journal of computer assisted radiology and surgery*, vol. 8, no. 6, pp. 1063–1072, 2013.

[8] E. D. Ehler, B. M. Barney, P. D. Higgins, and K. E. Dusenbery, "Patient specific 3d printed phantom for imrt quality assurance," *Physics in Medicine & Biology*, vol. 59, no. 19, p. 5763, 2014.

[9] C. Kollmann, D. Dubravskỳ, and B. Kraus, "An easy-to-handle speed of sound test object for skills labs using additive manufacturing (raptus-sos)," *Ultrasonics*, vol. 94, pp. 285–291, 2019.

[10] T. D. Ngo, A. Kashani, G. Imbalzano, K. T. Nguyen, and D. Hui, "Additive manufacturing (3d printing): A review of materials, methods, applications and challenges," *Composites Part B: Engineering*, vol. 143, pp. 172–196, 2018.

[11] Wikipedia contributors, "Fused filament fabrication — Wikipedia, the free encyclopedia," 2020. [Na spletu; dostopano 20. junij 2020].

[12] M. O. Culjat, D. Goldenberg, P. Tewari, and R. S. Singh, "A review of tissue substitutes for ultrasound imaging," *Ultrasound in medicine & biology*, vol. 36, no. 6, pp. 861–873, 2010.

Sklopljena numerična simulacija temperaturnega profila linearne pretočne komore med dovajanjem električnih pulzov

Peter Lombergar[1], Samo Mahnič-Kalamiza[1], Karel Flisar[1], Damijan Miklavčič[1]

[1] Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: peter.lombergar@fe.uni-lj.si

Coupled simulations of the temperature profile in a parallel plate treatment chamber during Pulsed Electric Fields treatment

The main goal of coupled simulations in Pulsed Electric Field (PEF) application is to study spatial and temporal temperature profile in treatment chamber during the PEF process, which is critical, since local increases in temperature can affect quality of end product, e.g. food.

In the study we present the use of a coupled time-dependent simulation to investigate the temperature profile in a parallel plate treatment chamber. Each pulse was simulated separately, which provided insight into temperature changes in fluid during the electrical pulse, and accurate analysis of the temperature profile.

Experiments were performed with 0.18 % NaCl solution, syringe pump, parallel plate treatment chamber, optical thermometer, and pulse generator, which was used to generate a train of electric pulses.

The model was validated by comparing the increase in temperature at the outlet of the treatment chamber in the model to data obtained in experiment. Effect of different maximum pulse voltages and different flow velocities on the increase in temperature at the outlet was simulated. The temperature profile inside the treatment chamber during application of the last electrical pulse in a pulse train was studied, where high local increases of temperature were observed.

1 Uvod

Izpostavitev celic ali tkiva kratkim električnim pulzom z dovolj visoko amplitudo, da dosežemo povišanje prepustnosti celične membrane, imenujemo elektroporacija, oziroma na področju prehranske industrije, obdelava s pulzirajočim električnim poljem (angl. PEF – Pulsed Electric Field treatment) [1].

Pomembna lastnost elektroporacije je, da je univerzalna za vse vrste celic (evkarionti, bakterije, arheje) v vseh različnih konfiguracijah (celice pritrjene na površini, suspenziji, tkivu) [2]. Prav zaradi te univerzalnosti se danes elektroporacija uporablja na različnih področjih; v medicini [3], biotehnologiji [4], okoljevarstvu [5], [6] in prehranski industriji [7].

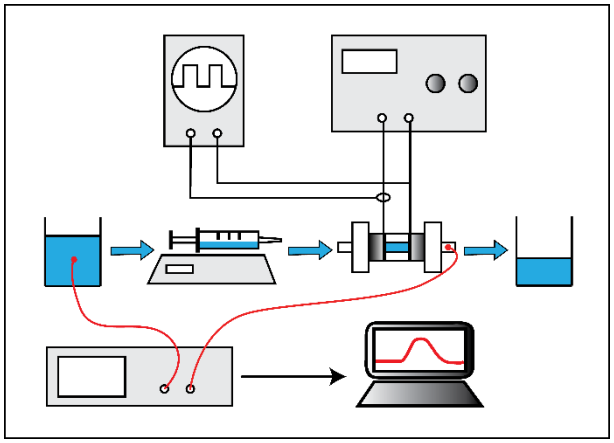
V prehranski industriji obdelava s PEF zajema ekstrakcijo sokov in ostalih pomembnih komponent iz rastlinskih tkiv in mikroorganizmov, dehidracijo tkiv in ne-toplotno konzervacijo tekoče hrane z inaktivacijo mikrobov [8]. Čeprav pri obdelavi hrane s PEF pride do segrevanja zaradi ohmskega gretja, je temperatura procesa večinoma dovolj nizka, da lahko PEF štejemo med ne-toplotne metode obdelave hrane. Tako je veliko bolj privlačna za obdelavo hrane kot konvencionalne metode, saj so pri obdelavi hrane s PEF izgube teksture, vonja in okusa hrane zanemarljive [9].

PEF sistem za obdelavo hrane sestavljajo pulzni generator, pretočna komora, pretočni sistem in sistem za nadzorovanje temperature in električnih pulzov [10]. Pretočna komora predstavlja eno glavnih komponent PEF sistema, saj je v njej hrana izpostavljena električnim pulzom. Sestavljena je iz vsaj dveh elektrod in izolacijskega materiala v različnih geometrijskih konfiguracijah, ki privedejo do različnih razporeditev električnega polja v komori. Najbolj uporabljene oblike pretočnih komor so linearna, ko-linearna in koaksialna, pri čemer se linearna pretočna komora ponaša z najbolj homogenim električnim poljem v kanalu med obema elektrodama [7].

Numerične simulacije na področju obdelave hrane s PEF lahko v grobem ločimo v dve skupini, pri čemer se prva ukvarja s simulacijami električnega polja v različnih pretočnih komorah. Namen simulacij električnega polja je optimizacija parametrov geometrije pretočnih komor za izboljšanje homogenosti električnega polja znotraj pretočne komore. Druga skupina simulacij na področju PEF, ki je za naš prispevek bolj relevantna, pa se ukvarja s sklopljenimi simulacijami električnega polja, pretoka tekočine in segrevanja. Namen teh simulacij je boljše poznavanje temperaturnega profila pretočnih komor med PEF procesom, kar je kritično, saj lahko lokalni dvigi temperature znotraj pretočne komore (kot posledica nehomogenosti električnega polja ali pretoka tekočine) poslabšajo kvaliteto hrane [11]. Pregled literature kaže, da so bile do sedaj na področju sklopljenih simulacij uporabljene večinoma simulacije stacionarnega stanja, pri čemer pri simulaciji segrevanja tekočine zaradi dovajanja električnih pulzov niso bili simulirani posamezni električni pulzi, ampak je bila uporabljena konstantna maksimalna napetost pulzov pomnožena s ponavljalno frekvenco (f) in dolžino pulzov (τ) [12]–[14].

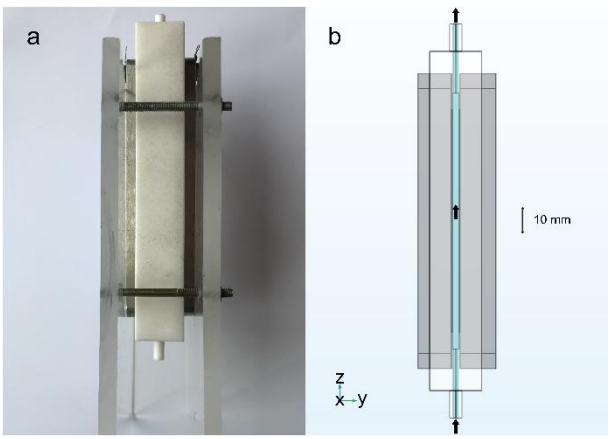
Namen tega prispevka je predstaviti uporabo sklopljene časovno odvisne simulacije v COMSOLu za analizo temperaturnega profila pretočne komore. Vsak pulz je simuliran posebej, kar omogoča vpogled v dogajanje v pretočni komori med samim električnim pulzom in tako bolj točno analizo temperaturnega profila. Rezultate simulacij smo primerjali z izmerjeno temperaturo s toplotnim senzorjem na izhodu (v iztoku) linearne pretočne komore.

2 Materiali in metode



Slika 1. Shema eksperimentalne postavitve. Zgoraj levo osciloskop, desno pulzni generator. Na sredini od leve proti desni: 0,18 % raztopina NaCl, injekcijska črpalka, pretočna komora, zbiralnik tekočine. Spodaj levo optični termometer s temperaturnimi senzorji in desno računalnik za zajem signalov.

Za poskuse je bila izbrana 0,18 % raztopina NaCl, saj je njena električna prevodnost približno enaka električni prevodnosti sadnih sokov, zlasti svežega pomarančnega soka. Koncentracijo 0,18 % raztopine NaCl smo pridobili z mešanjem fiziološke raztopine (0,9 %) in destilirane vode v razmerju 1:4. Raztopina NaCl je bila z injekcijsko črpalko preko silikonskih cevk črpana v pretočno komoro. Pravokotni unipolarni električni pulzi so bili dovedeni s pulznim generatorjem razvitem v Laboratoriju za Biokibernetiko [15]. Električni tok in napetost sta bila posneta v sekvenčnem načinu z osciloskopom z diferencialno visokonapetostno in tokovno sondo. Temperatura na izhodu pretočne komore je bila zajeta s pomočjo optičnega termometra, meritve pa nadalje obdelane z računalnikom. Temperatura 0,18 % raztopine NaCl in sobna temperatura, sta bili prav tako vseskozi beleženi.
Shemo eksperimentalne postavitve lahko vidimo na sliki 1.

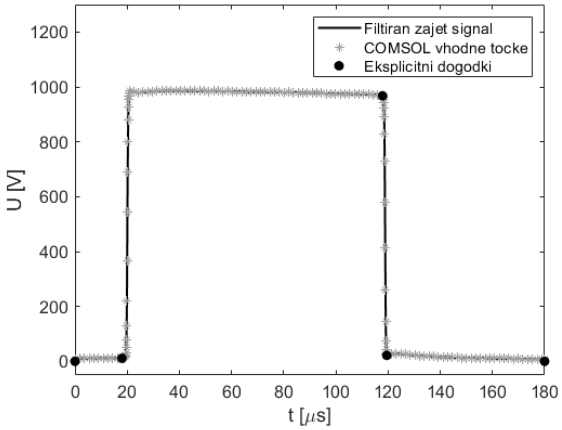


Slika 2. Fotografija linearne pretočne komore (a). Model linearne pretočne komore v COMSOLu (b).

Za poskuse smo uporabili linearno pretočno komoro (slika 2) [16], sestavljeno iz teflonskega nosilca in dveh elektrod iz nerjavnega jekla. Razdalja med elektrodami je bila 3,4 mm, površina ene elektrode 288 mm^2, prostornina komore $v = 0{,}826$ ml.

Za primerjavo meritev in simulacije smo uporabili vlak desetih pulzov z maksimalno amplitudo pulza $U = 1000\,V$, širino pulza $\tau = 100\,\mu s$ in ponavljalno frekvenco $f = 1\,Hz$ pri volumskem pretoku skozi komoro $F = 10\,ml\,min^{-1}$. Filtrirano obliko dovedenih električnih pulzov lahko vidimo na sliki 3.

2.1 Numerično modeliranje



Slika 3. Unipolarni pravokotni električni pulz uporabljen v eksperimentu in za simulacije. Na filtriranem signalu pulza so označene točke uporabljene za model pulza v COMSOLU s posebej označenimi točkami, uporabljenimi kot eksplicitni dogodki v COMSOLU (angl. »Explicit events«).

Za numerično modeliranje segrevanja tekočine znotraj pretočne komore med dovajanjem električnih pulzov smo uporabili COMSOL Multyphysics (COMSOL AB, Sweden).

Pretočno komoro smo dobro premerili in v COMSOLu postavili natančen model komore v treh dimenzijah (3D), slika 2 a. Materiale za model smo izbrali iz obsežne COMSOL baze materialov tako, da so bile njihove fizikalne lastnosti čim bolj podobne lastnostim dejanskih materialov uporabljenih v poskusih.

V primeru 0,18 % raztopine NaCl je bila temperaturna odvisnost električne prevodnosti izračunana po modelu opisanem v [17], za opis ostalih fizikalnih lastnosti pa je bil uporabljen model vode v COMSOLu.

Za modeliranje pretoka tekočine skozi linearno pretočno komoro smo zaradi nizkih povprečnih hitrosti tekočine uporabili model za laminaren tok (»Laminar Flow Interface«).

Za modeliranje električnih pulzov smo vlak pulzov uporabljen v poskusu najprej filtrirali, in nato decimirali v MATLABu (Mathworks, USA), ter točke pulza vnesli v COMSOL kot interpolacijsko funkcijo (slika 3), ki smo jo uporabili v analitični funkciji s periodičnim

ponavljanjem s periodo pulzov uporabljenih v poskusu $t_p = \frac{1}{f}$.

Za začetno temperaturo modela in začetno temperaturo tekočine na vhodu v pretočno komoro smo izbrali izmerjeno temperaturo v iztoku komore tik pred začetkom dovedenega vlaka električnih pulzov. Zaradi majhnih razlik med temperaturo segrete pretočne komore med dovajanjem električnih pulzov in temperaturo zraka v sobi (~1-2 °C) in kratkim trajanjem poskusa in simulacije (~ 10 s), je bilo v modelu odvajanje toplote iz površine pretočne komore kot posledica naravne konvekcije zanemarjeno.

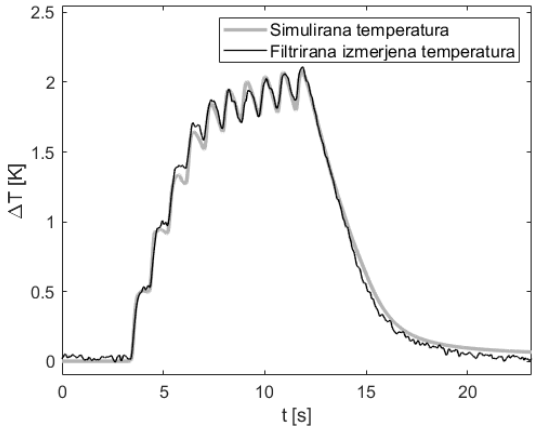
Da smo lahko s COMSOLom natančno simulirali vsak pulz, smo uporabili »Events interface«, v katerega smo vnesli eksplicitne dogodke. Eksplicitni dogodki prisilijo COMSOL, da zmanjša časovne korake okoli podane točke v času, ter tako ne spregleda hitrih sprememb napetosti. Za natančno sledenje napetosti uporabljenih električnih pulzov smo vnesli pet eksplicitnih dogodkov (slika 3) s periodo dogodka $t_p = \frac{1}{f}$.

3. Rezultati in razprava

Da bi preverili ustreznost opisanega modela linearne pretočne komore smo simulirali spremembo temperature na izhodu komore med dovajanjem unipolarnih pravokotnih električnih pulzov in jo primerjali s spremembo temperature izmerjeno med poskusom (slika 4).

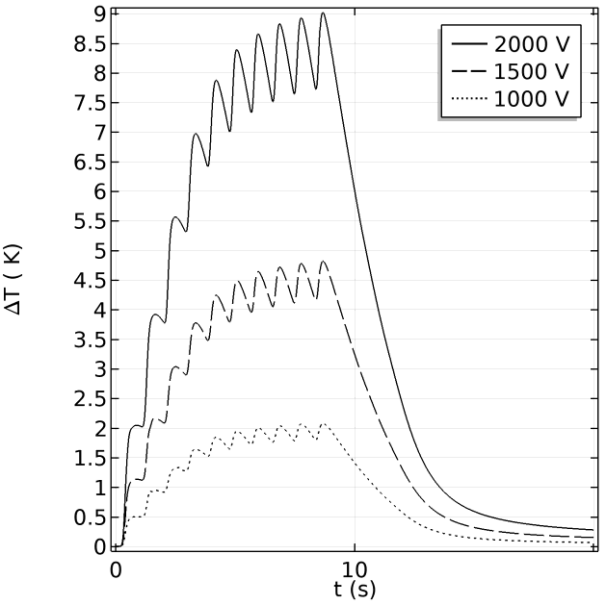
Spremembo temperature smo izračunali tako, da smo od meritve in simulacije temperature tekočine odšteli temperaturo tekočine, izmerjeno tik pred dovajanjem električnih pulzov. Temperatura v modelu je bila izrisana v točki, izbrani na podlagi ocene pozicije temperaturnega senzorja v odtočnem kanalu med samim poskusom.

Na sliki 4 lahko vidimo, da se časovna dinamika temperature določene v modelu, zelo dobro ujema s spremembo temperature izmerjene v poskusu. Prav tako lahko v poteku obeh temperatur opazimo nihanje, ki sovpada s številom in frekvenco dovedenih električnih pulzov.

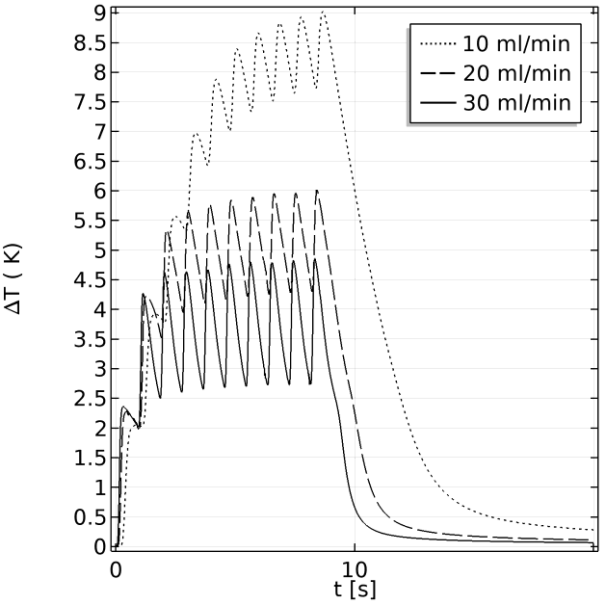


Slika 4. Primerjava spremembe simulirane temperature v iztoku linearne pretočne komore s spremembo filtrirane izmerjene temperature med dovajanjem električnih pulzov.

Z numeričnim modelom linearne pretočne komore lahko simuliramo temperaturne profile pri različnih parametrih, kot sta napetost in pretok, ki so bili zaradi omejitev eksperimentalne opreme težko izvedljivi. Na sliki 5 je prikazana sprememba temperature v iztoku komore pri različnih maksimalnih napetostih električnih pulzov in fiksnem pretoku, slika 6 pa podaja rezultate simulacij za različne pretoke pri fiksni maksimalni napetosti električnih pulzov.



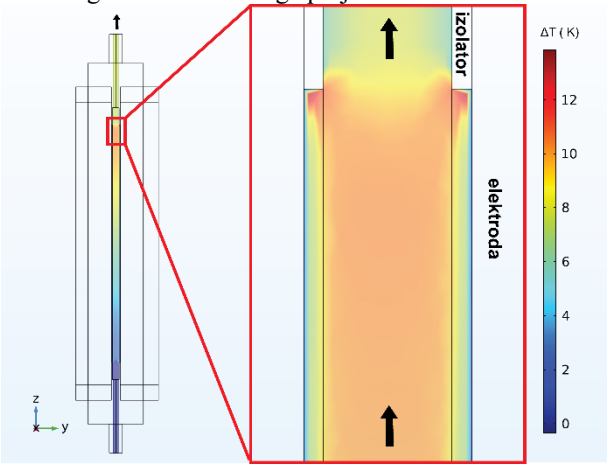
Slika 5. Simulirana temperatura v iztoku linearne pretočne komore pri različnih maksimalnih napetostih električnih pulzov pri pretoku $F = 10\ ml\ min^{-1}$.



Slika 6. Simulirana temperatura v iztoku linearne pretočne komore pri različnih pretokih tekočine, pri maksimalni napetosti električnih pulzov $U = 2000\ V$.

Napetosti na sliki 5 so bile dobljene z množenjem analitične funkcije, uporabljene za opis realnih pulzov dovedenih med poskusom s faktorjem 1,5 ($U = 1500\ V$), oziroma 2 ($U = 2000\ V$).

Glavna prednosti numerične simulacije je, da nam omogoča določiti temperaturni profil katerega koli dela pretočne komore v vsakem trenutku simulacije. Na sliki 7 lahko vidimo simulirani temperaturni profil tekočine v linearni pretočni komori takoj po koncu zadnjega izmed desetih dovedenih električnih pulzov. Na sliki lahko opazimo močno segrevanje tekočine v bližini stika med elektrodo, tekočino in izolatorjem, ki je posledica nehomogenosti električnega polja.



Slika 7. Temperaturni profil tekočine v linearni pretočni komori. Maksimalna napetost električnega pulza je $2000\,V$ in pretok $F = 10\,ml\,min^{-1}$.

Če primerjamo spremembo temperature v iztoku linearne pretočne komore (slika 5; $U = 2000\,V$) s temperaturnim profilom v celotni pretočni komori (slika 7), lahko opazimo, da je sprememba temperature tekočine na posameznih območjih v komori veliko večja od spremembe temperature v iztoku, kjer jo lahko merimo eksperimentalno.

4. Zaključek

V prispevku smo opisali sklopljene numerične simulacije v COMSOLU za namen analize temperaturnega profila linearne pretočne komore med dovajanjem električnih pulzov. Točnost modela smo ovrednotili s poskusom in dosegli dobro ujemanje med spremembo temperature v iztoku pretočne komore v simulaciji in v poskusu.

V prihodnosti bi bilo smiselno preizkusiti in ovrednotiti simulacije še za različne protokole električnih pulzov. Nadvse pomembna bi bila simulacija večjega števila pulzov, ki bi omogočala vpogled v temperaturni profil pretočne komore med kontinuiranim delovanjem, kar je za obdelavo hrane s PEF ključno. Numerični model bi bilo mogoče razširiti tudi na različne pretočne komore in uporabiti kot komplement k poskusom, ali kot orodje za optimizacijo procesa obdelave hrane.

Literatura

[1] T. Kotnik, L. Rems, M. Tarek, and D. Miklavčič, "Membrane Electroporation and Electropermeabilization: Mechanisms and Models," *Annu. Rev. Biophys.*, 48, 1, 63–91, 2019

[2] L. Rems and D. Miklavčič, "Tutorial: Electroporation of cells in complex materials and tissue," *J. Appl. Phys.*, 119, 20, 201101, 2016

[3] M. L. Yarmush, A. Golberg, G. Serša, T. Kotnik, and D. Miklavčič, "Electroporation-Based Technologies for Medicine: Principles, Applications, and Challenges," *Annu. Rev. Biomed. Eng.*, 16, 1, 295–320, 2014

[4] T. Kotnik, W. Frey, M. Sack, S. Haberl Meglič, M. Peterka, and D. Miklavčič, "Electroporation-based applications in biotechnology," *Trends Biotechnol.*, 33, 8, 480–488, 2015

[5] G. A. Ballash *et al.*, "Pulsed electric field application reduces carbapenem- and colistin-resistant microbiota and blaKPC spread in urban wastewater," *J. Environ. Manage.*, 265, 110529, 2020

[6] A. Golberg *et al.*, "Energy-efficient biomass processing with pulsed electric fields for bioeconomy and sustainable development," *Biotechnol. Biofuels*, 9, 1, 94, 2016

[7] S. Toepfl, C. Siemer, G. Saldaña-Navarro, and V. Heinz, "Overview of Pulsed Electric Fields Processing for Food," in *Emerging Technologies for Food Processing*, Elsevier, 2014, 93–114.

[8] S. Mahnič-Kalamiza, E. Vorobiev, and D. Miklavčič, "Electroporation in Food Processing and Biorefinery," *J. Membr. Biol.*, 247, 12, 1279–1304, 2014

[9] P. Mañas and R. Pagán, "Microbial inactivation by new technologies of food preservation," *J. Appl. Microbiol.*, 98, 6, 1387–1399, 2005

[10] S. Min, G. A. Evrendilek, and H. Q. Zhang, "Pulsed Electric Fields: Processing System, Microbial and Enzyme Inhibition, and Shelf Life Extension of Foods," *IEEE Trans. Plasma Sci.*, 35, 1, 59–73, 2007

[11] D. Gerlach, N. Alleborn, A. Baars, A. Delgado, J. Moritz, and D. Knorr, "Numerical simulations of pulsed electric fields for food preservation: A review," *Innov. Food Sci. Emerg. Technol.*, 9, 4, 408–417, 2008

[12] A. Fiala, P. C. Wouters, E. van den Bosch, and Y. L. M. Creyghton, "Coupled electrical-fluid model of pulsed electric field treatment in a model food system," *Innov. Food Sci. Emerg. Technol.*, 2, 4, 229–238, 2001

[13] M. Lindgren, K. Aronsson, S. Galt, and T. Ohlsson, "Simulation of the temperature increase in pulsed electric field (PEF) continuous flow treatment chambers," *Innov. Food Sci. Emerg. Technol.*, 3, 3, 233–245, 2002

[14] K. Huang, L. Yu, L. Gai, and J. Wang, "Coupled Simulations in Colinear and Coaxial Continuous Pulsed Electric Field Treatment Chambers," *Trans. ASABE*, 1473–1484, 2013

[15] K. FLISAR, „VISOKOENERGIJSKA NAPRAVA ZA REVERZIBILNO IN IREVERZIBILNO ELEKTROPORACIJO CELIČNE MEMBRANE". 2017.

[16] G. Pataro, G. Ferrentino, C. Ricciardi, and G. Ferrari, "Pulsed electric fields assisted microbial inactivation of S. cerevisiae cells by high pressure carbon dioxide," *J. Supercrit. Fluids*, 54, 1, 120–128, 2010

[17] R. B. McCleskey, "Electrical Conductivity of Electrolytes Found In Natural Waters from (5 to 90) °C," *J. Chem. Eng. Data*, 56, 2, 317–327, 2011

Electrical Impedance Spectroscopy insights into plant tissues treated by Pulsed Electric Fields

Jessica Genovese[1,2], Samo Mahnič-Kalamiza[2], Matej Kranjc[2], Pietro Rocculi[1], Damijan Miklavčič[2]

[1]*University of Bologna, Department of Agricultural and Food Sciences, P.zza Goidanich 60, 47521 Cesena, Italy*
[2]*Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržarška cesta 25, 1000 Ljubljana*
E-pošta: jessica.genovese3@unibo.it

Electrical Impedance Spectroscopy insights into plant tissues treated by Pulsed Electric Fields

Abstract. *Electroporation or pulsed electric field (PEF) treatment is known to cause an increase of cell membrane permeability and consequently an increase of the cell membrane conductivity. This is explained by the creation of aqueous pathways in the lipid domain of the cell membrane exposed to the external electric field. Since the cell membrane exhibits relatively high impedance, any permeabilization will result in a drop in impedance of the single cell and consequently of the tissue. Hence, the electroporation effect on biological matrices can be assessed by measurements of their electrical properties. The electrical impedance spectroscopy (EIS) has been suggested as a reliable method to estimate the extent of tissue damage due to high voltage treatment. This study reports on results of the bioimpedance measurements performed on different PEF-treated plant tissues (i.e. apples and potatoes). Furthermore, since bioimpedance depends on several physiological parameters, and changes in electrical properties can be masked by other processes, EIS was performed on a model system, i.e. an agarose phantom, lacking any cell structures and constituents. As expected, no changes of the measured electrical parameters were detected in the agarose samples. On the contrary, plant tissues showed a pronounced drop of the normalized impedance proportional to the electric field amplitude applied to the tissue.*

1 Introduction

The electrical impedance spectroscopy (EIS) is a powerful technique for material characterization by the investigation of its electrical properties. The majority of food matrices are composed of cells, and by the simplest electrical model of an individual cell, the cell is represented as an insulating membrane, while the intracellular and the extracellular media are modelled as ionic solutions (electrolytes). Electrolytes behave as a resistive (ohmic) load up to hundreds of MHz, while, in contrast, membranes exhibit high resistance but also considerable capacitance. Typically, passive electric properties (conductivity, σ, and permittivity, ε) are strongly dependent on the frequency [1]. Looking at one single cell or a suspension of cells between electrodes, several dispersions related to different polarization mechanisms can be identified. At low frequencies ($f = 1 - 10^4$ Hz), the α-dispersion is related to the lateral movement of ions along the insulating membrane; the β-dispersion ($f = 10^4 - 10^8$ Hz) reflects the polarization of the cell membrane; the rotation of molecules having a permanent dipole (water, proteins) cause the γ-dispersion ($f > 10^9$ Hz) [2-3]. Therefore, cells with their membranes will influence tissue impedance in a frequency range up to several MHz. Since the cell membrane acts as a capacitance, low-frequency currents cannot pass via the intracellular route, thus the magnitude of impedance will be higher across the low-frequency range. High-frequency currents are, on the contrary, free to flow through the cell membrane, i.e. the membrane is "invisible" to high-frequency fields. Typically, for most vegetable and animal tissues, the transition between low- and high-frequency behaviour arises at the spectral band starting from about 50 Hz and ending at about 10 MHz [4-5]. Being that the electrical impedance is very sensitive to the permeability of cell membranes, measurements of changes in the complex impedance have been suggested as a method to estimate the degree of tissue damage due to PEF treatment [6]. The utility of impedance measurement has already been demonstrated for various biological systems [7-9].

The aim of our study was to investigate the response of different plant tissues exposed to various amplitudes of the electric field by measuring changes in electrical properties of tissue occurring after pulse application. Additionally, bioimpedance measurements were performed on a model system, i.e. an agarose phantom, which lacks the cell structure, and therefore no changes in its impedance were expected to be seen.

2 Materials and Methods

Measurements of the electrical properties before and after electroporation were performed on different plant tissues and on an agarose phantom.

2.1 Plant tissues

Three plant tissues were used for this study:
- Apples (*Malus domestica*, cv 'Golden Deliciuos');
- Apples (*Malus domestica*, cv 'Idared');
- Potatoes (*Solanum tuberosum*, cv 'Liberta').

The raw materials were purchased at a local supermarket and, before trials, they were manually cut

using a sharp cork-borer to obtain cylindrical-shape samples of 26 mm in diameter and 6 mm in height.

2.2 Agarose phantom

The agarose phantom was made of agarose powder (Thermo Scientific, USA) and distilled deionized water (Braun, Germany). The electrical conductivity of the agarose mixture was adjusted (with the addition of NaCl saline solution) to the initial electrical conductivity of the vegetable tissues used in this study.

The agarose samples were cut into cylindrical-shape of 26 mm in diameter and 6 mm in height.

2.3 Electroporation protocol

Cylindrical samples were placed in a treatment chamber consisted of two stainless-steel parallel plate electrodes spaced at 6 mm from each other, and used to apply 8 rectangular monopolar pulses of 100 µs duration to the tissue with a pulse repetition rate of 1 Hz. For each treatment, different pulse amplitudes were used (from 50 V to 1500 V), resulting in different electric field strengths (V/cm, calculated as the ratio between the applied voltage and the electrodes distance). The delivered voltage and current in the chamber were measured and recorded by a high-voltage probe (mod. HVD3206A, LeCroy USA) and a current probe (mod. CP031A, LeCroy USA) connected to a sequencing digital storage oscilloscope (mod. HDO6104A-MS, LeCroy USA).

2.4 Impedance measurements

The bioimpedance measurement system was assembled by connecting the electroporation chamber that holds the sample and the electrodes to a precision LCR meter (mod. E4980A, Keysight Inc. USA). The same electrodes that were employed to deliver the high-voltage electroporation pulses were also used for the impedance measurements. Switching relays were adopted to alternate the electrical connections of the electrodes between the generator and the LCR meter. Multi-frequency parallel capacitance (C_p) and parallel resistance (R_p) measurements of untreated and treated samples (immediately, i.e. between 3-5 seconds, after the pulse application) were performed in the frequency range of 50 Hz to 1 MHz by applying a 100 mV peak potential difference to the electrodes. Data were acquired using an in-house developed software for the LCR meter control and data capture (based on the Arduino and National Instruments LabVIEW software platforms). The magnitude of impedance was calculated according to equation 1.

$$|Z| = \sqrt{R^2 + X^2} \quad (1)$$

where R is the real component (resistance) and X the imaginary component (reactance) of impedance.

3 Results

For various applied voltage magnitudes and different materials analysed, the ratio between the absolute impedance value after the electroporation ($|Z_a|$) and before the electric treatment ($|Z_b|$) is reported.

As expected, no changes in the magnitude of impedance were observed when agarose phantoms were exposed to pulses of 500 V, 1500 V, or 2000 V amplitude (Fig.1).

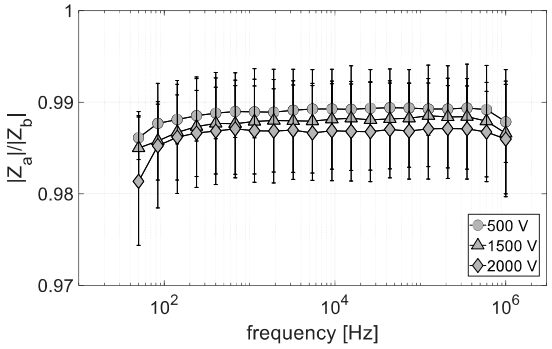


Figure 1. Normalised absolute impedance versus frequency at different applied voltage magnitudes for agarose phantom. Results are expressed as means ± standard deviations (error bars) of n = 4

Conversely, as a result of β-dispersion, the frequency dependence of the impedance can be observed in plant tissues (Fig. 2,3,4 A). Considering that low-frequency currents cannot penetrate the intact cell membrane and are restricted to pass via the extracellular space; when electroporation occurs, it will cause an increase of cell membrane permeability, resulting in a drop in the magnitude of electrical impedance (i.e. low-frequency currents will be able to pass through the intracellular space). Hence, for a better visualization of the dependency of the impedance drop on the magnitude of electric field applied, the normalized absolute impedance at sampling frequency of 5 kHz has been selected (Fig. 2,3,4 B). The indicated frequency has been selected considering that two-electrode impedance measurements suffer from electrode polarization effects, thus a high enough frequency has to be used for comparison.

For both apple cultivars (Golden Delicious and Idared), the curves show a pronounced drop in the normalized impedance when 500 V/cm voltage-to-distance ratio pulses were applied (Fig. 2,3). Dissimilarly, in the potato tissue, a consistent drop of the normalized impedance from ~ 1 (i.e. no differences in the impedance measured after electroporation compared to the value before electroporation) to 0.53 ± 0.02 has been attained after applying 250 V/cm (Fig.4). Electric fields above 500 V/cm did not result in a further decrease in impedance, suggesting the achievement of maximum cell membrane permeabilization using this particular experimental protocol (i.e. number of pulses and pulse duration) (Fig.4 B).

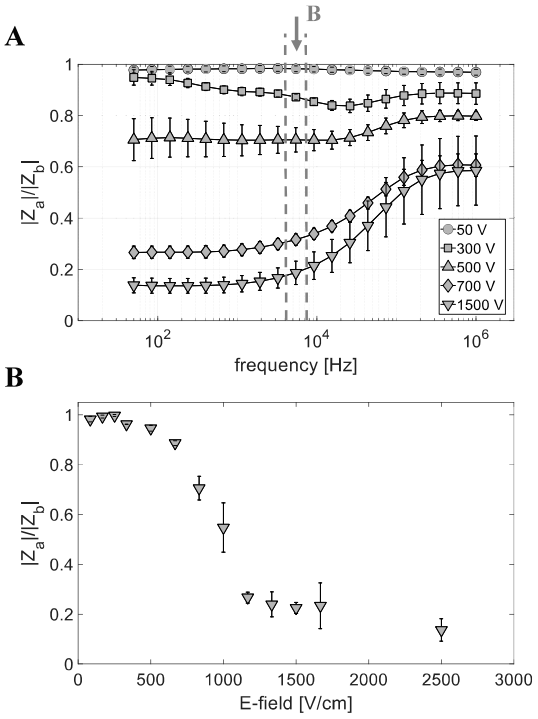


Figure 2. (A) Normalized absolute impedance versus frequency at different applied voltage magnitudes for apple cv. Golden Delicious (not all the data are shown for clarity of presentation). (B) Normalized absolute impedance at sampling frequency of 5 kHz versus the electric field magnitude. Results are expressed as means ± standard deviations (error bars) of n = 4

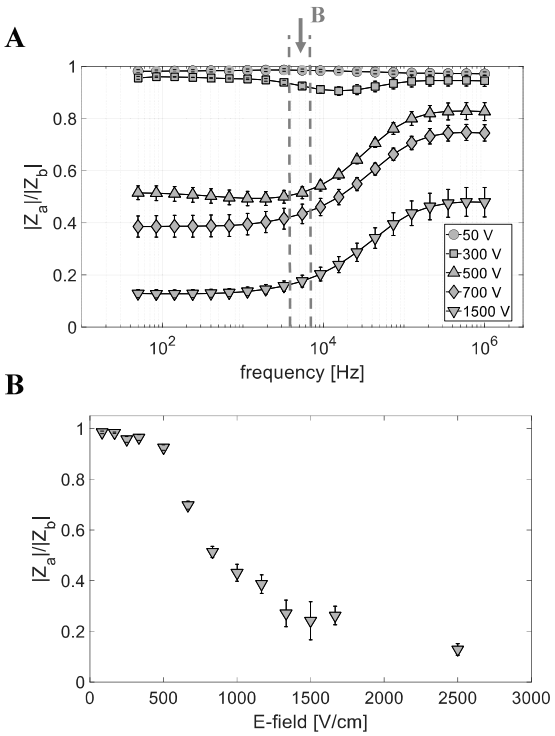


Figure 3. (A) Normalized absolute impedance versus frequency at different applied voltage magnitudes for apple cv. Idared (not all the data are shown for clarity of presentation). (B) Normalized absolute impedance at sampling frequency of 5 kHz versus the electric field magnitude. Results are expressed as means ± standard deviations (error bars) of n = 4

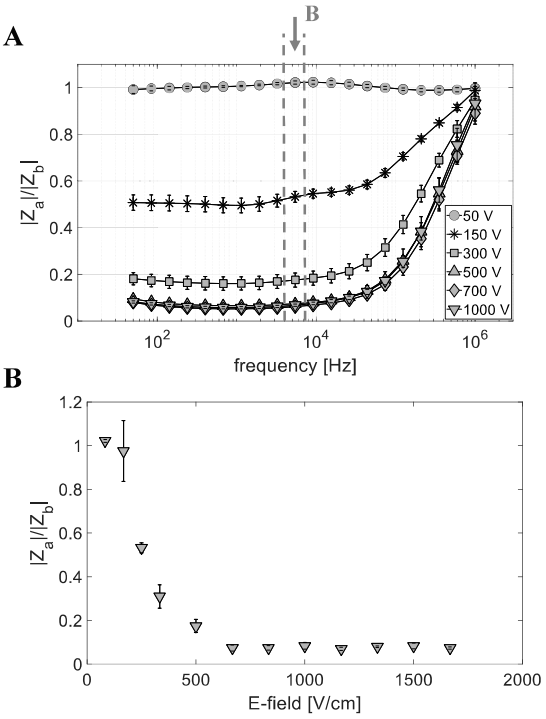


Figure 4. (A) Normalized absolute impedance versus frequency at different applied voltage magnitudes for potato (not all the data are shown for clarity of presentation). (B) Normalized absolute impedance at sampling frequency of 5 kHz versus the electric field magnitude. Results are expressed as means ± standard deviations (error bars) of n = 4

4 Discussions

The aim of this study was to assess the response of plant tissues subjected to pulsed electric field treatments of various pulse amplitudes by measuring the tissue electrical properties. The bioimpedance analysis of PEF-treated plant tissues has shown to be strongly related with the magnitude of the applied voltage.

In general, it can be observed that all vegetable samples exhibit a consistent decrease in their impedance following PEF treatment and the trend is non-linearly related to the magnitude of the electric field applied, suggesting the thresholding nature of the phenomenon. The decrease in impedance can be attributed to the permeabilization of the cell membrane, leading to an increase in the membrane conductivity. On the contrary and as expected, no matter the applied electric field amplitude, in agarose phantom the pulses did not cause any decrease in electrical impedance. The agarose was chosen deliberately as a control/model system, as it is not affected by electroporation due to lack of presence of any of the biological cell structures. Therefore, the changes detected in the select biological tissues by bioimpedance measurements must be related to changes that have occurred due to the application of pulses of certain electric field amplitude. In fact, as it can be discerned from Fig. 2,3,4 (B), a dramatic decrease of the normalized impedance is observed above a certain electric field strength threshold. The latter has shown different values in relation to the different plant tissues analysed in this study (E ~ 500 V/cm for apples; E ~

250 V/cm for potatoes). This could be explained by the different characteristics of the biological material that could affect the electrical parameters, such as the size of cells, the packing density, the spatial distribution of air, the overall tissue moisture content, and initial cell turgor pressure. For this reason, further studies should focus on the relation between the changes of the electrical parameters and changes of other physical characteristics (e.g. mechanical properties) of the biological material subject to PEF treatment.

Acknowledgments

The work was funded through research programmes and projects funded by the Slovenia Research Agency (ARRS), namely research programme P2-0249 and postdoctoral project Z7-1886. The work was performed within the scope of the Infrastructural Centre "Cellular Electrical Engineering" (I0-0022).

References

[1] S. Grimnes, Ø. G. Martinsen, S. Grimnes, and Ø. G. Martinsen, "Chapter 3 – Dielectrics," in *Bioimpedance and Bioelectricity Basics*, 2015.

[2] N. Lebovka and E. Vorobiev, "Techniques to detect electroporation in food tissues," in *Handbook of Electroporation*, 2017.

[3] D. Miklavčič, N. Pavšelj, and F. X. Hart, "Electric Properties of Tissues," in *Wiley Encyclopedia of Biomedical Engineering*, 2006.

[4] Q. Castellví, B. Mercadal, and A. Ivorra, "Assessment of electroporation by electrical impedance methods," in *Handbook of Electroporation*, 2017.

[5] F. Donsì, G. Ferrari, and G. Pataro, "Applications of pulsed electric field treatments for the enhancement of mass transfer from vegetable tissue," *Food Eng. Rev.*, vol. 2, no. 2, pp. 109–130, 2010.

[6] A. Angersbach, V. Heinz, and D. Knorr, "Evaluation of process-induced dimensional changes in the membrane structure of biological cells using impedance measurement," *Biotechnol. Prog.*, 2002.

[7] A. Angersbach, V. Heinz, and D. Knorr, "Electrophysiological model of intact and processed plant tissues: Cell disintegration criteria," *Biotechnol. Prog.*, 1999.

[8] D. Cukjati, D. Batiuskaite, F. André, D. Miklavčič, and L. M. Mir, "Real time electroporation control for accurate and safe in vivo non-viral gene therapy," *Bioelectrochemistry*, 2007.

[9] S. Mahnič-Kalamiza, D. Miklavčič, and E. Vorobiev, "Dual-porosity model of mass transport in electroporated biological tissue: Simulations and experimental work for model validation," *Innov. Food Sci. Emerg. Technol.*, 2015.

[10] M. Kranjc, F. Bajd, I. Serša, and D. Miklavčič, "Magnetic resonance electrical impedance tomography for measuring electrical conductivity during electroporation," *Physiol. Meas.*, 2014.

Numerično modeliranje porazdelitve pH sprememb v tkivu med dovajanjem elektroporacijskih pulzov

Rok Šmerc, Samo Mahnič-Kalamiza, Damijan Miklavčič

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: rok.smerc@fe.uni-lj.si

Numerical modelling of pH changes distribution in tissue during delivery of electroporation pulses

Abstract. *During and after delivery of electroporation pulses, tissue damage, caused by various electric current and field effects, may occur. For optimal gene electrotransfer, tissue damage needs to be minimised. One of the possible causes for tissue damage is altered pH, caused by electrolysis of water, i.e. hydrolysis. Numerical modelling can be used to assess these pH changes in tissue. In the present study we built a two-dimensional numerical model, based on Nernst-Planck system of equations for the concentration of ions in a four-component electrolyte, and used it to calculate the distribution of pH changes in tissue during delivery of gene electrotransfer pulses (applied voltage of* $100\,\mathrm{V}$ *and total pulse time of* $60\,\mathrm{ms}$*). Two mechanisms were taken into account: diffusion, caused by concentration gradients; and migration, caused by potential gradients. COMSOL Multiphysics software was used for calculations of pH changes distribution in tissue with finite element method. Our results show that during delivery of pulses, an acidic pH front forms in the vicinity of the anode, and an alkaline pH front forms in the vicinity of the cathode. Extreme pH values below* 2.5 *and above* 13 *are reached and sharp delineations between the normal pH and altered pH are observed. The model at this stage has difficulties with convergence, which are caused by relatively high voltage used in electroporation, very high concentration gradients, and the fact that five, highly non-linear, coupled partial differential equations need to be solved.*

1 Uvod

Elektroporacija oz. elektropermabilizacija je pojav, kjer z izpostavitvijo celice električnemu polju dovoljšnje jakosti povzročimo spremembe v njeni membrani, zaradi česar le-ta začasno postane bolj prepustna za molekule, za katere je sicer slabo ali neprepustna [1]. Elektroporacija se uporablja na različnih področjih, od biomedicine [2] do biotehnologije [3] in predelave hrane [4].

V biomedicini je ena izmed najpogosteje uporabljanih metod, temelječih na elektroporaciji, genska elektrotransfekcija. Pri genski elektrotransfekciji olajšamo prehajanje genskega materiala (običajno plazmidne DNK) v celice z uporabo metode reverzibilne elektroporacije [5]. Za dosego čim boljše genske izražave je potrebno minimizirati poškodbe tkiva, ki jih povzročijo različni učinki uporabe elektroporacije in dovajanja električnih pulzov. Eden izmed teh so spremembe pH v tkivu, povzročene z elektrolizo vode, tj. hidrolizo [6]. Te spremembe lahko ocenimo z uporabo numeričnega modeliranja. Pri tem uporabimo Nernst-Planckov sistem enačb, ki popisuje gibanje nabitih ionov v mediju, ki jih rešimo z izbrano numerično metodo, najpogosteje z metodo končnih razlik ali z metodo končnih elementov. Težava pri tem je, da je potrebno rešiti več, močno nelinearnih, sklopljenih parcialnih diferencialnih enačb – eno, ki popisuje prostorsko porazdelitev električnega polja ter dodatno enačbo za vsako vrsto ionov, ki jo v modelu upoštevamo. Objavljene študije s tega področja se osredotočajo bodisi na elektrokemijsko zdravljenje tumorjev, kjer se uporabljajo znatno nižje vrednosti električnih tokov in napetosti [7-11] bodisi na pojave med elektroporacijo v le enodimenzionalnem prostoru [6, 12-14]. Študija, kjer bi uspešno implementirali model v več dimenzijah z elektroporacijskimi pogoji, kolikor nam je znano, še ni objavljena.

V pričujoči raziskavi predstavljamo prvi korak v izdelavi numeričnega modela, ki bi omogočal izračun sprememb pH v tkivu pri vrednostih električne napetosti in toka, značilnih za elektroporacijo. Model smo izdelali v programskem okolju COMSOL Multiphysics, ki omogoča numerično reševanje diferencialnih enačb z metodo končnih elementov.

2 Materiali in metode

2.1 Nernst-Planckov sistem enačb

Enačbe, ki smo jih uporabili za numerično modeliranje porazdelitve pH sprememb v tkivu med dovajanjem elektroporacijskih pulzov temeljijo na delu Eve Nilsson s sodelavci [15, 16] s področja numeričnega modeliranja elektrokemijskega zdravljenja tumorjev. Model temelji na Nernst-Planckovem sistemu enačb za koncentracijo ionov v štirikomponentnem elektrolitu. Upoštevamo dva mehanizma: difuzijo vsled koncentracijskih gradientov in migracijo zaradi potencialnih gradientov. Za konvekcijo predpostavimo, da je zanemarljiva in je torej ne upoštevamo. Model vsebuje pet neznanih spremenljivk: koncentracije vodikovih, hidroksidnih, kloridnih in natrijevih ionov ter prostrosko porazdelitev električnega polja.

Materialno ravnovesje ionov komponente i je podano z:

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot \mathbf{N}_i + R_i, \tag{1}$$

kjer je c_i koncentracija ionov komponente i (kjer je i H^+, OH^-, Cl^- ali Na^+ ion), t je čas, R_i je sproščanje ionov komponente i s kemijskimi reakcijami v elektrolitu, $\mathbf{N}_i$ pa je molarni tok ionov komponente i:

$$\mathbf{N}_i = -D_i \nabla c_i - \frac{z_i}{|z_i|} u_i c_i \nabla \phi, \tag{2}$$

kjer je D_i difuzijski koeficient ionov komponente i, z_i je (predznačen) naboj ionov komponente i, u_i je gibljivost ionov komponente i, ϕ pa je električni potencial v elektrolitu. Pogoju elektronevtralnosti je zadoščeno z naslednjo enačbo:

$$\sum_{i=1}^{4} z_i c_i = 0. \tag{3}$$

Gibljivost ionov in difuzijski koeficient sta povezana z Nernst-Einsteinovo enačbo:

$$D_i = \frac{RT}{|z_i|F} u_i, \tag{4}$$

kjer je R splošna plinska konstanta, T je temperatura in F je Faradayeva konstanta. Gostoto električnega toka $\mathbf{j}$ v elektrolitu izračunamo z uporabo Faradayevega zakona:

$$\mathbf{j} = F \sum_{i=1}^{4} z_i \mathbf{N}_i. \tag{5}$$

Enačbo za ohranitev električnega naboja

$$\nabla \mathbf{j} = 0 \tag{6}$$

lahko, upoštevaje enačbi (2) in (5), zapišemo kot:

$$\nabla \left(\sum_{i=1}^{4} |z_i| u_i c_i \nabla \phi \right) + \sum_{i=1}^{4} z_i D_i \nabla^2 c_i = 0, \tag{7}$$

kar je enačba, uporabljena za izračun porazdelitve električnega polja.

Dve reakciji potekata na anodi, evolucija kisika

$$2\,H_2O \rightleftharpoons O_2 + 4\,H^+ + 4\,e^- \tag{8}$$

in evolucija klorida

$$2\,Cl^- \rightleftharpoons Cl_2 + 2\,e^- \tag{9}$$

ter ena na katodi, evolucija hidroksida

$$2\,H_2O + 2\,e^- \rightleftharpoons H_2 + 2\,OH^-. \tag{10}$$

Kinetiko reakcij na elektrodah računamo z uporabo koncentracijsko osnovane Butler-Volmerjeve enačbe. Edina reakcija v elektrolitu, ki jo upoštevamo, je protolitska reakcija vode:

$$H^+ + OH^- \rightleftharpoons H_2O. \tag{11}$$

Sproščanje natrijevih in kloridnih ionov je enako 0 (torej njuna koncentracija ostane konstantna), vodikovi in hidroksidni ioni pa se sproščajo in porabljajo skladno s protolitsko reakcijo vode. Tako lahko zapišemo naslednja izraza za sproščanje ionov:

$$R_{Na^+} = R_{Cl^-} = 0, \tag{12}$$

$$R_{H^+} = R_{OH^-} = k_{w,b} c_{H_2O} - k_{w,f} c_{H^+} c_{OH^-}, \tag{13}$$

kjer sta $k_{w,b}$ in $k_{w,f}$ hitrostni konstanti protolitske reakcije vode v smeri nazaj in naprej.

Vrednosti parametrov, ki smo jih uporabili v numeričnih izračunih, so navedene v tabeli 1.

Tabela 1: Vrednosti parametrov, uporabljene v numeričnih izračunih [15].

Parameter	Vrednost
$k_{w,f}$	$1{,}5 \cdot 10^{11}\ \mathrm{l\,mol^{-1}}$
$k_{w,b}$	$2{,}7 \cdot 10^{-5}\ \mathrm{s^{-1}}$
D_{H^+}	$9{,}31 \cdot 10^{-5}\ \mathrm{cm^2\,s^{-1}}$
D_{OH^-}	$5{,}26 \cdot 10^{-5}\ \mathrm{cm^2\,s^{-1}}$
D_{Na^+}	$1{,}33 \cdot 10^{-5}\ \mathrm{cm^2\,s^{-1}}$
D_{Cl^-}	$2{,}03 \cdot 10^{-5}\ \mathrm{cm^2\,s^{-1}}$
$c^0_{H^+}$	$1{,}0 \cdot 10^{-7}\ \mathrm{mol\,l^{-1}}$
$c^0_{OH^-}$	$1{,}0 \cdot 10^{-7}\ \mathrm{mol\,l^{-1}}$
$c^0_{Na^+}$	$0{,}16 \cdot 10^{-7}\ \mathrm{mol\,l^{-1}}$
$c^0_{Cl^-}$	$0{,}16 \cdot 10^{-7}\ \mathrm{mol\,l^{-1}}$
c_{H_2O}	$55{,}5 \cdot 10^{-7}\ \mathrm{mol\,l^{-1}}$

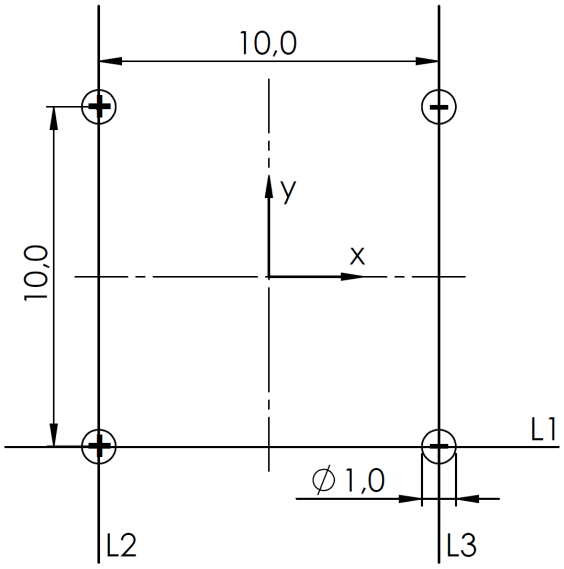
2.2 Numerično modeliranje

Z opisanimi enačbami smo izvedli numerični izračun pH sprememb v tkivu med dovajanjem elektroporacijskih pulzov, kakršni se uporabljajo pri genski elektrotransfekciji [5]. Uporabili smo štiri igelne elektrode premera $1\,\mathrm{mm}$, ki so razporejene v ogliščih kvadrata s stranico $10\,\mathrm{mm}$. Postavitev elektrod je prikazana na sliki 1.

Predpostavimo, da je glavni mehanizem, ki povzroči spremembe pH v tkivu migracija ionov v času apliciranega električnega polja, difuzija, ki poteka tekom pavz med pulzi, pa je v primerjavi s tem zanemarljiva. Ta predpostavka nam omogoči, da izračune izvedemo z uporabo konstantne napetosti, torej ne upoštevamo dejanske oblike pulzov in pavz med njimi. S tem se v modelu znebimo enega izmed izvorov nelinearnosti, s čimer postane model numerično nekoliko lažje obvladljiv. Uporabili smo napetost $100\,\mathrm{V}$ in pulze skupne dolžine $60\,\mathrm{ms}$. Porazdelitev pH vrednosti v tkivu izračunamo na podlagi porazdelitve vodikovih ionov:

$$pH = -\log_{10} c_{H^+}. \tag{14}$$

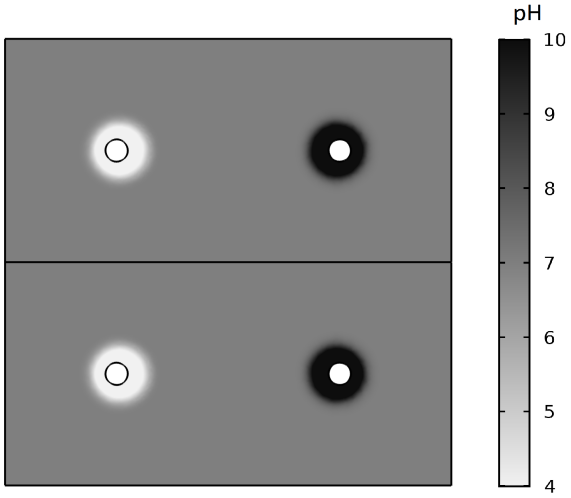
Enačbe, navedene v prejšnjem in v tem podpoglavju, smo uporabili za izdelavo numeričnega modela v dveh dimenzijah v programskem okolju COMSOL Multiphysics (COMSOL Inc., Švedska, različica 5.5), ki omogoča numerično analizo na osnovi metode končnih elementov (MKE).



Slika 1: Postavitev elektrod, vstavljenih v tkivo, z označenimi črtami L1, L2 in L3, ki smo jih uporabili za natančnejši prikaz porazdelitve pH sprememb v tkivu. Vse označene dimenzije so v milimetrih.

3 Rezultati

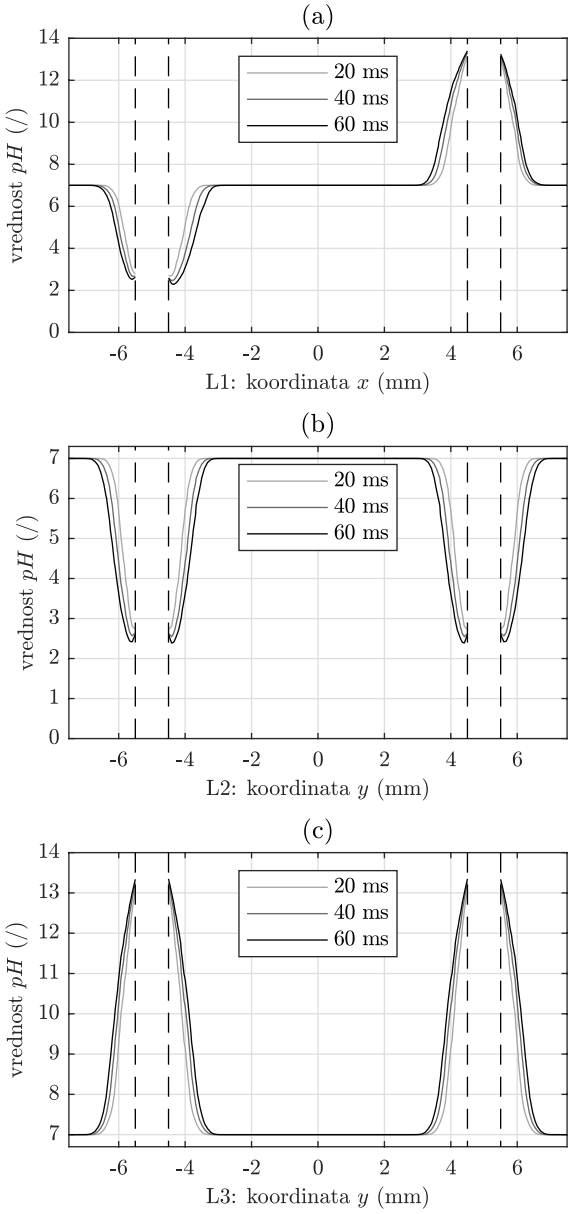
Izračunano porazdelitev pH sprememb v tkivu po dovajanju elektroporacijskih pulzov napetosti $100\,\mathrm{V}$ in skupne dolžine $60\,\mathrm{ms}$ prikazuje slika 2.



Slika 2: Porazdelitev pH sprememb v tkivu po dovajanju elektroporacijskih pulzov napetosti $100\,\mathrm{V}$ in skupne dolžine $60\,\mathrm{ms}$.

Za natančnejši prikaz porazdelitve pH sprememb v tkivu smo si izbrali tri črte, ki so označene na postavitvi elektrod, prikazani na sliki 1. Črta L1 poteka skozi eno izmed anod in eno izmed katod, črta L2 poteka skozi obe anodi in črta L3 poteka skozi obe katodi. Porazdelitev pH sprememb v tkivu po dovajanju elektroporacijskih pulzov napetosti $100\,\mathrm{V}$ in skupne dolžine 20, 40 ali $60\,\mathrm{ms}$ vzdolž teh črt je prikazana na sliki 3.

V tabeli 2 so navedene izračunane površine območij



Slika 3: Porazdelitev pH sprememb v tkivu po dovajanju elektroporacijskih pulzov napetosti $100\,\mathrm{V}$ in skupne dolžine 20, 40 ali $60\,\mathrm{ms}$ (a) vzdolž črte L1, ki poteka skozi eno anodo in eno katodo, (b) vzdolž črte L2, ki poteka skozi obe anodi in (c) vzdolž črte L3, ki poteka skozi obe katodi. Črtkane črte označujejo lokacije elektrod.

okoli vsake izmed elektrod, kjer je prišlo do spremembe vrednosti pH (vrednost pH $< 6{,}9$ oz. $> 7{,}1$) po dovajanju elektroporacijskih pulzov napetosti $100\,\mathrm{V}$ in skupne dolžine 20, 40 ali $60\,\mathrm{ms}$. Kot meje območij, kjer pride do sprememb vrednosti pH, smo izbrali vrednosti 6,9 oz. 7,1 zato, da v območje ne zajamemo vrednosti, ki se od 7,0 razlikujejo zaradi numerične napake. Marino in sodelavci navajajo, da poškodbe plazmidne DNK nastopijo, kadar je le ta izpostavljena kislemu mediju s pH nižjim od $4{,}5$, ali bazičnemu mediju s pH višjim od $8{,}4$. Izračunane vrednosti območij okoli vsake izmed elektrod, ki ustrezajo temu pogoju, so prav tako navedene v tabeli 2.

Tabela 2: Izračunane površine območij, kjer je prišlo do spremembe vrednosti pH (vrednost pH $<$ 6,9 oz. $>$ 7,1) in območij, kjer pričakujemo poškodbe plazmidne DNK (vrednost pH $<$ 4,5 oz. $>$ 8,4) okoli vsake izmed elektrod (tj. ene izmed štirih elektrod, anode oz. katode) po dovajanju elektroporacijskih pulzov napetosti $100\,\mathrm{V}$ in skupne dolžine 20, 40 ali $60\,\mathrm{ms}$. Kot meje območij, kjer pride do sprememb vrednosti pH, smo izbrali vrednosti 6,9 oz. 7,1 zato, da v območje ne zajamemo vrednosti, ki se od 7,0 razlikujejo zaradi numerične napake.

$t\,(\mathrm{ms})$	20	40	60
$pH < 4{,}5\,(\mathrm{mm}^2)$	1,583	2,528	3,437
$pH < 6{,}9\,(\mathrm{mm}^2)$	4,973	6,777	8,301
$pH > 7{,}1\,(\mathrm{mm}^2)$	6,013	7,353	8,384
$pH > 8{,}4\,(\mathrm{mm}^2)$	2,870	3,872	4,693

4 Razprava

Na grafih, ki jih prikazuje slika 3 opazimo, da se med dovajanjem elektroporacijskih pulzov v bližini anod pojavi kislo območje, v bližini katod pa bazično. V bližini anod se pojavi skrajna vrednost pH, nižja od $2{,}5$, v bližini katod pa višja od 13. Opazimo ostre razmejitve med nespremenjenimi in spremenjenimi vrednostmi pH, tj. področja pH front. Kot je pričakovano, se z daljšanjem skupnega časa dovedenih pulzov povečuje območje spremenjenih vrednosti pH, prav tako pa se znižujejo oz. zvišujejo skrajne dosežene vrednosti v bližini elektrod. Opazimo, da je v bližini katod površina območja spremenjenega pH večja, kot v bližini anod, prav tako pa je v bližini katod večja površina območja, kjer pričakujemo poškodbe plazmidne DNK.

V tem prispevku smo predstavili prvi korak pri izdelavi numeričnega modela, ki omogoča izračun porazdelitve pH sprememb v tkivu med dovajanjem elektroporacijskih pulzov. Model smo uporabili za izračun porazdelitve pH sprememb v tkivu med dovajanjem elektroporacijskih pulzov, ki se uporabljajo pri genski transfekciji. Numerični model ima na tej stopnji razvoja težave s konvergenco in potrebuje nadaljnji razvoj. Težave povzročajo relativno visoke napetosti, ki se uporabljajo v elektroporaciji, visoki koncentracijski gradienti in pa dejstvo, da rešujemo sistem petih močno nelinearnih sklopljenih parcialnih diferencialnih enačb. Prav tako v tej fazi še niso upoštevane naravne sposobnosti tkiva za nevtralizacijo povzročenih pH sprememb, rezultati izračunov pa niso bili validirani z rezultati realnih eksperimentov.

Zahvala

Študijo sta delno financirali Javna agencija za raziskovalno dejavnost Republike Slovenije in podjetje RenBio.

Literatura

[1] T. Kotnik, L. Rems, M. Tarek, in D. Miklavčič, „Membrane Electroporation and Electropermeabilization: Mechanisms and Models“, Annu. Rev. Biophys., let. 48, št. 1, str. 63–91, maj 2019.

[2] M. L. Yarmush, A. Golberg, G. Serša, T. Kotnik, in D. Miklavčič, „Electroporation-Based Technologies for Medicine: Principles, Applications, and Challenges“, Annu. Rev. Biomed. Eng., let. 16, št. 1, str. 295–320, jul. 2014.

[3] T. Kotnik, W. Frey, M. Sack, S. Haberl Meglič, M. Peterka, in D. Miklavčič, „Electroporation-based applications in biotechnology“, Trends in Biotechnology, let. 33, št. 8, str. 480–488, avg. 2015.

[4] S. Mahnič-Kalamiza, E. Vorobiev, in D. Miklavčič, „Electroporation in Food Processing and Biorefinery“, J Membrane Biol, let. 247, št. 12, str. 1279–1304, dec. 2014.

[5] C. Rosazza, S. Haberl Meglič, A. Zumbusch, M.-P. Rols, in D. Miklavčič, „Gene Electrotransfer: A Mechanistic Perspective“, CGT, let. 16, št. 2, str. 98–129, apr. 2016.

[6] N. Olaiz idr., „Tissue damage modeling in gene electrotransfer: The role of pH“, Bioelectrochemistry, let. 100, str. 105–111, dec. 2014.

[7] P. Turjanski, N. Olaiz, P. Abou-Adal, C. Suárez, M. Risk, in G. Marshall, „pH front tracking in the electrochemical treatment (EChT) of tumors: Experiments and simulations“, Electrochimica Acta, let. 54, št. 26, str. 6199–6206, nov. 2009.

[8] E. Luján idr., „Optimal dose-response relationship in electrolytic ablation of tumors with a one-probe-two-electrode device“, Electrochimica Acta, let. 186, str. 494–503, dec. 2015.

[9] A. Soba idr., „Integrated analysis of the potential, electric field, temperature, pH and tissue damage generated by different electrode arrays in a tumor under electrochemical treatment“, Mathematics and Computers in Simulation, let. 146, str. 160–176, apr. 2018.

[10] A. Mokhtare, M. Shiv Krishna Reddy, V. A. Roodan, E. P. Furlani, in A. Abbaspourrad, „The role of pH fronts, chlorination and physicochemical reactions in tumor necrosis in the electrochemical treatment of tumors: A numerical study“, Electrochimica Acta, let. 307, str. 129–147, jun. 2019.

[11] E. M. Calzado, H. Schinca, L. E. B. Cabrales, F. M. García, P. Turjanski, in N. Olaiz, „Impact of permeabilization and pH effects in the electrochemical treatment of tumors: Experiments and simulations“, Applied Mathematical Modelling, let. 74, str. 62–72, okt. 2019.

[12] P. Turjanski idr., „The Role of pH Fronts in Reversible Electroporation“, PLoS ONE, let. 6, št. 4, str. e17303, apr. 2011.

[13] F. Maglietti, S. Michinski, N. Olaiz, M. Castro, C. Suárez, in G. Marshall, „The Role of Ph Fronts in Tissue Electroporation Based Treatments“, PLoS ONE, let. 8, št. 11, str. e80167, nov. 2013.

[14] M. Marino idr., „pH fronts and tissue natural buffer interaction in gene electrotransfer protocols“, Electrochimica Acta, let. 255, str. 463–471, nov. 2017.

[15] E. Nilsson, J. Berendson, in E. Fontes, „Electrochemical treatment of tumours: a simplified mathematical model“, Journal of Electroanalytical Chemistry, let. 460, št. 1–2, str. 88–99, jan. 1999.

[16] E. Nilsson in E. Fontes, „Mathematical modelling of physicochemical reactions and transport processes occurring around a platinum cathode during the electrochemical treatment of tumours“, Bioelectrochemistry, let. 53, št. 2, str. 213–224, mar. 2001.

Didaktika

Didactics

ICT engineering study programs to meet modern society needs: Erasmus+ project BENEFIT

Matej Zajc[1], Aljo Mujčić[2], Andrea Tonello[3], Vlado Delić[4], Mladen Koprivica[5], Snježana Rimac Drlje[6], Urban Burnik[1], Nermin Suljanović[2]

[1]University of Ljubljana, FE, Tržaška 25, SI-1000 Ljubljana, Slovenia
[2] University of Tuzla, FET, Franjevačka 2, 75000 Tuzla, B&H
[3] University of Klagenfurt, NES, Lake Side Park B02 – A-9020 Klagenfurt, Austria
[4] University of Novi Sad, FTN, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
[5] University of Belgrade, ETF, Bulevar kralja Aleksandra 73, Beograd, Serbia
[6] University of Osijek, FERIT, Kneza Trpimira 2b, 31000, Osijek, Croatia

E-pošta: matej.zajc@fe.uni-lj.si

Abstract **-** In this paper, we briefly present Erasmus+ KA2 Capacity Building project BENEFIT, targeting ICT engineering study programs in Western Balkan region. Project partners work on improving ICT engineering education at participating HEIs, while following international guidelines, industry needs and student expectations. Paper outlines project goals and highlights some significant efforts and achievements.

1 Introduction

Engineering education is consistently evolving to meet the demands of society, with a focus on industry needs [1, 2]. There are many challenges and constraints to global progress of engineering education, among others related to interdisciplinarity of programs, implementation of efficient approaches to active learning, and focus on teaching achievements. Constantly changing conditions in the job market require that Higher Education Institutions (HEIs) continuously improve engineering education.

For the last decade, the BENEFIT project partners identified lowered interest in telecommunication studies in the south-eastern Europe – Western Balkan region. This trend is at least partially a consequence of telecommunication industry passing through significant change [3].

Furthermore, novel technologies such as Internet of things and cloud technologies are demanding skills and expertise reaching beyond the traditional telecommunication engineer profile towards the information and communication technologies (ICT) expert trained to understand and respond to new information-centric era.

While ICT is a pillar of modern society, the industry has difficulties to find sufficiently trained professionals [1]. Industry is finding alternative ways to rising demand in the ICT sector, while universities strive to modernize teaching methodologies and infrastructures. To properly address these circumstances, telecommunications engineering studies require evolution towards more ICT oriented studies in the Western Balkan region and better alignment with the regional industry needs. A comprehensive approach requires that all stakeholders are involved, namely teachers, industry, public authorities as well as students, whose expectations need to be addressed [1-3]. Institutional culture greatly affects how teaching staff address these issues [2]. To overcome such demanding challenges the teachers in Western Balkan need opportunities for inter-institutional collaboration, as for example provided via Erasmus+ KA2 Capacity Building program.

In this paper, we present Erasmus+ KA2 Capacity Building project BENEFIT (2017-2021), targeting ICT engineers' profile to meet modern society and industry needs [3]. Paper presents project goals and scope, as well as introduces the consortium in Section 2. In Section 3 some significant efforts and results are highlighted. Section 4 provides discussion and conclusions.

2 Project BENEFIT overview

2.1 Goals and aims

With the aim to collaboratively work on above mentioned issues the BENEFIT partners titled the project Boosting the telecommunications engineer profile to meet modern society and industry needs, focusing on ICT engineering education in the region, while following international guidelines, industry needs and student expectations. University of Klagenfurt is coordinating the project [3].

Motivation for BENEFIT project is to connect regional universities that provide telecommunication and ICT engineering programs with the goal to make their study programs more attractive and adapted to regional industry. A common HEI-industry e-platform is established acting as a web repository of class and other materials [4].

Main project goals can be summarized as:

1. Modernization of telecommunications/ICT engineering study programs,
2. Adoption of modern teaching methodologies,
3. Improved cooperation between HEIs and the industry,
4. Creation of joint university-industry labs.

Within BENEFIT project, we have identified several paths to achieve the project goals, among others:

- Extension of cooperation between HEIs and regional industry beyond current industry project partners
- Modernization of existing study programs including enhancement of existing and introduction of new courses
- Development and adoption of new teaching methods and tools
- Implementation of trainings for both teachers and students as a continuous process.

2.2 BENEFIT consortium

The consortium consists of academic partners from four countries [3]. From Western Balkan six universities from Bosnia and Hercegovina - University of Tuzla, University of Sarajevo, and University of Banja Luka, and from Republic of Serbia - University of Belgrade, University of Novi Sad, and University of Niš. From EU three universities are leading the process, namely, from Austria University of Klagenfurt, from Slovenia University of Ljubljana, and from Croatia University of Osijek.

Besides HEIs, there are also important regional industry partners involved in the project, namely, Ericsson Nikola Tesla Zagreb, Cisco Belgrade, Bicom Zenica, BIT Centar Tuzla, NiCAT Niš and RT-RK Novi Sad. Several companies are also serving as associated partners AlfaNum, Vojvodina ICT Cluster and Zesium Mobile.

3 Study programs in the field of telecommunications and ICT

As one of the first steps current study programs at all partner HEIs were carefully analyzed. The acquired data has been analyzed aiming at identifying current status of the academic study programs. The analyzed data gives an insight into main similarities and differences between the programs, as well as existing deficiencies. For example, EU HEIs programs consist of 3 years for the first cycle (bachelor) and 2 years for the second cycle (masters), while all programs in Serbia as well as in Bosnia and Herzegovina (besides program at the University of Sarajevo) 4 years for the first cycle and 1 year for the second cycle. This fact can be an obstacle for harmonization of programs and future establishment of joint studies between EU and WB partner institutions.

The analysis showed that although study programs provide good theoretical background, there is a lack of appropriate laboratory equipment and new methodologies aimed at active learning, as well as adequate collaboration with industry to provide the students with the practical skills needed today [3, D1.1]. Another deficiency of existing programs is that there is no first-year course directly related to modern telecommunications technologies that would encourage students to enroll in the ICT field of study. Furthermore, students usually perceive current programs as classic telecommunications programs because state of the art ICT topics are not sufficiently emphasized.

3.1 Bologna process

The modernization of telecommunications engineering study programs must provide international perspectives and reflect a global view of new telecommunications technologies related to developments in electronic, computer and software engineering.

There are several curriculum guidelines provided by national and international institutions such as ABET, ACM, where within the EU the Bologna process is most relevant. The aim of the BENEFIT project is to follow the Bologna process considering autonomy of the universities while capturing changing needs of industry, expectations of students and society. These inputs were collected during the project, among others via open day events, as for example, Open event Ljubljana 2018 with representatives of students, University of Ljubljana career center, Slovenian Chamber of commerce and HR personnel from the industry as well CMPEPIUS [3].

Main aspects of the Bologna process relevant for the BENEFIT project are efforts towards comparable degrees, mechanisms for transferable academic credits, positioning students as essential partners, promotion of inter-institutional cooperation, and various mobility schemes. In addition, other guidelines are also consulted, as for example following the outline of ACM computer engineering body of knowledge as a framework for our work [5]. As a result of analysis of current study programs from all participating HEIs a three-level hierarchical structure was developed with seventeen knowledge areas. Based on this body of knowledge later knowledge units were identified and implemented in selected existing and new courses.

3.2 Framing the regional industry needs

To have better insights into local and regional industry needs related to ICT engineering studies a survey was conducted among 49 companies in the region. The survey has been prepared to collect information on job market/needs and to map skills and knowledge areas required from ICT engineers and specialists in the ICT sector. The aim was to gather relevant information from industry in the region so that academic partners would have current information in the process of modernization of the study programs [3, D1.1].

The survey included questions related to professional skills, which are most important for an ICT engineer and skills that current engineers are lacking most. Companies have reported that majority are facing problems with ensuring adequate skills of their employees as well that they are facing difficulties in filling vacancies.

Related to the ICT engineering profile they report insufficient supply of qualified candidates with adequate skills, following lack of work experience and on the other side insufficient salaries to attract most qualified candidates.

We have also investigated relevance of individual knowledge areas covered by current study programs in telecommunications and ICT engineering to better inform current and future improvements of the study programs.

3.3 Updated and novel courses

The body of knowledge was created on the existing university study programs with a vision of modernization of these study programs at partner universities. As part of the first preparatory work package seventeen knowledge areas were defined based on analysis of existing study programs [4, 5]. The knowledge areas include topics in telecommunication/ICT engineering as well as core sciences. Each knowledge area consists of knowledge units further composed of topics and learning outcomes.

The body of knowledge is the basis for creating new courses and modernizing existing ones. The course may include knowledge units from one or more knowledge areas, depending on the learning outcomes for a given course.

As writing the learning outcomes is one of the fundamental elements in the Bologna process all courses within BENEFIT project were prepared using the guide for writing and using learning outcomes [6]. As part of the project activities, a teacher training was conducted in 2019 as a full day workshop at University of Ljubljana, University of Banja Luka and University of Osijek. As part of the internal evaluation of the two-day event, comments from the participants were very positive, indicating that such trainings are needed and well accepted by the academic staff. After the event certificates of attendance of the workshop were provided to the participants.

At first cycle level, 32 courses were modernized, and 8 new courses were introduced. These courses covered 10 knowledge areas. At second cycle degree level, 11 courses were modernized, and 13 new courses were introduced. These courses fit into 8 knowledge areas. Some of these novel and modernized courses for ICT study programs are also used as elective courses at other study programs at universities.

The process of creating a new course or upgrading an existing course was defined in the following steps:

1. Determine Knowledge Areas that cover objectives and outcomes of the course,
2. Select knowledge units and sub-units from Telecommunication/ICT Body of Knowledge,
3. Prepare course materials and select appropriate teaching methodologies,
4. Upload material to the e-platform.

The preparation of new educational materials also involves introduction of new teaching methodologies (Section 3.4) and requires modern laboratory equipment for the realization of accompanying lab sessions (Section 3.5). All teaching materials are uploaded to the e-platform and are accessible to all students from participating universities. The process was developed and trained within local and regional workshops.

3.4 Teaching methodologies

In modern teaching, teachers become facilitators. They need to provide additional class readings to facilitate discussions, plan individual and group project work in the class and outside the class, facilitate interaction with the industry and society, and give opportunities to students to present their work and to tutor other students [2].

Therefore, the process of updating the existing courses as well as the development of new courses also require introduction of new, more active approaches to teaching and learning. Identification of appropriate teaching methodologies for ICT engineering studies resulted in a collection of nearly 30 innovative teaching/learning methodologies, grouped into three groups [3, D3.1]:

- Student oriented,
- Technology oriented,
- Activity oriented (the largest group).

Several workshops were organized to present and demonstrate selected teaching methodologies to teachers and assistants, including the project meetings and conferences in Tuzla (InTsinkt), Novi Sad (ZINC), Osijek, Banja Luka, Belgrade (TELFOR) and Sarajevo.

From investigated teaching methodologies teachers have selected appropriate teaching methodologies for 62 novel or enhanced courses to be implemented and investigated in the classroom. The most frequently selected teaching methodologies were: Project-based learning, Active learning, Research-related teaching, The case method, and Teaching support via websites and social media. However, the selection of teaching methodologies has been made before the Covid-19 pandemic time. An analysis of changes in selection of teaching methods has been conducted and proved that online courses (video lectures), creative assignments, self-learning, pre-lecture based learning, as well as audio-library have some advantages under the circumstances of remote-online classes. More details on our experience with the teaching methodologies for ICT engineering education are available in [7].

In order to make all relevant information and materials available to students and other interested stakeholders, a web platform with an official project website and specialized e-platforms was created [3, 4, 8]: Study programs web portal, Teaching material repository, Industry catalog and Industry information portal.

3.5 New university-industry labs

An important part of the project activities is creation of novel thematic joint university/industry labs. All WB HEIs are involved in the process of joint university-industry lab development: University of Banja Luka: “Signal Processing in Telecommunications Lab” in collaboration with Bicom and AlfaNum; University of Sarajevo: “Telecommunications Lab” in collaboration with BIT Centar; University of Tuzla: “VoIP Services Lab” in collaboration with Bicom and BIT Centar; University of Belgrade: “Networks and IoT Lab” in collaboration with CISCO; University of Niš: “Machine-

to-Machine Communication Lab" in collaboration with NiCAT; University of Novi Sad: "Wireless Communications and Information Processing Lab" in collaboration with RT-RK and Saga – the laboratory will be located in the Scientific and Technology Park with opportunity to collaborate with many other ICT industrial partners.

The aim of these new joint university-industry labs is to support laboratory practice with modern state of the art equipment. Joint activities with the industry will be implemented through the specific agreements related to operation of these joint labs, as for example industry trainings for students and upgrading the capacity of teaching staff. The partners collaborate in the development of teaching methodologies that involve participation of industrial partners, including joint project tasks for students, implementation/development challenges, hackathons, and team competitions.

Practicums and promotional video lectures are under preparation; some of them are already posted on e-platform [4, 8]. Most of new courses have defined topics for student projects and theses in collaboration with the industry. Moreover, several summer schools, internships, hackathons, and student competitions have been organized [3, D4.1], more in 2019 and less in 2020 due to pandemic. These joint labs activities will continue also after the project lifetime.

3.6 Quality management of the project

Relevant quality control measures are a formal requirement within Erasmus+ assessment criteria and ensure efficacy of project actions in all project stages. In order to persistently monitor the project actions and to foster continuing BENEFIT activities throughout and after the project duration, Quality Assurance Plan (QAP) was established [8]. QAP provides quality measures for monitoring project implementation based on project indicators. The Plan systematically addresses project documentation and reporting, defines assessment of deliverables, standardizes event reporting and analysis, provides publicity control, and ensures consistency of project documents through standardized templates.

4 Discussion and Conclusion

Project BENEFIT will influence the future development of ICT engineering study programs at the participating institutions by enhancing and modernizing engineering education at both first and second cycle levels. The activities conducted by the project partners since 2017 promote regional cooperation and mobility, training of teaching staff, inclusion of students and increased cooperation with local and regional industry.

Our work started with the consolidated ex-ante analysis of current study programs and industry needs and analysis of national and international guidelines aimed at transition of telecommunications engineer profile towards more ICT oriented studies. Current study programs are influenced with modernization of selected courses and introduction of new courses based on establishment of novel university-industry labs.

During the project, the importance of long-term regional relationship and collaboration between HEIs has become even more evident. Collaboration between regional universities as well as between universities and regional industry are supported through the common HEI-industry e-platform [4, 8]. Project offers increased opportunities to our students with connections with industry and regional participating universities.

The activities related to improvements of study programs need to remain continuous process. During the project, we improved our understanding of importance to frame industry needs, increase involvement of students in every activity, engage teaching staff, etc. Joint industry-university labs facilitate activities far beyond individual course needs enabling better cooperation with industry and offering support to research activities.

Acknowledgments

This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP.

References

[1] Aničić, K.P., Divjak, B. and Arbanas, K., 2016. Preparing ICT graduates for real-world challenges: results of a meta-analysis. IEEE Transactions on Education, 60(3), pp.191-197.

[2] Carberry, A.R. and Baker, D.R., 2018. The impact of culture on engineering and engineering education. In Cognition, metacognition, and culture in STEM education (pp. 217-239), Springer.

[3] Project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs, https://www.project-benefit.eu/, accessed August 2020.

[4] E-platform of the project BENEFIT, https://www.project-benefit.eu/eplatform/, accessed Aug. 2020.

[5] Curricula Recommendations: Computer Engineering Guidelines 2016. ACM, 2016. Available at: https://www.acm.org/binaries/content/assets/education/ce2016-final-report.pdf, accessed Aug. 2020.

[6] Kennedy, D. Writing and using learning outcomes, Quality Promotion Unit, UCC, 2007.

[7] Delić, V. et al. Teaching Methodologies for ICT Engineering Education before and after the Pandemic Developed in the BENEFIT Project, ERK, Portorož, Slovenia, Sep. 2020.

[8] Rimac-Drlje, S., Žagar, D. BENEFIT Industry information portal - a service for better cooperation between universities and companies in the teaching process, ERK, Portorož, Slovenia, Sep. 2020.

[9] Burnik, U. et al. Quality management of Erasmus+ BENEFIT project, ERK, Portorož, Slovenia, Sep. 2020.

Quality management of Erasmus+ BENEFIT project

Urban Burnik[1], Francesco Marcuzzi[2], Dušanka Bošković[3], Drago Žagar[4], Nataša P. Maleš-Ilić[5], Darko Huljenić[6], Matej Zajc[1]

[1]University of Ljubljana, FE, Tržaška 25, SI-1000 Ljubljana
[2] University of Klagenfurt, NES, Lake Side Park B02 – A-9020 Klagenfurt, Austria
[3] University of Sarajevo, FER, Zmaja od Bosne bb, 71000 Sarajevo, BiH
[4] University of Osijek, FERIT, Kneza Trpimira 2b, 31000, Osijek, Croatia
[5] University of Niš, FEE, Aleksandra Medvedeva 14, 18106 Niš, Serbia
[6] Ericsson Nikola Tesla d.d., Krapinska 45, HR-10 002 Zagreb, Croatia
E-mail: urban.burnik@fe.uni-lj.si

Abstract. *The paper presents the quality management system of the BENEFIT Erasmus+ project. Quality management is an important part of planning Erasmus+ projects. After a brief project introduction, we present general quality management aspects related to projects of this nature. We then present the project management structure followed by a detailed description of quality assurance plan and its development.*

1 Introduction

The members of the consortium have initiated project BENEFIT to modernize study programmes in the area of telecommunication engineering towards ICT engineering, intending to meet the demands of regional industry [1]. The project provides a framework supporting regional collaboration among the stakeholders in order to increase the quality and attractiveness of telecommunication studies and to harmonize the study programmes and methodologies with EU anticipations. The consortium consists of several universities, of which three are located in Bosnia and Hercegovina, and three in Serbia, while the action is supported by three universities from EU member states. Within the project, the consortium has introduced modernized study programmes, several new teaching mechanisms, combined with laboratory and industrial training. Industrial and educational experts evaluated the concept of BENEFIT and its new teaching and training methods. The main motivation of the participating Universities was to regain the student interest in study programs in telecommunication engineering, as the enrolment numbers recently decreased despite the dominant role of ICT industry in the modern economy. The labour market struggles for a new profile of telecom engineer (TE): a professional who is trained in a broad ICT domain, possessing both technical and market-oriented skills and consequently can contribute to the techno-economic society challenges. Accordingly, BENEFIT shapes a modern profile of graduates in telecom/ICT engineering, bridging the gap with the job market needs, specifically demanded by Western Balkan (WB) countries, which tend to be integrated in the EU.

The EU character of the project supports modernization of the telecommunication engineers' profile by bringing the experience of EU High Education Institutions (HEI) to enhance the quality of study programmes in WB. More skilled and competent graduates will reflect in better-qualified engineers who will contribute to the innovation process of the regional industry. Study programme improvements, innovative teaching and training methodologies, new labs, and internships will prepare students for a flexible international job market and enhance their mobility opportunities. The projects' idea success depends on the partners' active engagement in exchanging of good practices, enhancement of curricula and their contents, facilitation of mutual studies and degrees recognition as well as on active cooperation with industry. In order to persistently monitor the project actions and to foster continuing BENEFIT activities throughout and after the project duration, a formal Quality Assurance Plan was established, which will be presented in the following.

2 Quality management in Erasmus+ projects

Relevant quality control measures are among the required elements in Erasmus+ assessment criteria [2]. The existence of relevant quality control measures was already part of the project proposal. However, creation of a detailed Quality Assurance Plan was one of the project tasks that underwent within the first months of the project lifetime and were regularly updated to reflect the identified requirements as the project was in effect. The project consortium comprises 15 members: 9 universities and 6 industry partners. Project duration was originally foreseen for 36 months (October 2017 – October 2020), but due to the COVID-19 pandemic introducing some delays in implementation it is extended for additional 4 months (until February 2021). It is important to note the scope of the project and significance of an efficient and effective quality management system for the overall success of the project. We have resourced to formal project quality control requirements with a reference to former successful Erasmus+ projects in the region.

J. Kontio et al represent a systematic approach of quality assurance procedures where eight European universities identified a need for enhancement of self-evaluation and creation of processes with peers in order to reduce the level of inertia due to evaluation procedures in modernisation of their own study programmes [3]. T Hjeltnes et al present a comprehensive overview of quality assurance methodology in form of a case study, based on project "9-conversations", where they study internal, external and national evaluation of the

Erasmus+ projects based on interviews and empirical data from the project in order to advise on evaluation and quality assurance processes in general [4]. We have identified a highly relevant source of knowledge, including templates and sample documents in quality assurance documentation related to project "Re@WBC", which addressed management of human research potential at WB universities in line with national and EU strategies for researchers [5]. Another valuable resource is a well elaborated Quality Control and Monitoring manual of Erasmus+ project "PH-ELIM", available online at [6]. Apart from these resources, the following formal requirements and guidelines served as a resource in creation of the Quality Assessment Plan:

- EACEA – BENEFIT project Grant Agreement
- BENEFIT project Dissemination and Exploitation Plan
- BENEFIT project budget and task assignment
- Erasmus + Guidelines for the Use of Grants
- Erasmus + Frequently Asked Questions

3 BENEFIT Project Management and Quality Board

The BENEFIT Project Management (PM) is implemented via the following structure:

- Management Board (MB)
- Development Board (DB)
- Quality Control and Management Board (QB)

The MB consists of one representative per project partner country and is responsible to monitor the project progress, the achievement of milestones and the delivery of planned results as well as monitor the financial aspects and the use of resources. The DB consists of WP leaders and reports the progress of activities, achievements and possible difficulties. Quality Control and Management Board (QB) comprises of one representative of the participating universities per country, of three external experts outside the consortium and of two student representatives. It is responsible for quality control and monitoring throughout the project. In case of identified issues, MB analyses the situation and eventually proposes a contingency plan.

Quality Control and Monitoring (QCM) is being performed continuously during the project duration and includes:

- Internal control of project progress and outcomes.
- Monitoring and evaluation of deliverables.
- Monitoring and evaluation of the events, (management meetings, students and teacher trainings, Webinars).
- Suggestion of improvement strategies, including evaluation of offered modified and new courses, improved teaching and lab facilities as well as teaching staff and student training.
- Monitoring the achievement of objectives such as impact of the project at each university, profile of graduates, improvements in skills, cooperation with industry, response to industry needs, impact of the project at regional/national level, enrolment of students in new and updated courses and employability of the graduates resulting from the project actions.
- Collecting feedback from students and industry representatives via questionnaires and surveys, taking advantage of direct communication and social networks.
- Monitoring students' ability to enter the job market in the industry and through entrepreneurial initiatives as start-ups.

The QCM activities follow the requirements specified in the projects' Logical Framework Matrix and are coordinated by QB, which organizes monthly conference calls complemented by annual board meetings and assessment visits at selected participating institutions. The Board by cooperation and approval of each project partner prepared projects' Quality Assessment Plan, which details the required procedures regarding QCM.

4 BENEFIT Quality Assessment Plan

Following the state-of-the art findings and good practices in Erasmus+ project management, we have analysed the requirements and specifics of the BENEFIT project. We have identified the main domains of quality assurance for the specific objectives and actions identified within BENEFIT. The result was a Quality Assessment Plan, published as one of the key deliverables within the project. The core quality assessment requirements are categorised and described as follows.

4.1 Quality of project implementation

Most important project requirement is to meet the general objectives of the proposal, which is to render the existing telecommunications study programmes more attractive and to boost the TE profile in the region. In an operational setting, the goal depends on the three BENEFIT pillars (1) cooperation between HEIs and industry to modernize the study program in telecommunication engineering, (2) adoption of modern teaching methodologies and tools, the upgrade of the lab infrastructure, and the creation of joint university-industry labs and (3) implementation of training of both teachers and students.

4.2 Progress indicators

The effectiveness and impact are the main focus of BENEFIT quality control. In order to adequately track the qualitative and quantitative status of achievements on the project, a series of progress indicators have been defined. Scope of the project activities and outreach can be illustrated by providing current values for some of the key indicators:

- Number of modernized courses at 1st cycle of study: **45** /at 2nd cycle of study: **20**
- Number of study programs involved in modernization: 12
- Number of university-industry labs ruled according to an agreement: 6
- Number of events: **18** /project consortium meetings: **4**

- Number of direct beneficiaries in the Partner countries per year: academic staff **101** / students **1207**
- Number of partner country academic staff trained: average of **40** participants p 4 trainings face-to-face, average **50** participants at 3 online trainings

A complete list of project indicators is listed in Table 1.

Table 1.Project Indicators

1. Indicators of progress	
1.1	Number of milestones reached
1.2	Number of meetings
1.3	Number of webinars
1.4	Number of deliverables
2. Indicators of dissemination	
2.1	Project web page
2.2	Number of visits to project web
2.3	Central web platform
2.4	Number of visits to Central web platform
2.4	Designed, printed and disseminated project promotion material.
2.5	Newsletters, e-bulletins, newspapers.
2.6	Reports from presentation meetings, presentations for media and interested groups.
3. Indicators of quality	
3.1	QCB established
3.2	Quality plan
3.3	Quality reports
3.4	Feedback, questionnaire results from clusters and other industrial partners.
3.5	External audits
3.6	Field visits
3.7	Minutes from QCB meetings.
4. Indicators of direct impact	
4.1	Number of direct beneficiaries in the Partner country(ies) per year: academic staff from HEIs
4.2	Number of direct beneficiaries in the PCs (/year): administrative staff from HEIs
4.3	Number of direct beneficiaries in the PCs (/year): HE students
4.4	Number of direct beneficiaries in the PCs (/year): non HE individuals
5. Indicators of training activities	
5.1	Number of partner country "HEIs' students trained
5.2	Number of partner country "HEIs' academic staff" trained
5.3	Number of partner country "HEI's administrative staff" trained
5.4	Number of partner country "non-HEI individuals" trained (priv. sector, NGOs, civil servants, etc.)
5.5	Number of trainings
6. Indicators of modernized curricula	
6.1	Number of modernized courses at 1st cycle of study.
6.2	Number of modernized courses at 2nd cycle of study.
6.3	Number of novel specific-knowledge courses at 1st cycle of study.
6.4	Number of novel specific-knowledge courses at 2nd cycle of study.
6.5	Number of study programs involved in modernization
6.6	Number of knowledge areas
6.7	Number of knowledge areas with modernized courses
6.8	Number of study programs included in the web portal (central platform)
6.9	Number of courses on the web portal (BSc and MSc)
6.10	Number of textbooks published including guides for new equipment.
6.11	Number of new and modernized courses included into audio-library.
6.12	Number of study programs having implemented modernization
6.13	Number of delivered modernized courses (BSc and MSc)
6.14	Number of accredited study programs
6.15	Number of students enrolled to modernized study programs.
6.16	Satisfaction survey of new learning material quality reviewers.
6.17	Students' satisfaction survey reports.
7. Indicators of modernized teaching methodologies	
7.1	Number of new methods
7.2	Number of courses implementing new methods
7.3	Number of hackatons and team competitions
8. Indicators of industry co-operation	
8.1	Number of university-industry labs ruled according to an agreement.
8.2	Number of purchased equipment units (list in the inventory books) and material.
8.3	Number of students included into internships.
8.4	Number of theses co-supervised with industry.
8.5	Number of companies listed in the web catalogue
8.6	Number of collaborating enterprises
9. Indicators of management	
9.1	Partnership agreements
9.2	Internal semi-annual financial reports
9.3	PM board meetings and minutes
9.4	PM board reports
10. Indicators of sustainability	
10.1	Number of university-industry labs ruled according to an agreement.
10.2	Number of signed agreement about long-term cooperation between HEIs and non-academic partners.

4.3 Quality of project activities and documentation

The transparency of project results is best demonstrated through consistent reporting of activities and well-organized documentation. The Quality Board has imposed several good practices in order to ensure consistency and appearance of project activities and documentation. These practices are systematically described as follows.

Document deliverables

The tangible project deliverables are organised in a form of reports, publications, manuals, methodology, plans, printed and electronically available promotional material and follow the common project appearance guidelines. All the deliverables are being analysed and evaluated in terms of relevance (is the deliverables coherent with the planned content), effectiveness (did the deliverable contribute to the achievement of the general and specific objectives) and timing (was the deliverable delivered according to the planned timeline). Being among the most representative of the project results, the quality of deliverables is being maintained at four quality verification levels. These levels include (1) deliverable approval by responsible WP leader, (2) internal review of deliverable provided by project partner not involved in deliverable authoring, (3) quality evaluation of deliverable by a QB board member and (4) acknowledgment of deliverable by each projects partner.

Events

Organised events (training, conferences, info days, etc.) are considered intangible deliverables of the project and are therefore an important project result. The organizers of all the project events (working meetings, training abroad, etc.) should provide a full information package to the participants including the draft agenda, letter of invitation and a note on the logistics in due time. The organizers record the minutes of the meetings, which are distributed to all the project participants and made accessible via project portal. The quality of the project events is ensured by collecting a variety of information using visits, interviews, questionnaires, consultation, and other forms of activities. These will bring awareness of the satisfaction of beneficiaries and other target groups. Posters and other promotional materials are set-up during the event in order to increase visibility of the events. Each event should be documented using project website, Intranet site (Confluence), news, agenda, list of participants, list of trainees, report, gallery, presentations, and/or video materials. Reports include summary review of statistical data to help in final reporting.

BENEFIT websites and other electronic tools

The BENEFIT project foresees an immediate set up of a dedicated project website for dissemination purposes. A dedicated intranet portal is set up and maintained by project coordinator. It can be accessed by all partners depending on their assigned tasks and roles and represents a single point of reference for the project documentation and communication among partners. All partners are asked to promote BENEFIT project on their websites and other electronic tools (such as: Facebook, Twitter and LinkedIn profiles/groups, newsletters, etc.) by providing short description of the project, logo and link to BENEFIT website.

Publicity control

The project coordinator (UNI-KLU) is responsible for the design of the promotional material. The draft version will be sent to all partners for comments and suggestions, before printing, publishing and distribution. The materials will be disseminated by all the partners at appropriate events in order to reach the project's target group.

Document templates

Quality board has prepared and maintained a repository of document templates, which ensure common project appearance of public and internal project documents. The templates were made available via an intranet project repository and included templates of (1) minutes of the meetings (2) project deliverables, (3) PowerPoint presentations, (4) event attendance sheets, (5) meeting evaluation forms and (6) event reports. In addition, templates for internal milestone verification and deliverable approval at several levels were provided. The common user of provided templates promotes the graphical identity of the project and ensures the publicity guidelines of EU are strictly followed. This project logo serves to increase the visibility of the project and has to be used for all the project deliverables and its official project documents.

4.4 Risk Management

Regular risk assessment is to be carried out, which shall lead to corrective actions and potential adaptations of the workplan. This assessment will take care of issues that could endanger the project achievements. In case of serious risks, MB should suggest alternatives, workarounds and the proposed corrective actions that will make the risk consequences acceptable for the consortium. The identification and assessment of new risks is a joint responsibility of all project partners and of external evaluation. Partners may propose preventive actions (avoiding that the risk occurs) and corrective actions (decreasing the severity and impact), specifying also the resources that would be needed.

5 Comments and conclusions

The paper presents a Quality Assurance plan of the BENEFIT project as it has been accepted and enhanced throughout the first two years of project duration. There were five plan revisions, based on the comments of internal and external reviewers and midterm project revision comments of the EC. The quality management activities of the project were commenced as planned. We plan to publish the actual results of project quality evaluation after the project is finished.

Acknowledgements

"This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP".

References

[1] "Project-Benefit," [Online]. Available: https://www.project-benefit.eu/. [Accessed 22 7 2020].

[2] "Capacity Building in the field of higher education - Erasmus+ - European Commission," [Online]. Available: https://ec.europa.eu/programmes/erasmus-plus/programme-guide/part-b/three-key-actions/key-action-2/capacity-building-field-higher-education_en. [Accessed 22 7 2020].

[3] J. Kontio, K. Heikkinen, F. Georgsson, J. Bennedsen and R. Clark et. al., "QA and Enhancement Marketplace for HEIs: An Erasmus+ project," in *CDIO 2015 : 11th international CDIO conference Conceive Design Implement Operate: Collaboration and Extension*, Chengdu, China, 2015.

[4] T. Hjeltnes, A. Fox, T. Hjeltnes, K. Strand and M. Storvik, "Quality assurance and evaluation in Erasmus+ projects, a case study," *INTED2019 Proceedings*, pp. 8758-8764, 2019.

[5] "Re@WBC," [Online]. Available: http://rewbc.ni.ac.rs/. [Accessed 22 7 2020].

[6] "QUALITY CONTROL AND MONITORING," [Online]. Available: http://ph-elim.net/activities/qcm/. [Accessed 22 7 2020].

BENEFIT Industry Information Portal - a Service for Better Cooperation Between Universities and Companies in the Teaching Process

Snježana Rimac-Drlje[1], Drago Žagar[1]

[1]*Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek , Kneza Trpimira 2b, 31000, Osijek, Croatia*
E-mail: snjezana.rimac@ferit.hr

Abstract. *The importance of cooperation between universities and companies not only in research but also in the teaching process is recognized by most European higher education institutions (HEIs). This cooperation is of special importance in the education of engineers and can be carried out through student internships, guest lectures given by industry experts, preparation of diploma theses in cooperation with companies and other activities.*

This paper presents the industry information portal developed within the BENEFIT project as a service for better cooperation of the Western Balkans (WB) HEIs with ICT companies in the region.

1 Introduction

The rapid development of information and communication technologies (ICT) requires adapting the education of telecommunications students to the trends and requirements of the industry. During engineering education, the focus is usually on theoretical knowledge and less attention is paid to developing practical skills professional engineers need [1]. In order for universities to be able to follow the rapid development of technology and provide adequate education to students, it is necessary to strengthen the cooperation with the industry not only in the field of research, but also in the teaching process [2], [3]. To achieve this, it is important to develop appropriate models and methods of cooperation, which includes student industrial practice (internship), preparation of diploma theses in cooperation with companies, application of a project approach as well as engaging industry experts in teaching. It is also necessary to provide an efficient service for the exchange of information between companies and students about the opportunities available to students. This paper presents one such service developed within the Erasmus+ project entitled *BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry need*.

The second section discusses the importance of university-industry collaboration and the activities of the BENEFIT project undertaken to enhance this cooperation in the Western Balkans region (WB). The third section presents the portal STUP, developed at the Faculty of Electrical Engineering, Computer Science and Information Technology Osijek (FERIT), whose purpose is a direct communication between companies and students. In the same section, the industry information portal developed within the BENEFIT project with the same purpose is presented. The fourth section gives concluding remarks.

2 Importance of university – industry cooperation in the teaching process

The importance of cooperation between the industry and universities in student education has been recognized by numerous higher education institutions (HEIs). One of the most important forms of this collaboration is industrial internship. It is an important element in preparing students for future jobs in a real environment, which gives them an opportunity to gain experience of teamwork as well as to improve practical knowledge and skills.

The University of South-Eastern Norway, Faculty of Technology, Natural Sciences, and Maritime Sciences (USN-TNM) introduced the course "Practical Engineering" in the undergraduate engineering study programs in 2015 [1]. In the following three years, 70 students have enrolled in the course and more than 40 companies have offered a work placement. The feedbacks from both students and companies were positive.

In [3], the authors reported on different forms of cooperation with the industry undertaken by DTU Diplom (a center at DTU - Technical University of Denmark), with the aim to provide the majority of its students with the opportunity to do internships and complete their diploma theses in the same company.

The research presented in [4] analyzes the compliance of university education with the requirements of the labor market in the Region of Madrid and the impact of cooperation between the industry and universities on student employability. The analysis results indicate internships have the greatest impact on the employability, which greatly facilitates finding the first job and entering the labor market.

2.1 The BENEFIT project

BENEFIT is a 3-year project co-funded by the ERASMUS+ KA2 Capacity Building Program [5]. The project gathers 6 WBs (3 from Serbia and 3 from Bosnia and Herzegovina), 3 EU universities as well as 11 industry partners. The aim of the project is to enhance the quality of traditional telecommunication studies and

improve the telecom engineer profile in WB region. Furthermore, the aims are to increase the telecom study program attractiveness by modernizing the curriculum, developing new course contents as well as new courses, modernizing laboratories and building joint university-industry labs, improving teaching methods by offering workshops for teachers and creating web platforms for efficient result dissemination.

Reinforcing and broadening the collaboration between universities and the industry are also the BENEFIT project objectives. The project activities with respect to the cooperation with the industry include:

- mapping the competencies and skills of a telecom engineer required by the labor market;
- engaging industry experts in the design of modernized study programs;
- engaging industry experts in the teacher training;
- creating joint laboratories of the industry and academia;
- increasing the number of offered internships;
- developing a web catalog with a list of companies in WB region working in the field of ICT;
- developing an industry information portal containing information on internships, scholarships and graduate theses offered by companies; lectures, workshops as well as vacancies.

At the beginning of the project, two surveys and analyses, academic and industrial, were conducted. The academic survey presents an overview of the curriculum of telecommunications engineers in the EU and WB HEIs, and the industrial survey provides an overview of the labor market needs with respect to the skills and knowledge ICT engineers are required to obtain. One of the conclusions of the academic survey was that there is a lack of a sustainable cooperation model with the main ICT companies in the region which should provide internships for students as well as teacher training. Among other things, the industry survey revealed that the majority of the interviewed companies cooperate with universities in terms of hiring graduates, providing work experience, scholarships and internships, but all respondents suggest “defining and updating educational profiles according to the labor market needs”, followed by “review and change the curriculum to adapt to technological change”, as well as “focus on practical training, organization of practice, practice in the company, etc.”

One of the tools for achieving better cooperation between universities and the industry, developed within the BENEFIT project, is the industry information portal. This portal is based on the portal STUP, developed at FERIT, and both portals allow companies to provide students with all the information they may find interesting in one place.

3 Industry information portal

3.1 FERIT's portal STUP as an example of good practice

Recognizing the need for students to acquire practical skills and knowledge in an industrial environment, FERIT introduced mandatory internships for all graduate students in the second study year starting from academic year 2016/2017. Given the large number of students (about 200), for whom an internship needs to be organized each year, the idea of a web portal that would facilitate this organization was born ultimately resulting in a portal for a direct communication between FERIT students and employers in general. In December 2015, the development of this portal, named STUP, began, and in parallel, different companies working in the field of electrical engineering and ICT were contacted and invited to sign an agreement to use this portal. By the end of April 2016, a total of 47 agreements were signed. STUP was publicly launched on May 1, 2016 (Figure 1), [7].

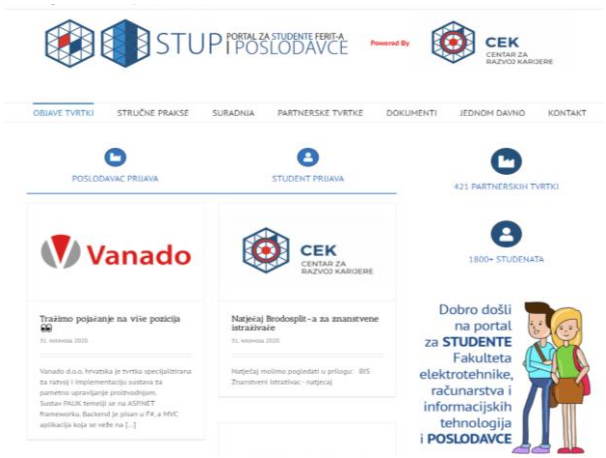


Figure 1. STUP – FERIT's portal for students and employers

Companies registered on STUP can post various notices they find interesting to students, and, on the other hand, FERIT students can open their profiles on STUP and communicate directly with companies. This way of communication, which directly connects companies as an employer and students as future employees, has proven to be very efficient and well accepted by both companies and students. By the end of 2017, 300 companies were registered at STUP and currently 421 companies use this portal. The majority of these companies are from Osijek and Slavonia and Baranja region but there are many from other parts of Croatia and some foreign ones. The increase in the number of companies registered at STUP is presented in Figure 2.

Furthermore, each year more than 200 students open their profiles on the portal and use it until the end of their studies (Figure 2).

Although the portal STUP was launched to support the organization of professional internships, companies

may find this portal more interesting for other things, such as head hunting. However, the portal does an excellent job for professional internships as well as presented in Figure 3. Even though not all companies have offered internships, the number of annually announced places exceeded the number of students who used those internships.

In addition to professional internships, companies post information on vacancies, guest lectures given by their experts as part of classes at FERIT, opportunities for additional education they organize for students, scholarships as well as topics for final and graduate theses. Some of these data are given in Figure 4.

The portal STUP is a good example of how to enable students to get information about all the opportunities offered by companies in one place and, on the other hand, to enable companies to directly contact students, participate in their education thus helping them to better prepare for the labor market. Companies being able to get to know their potential future employees better and train them during their studies for specific jobs, such as working with specialized software and hardware, is of great importance.

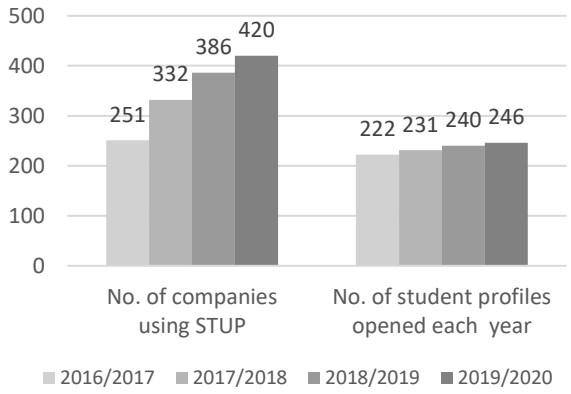
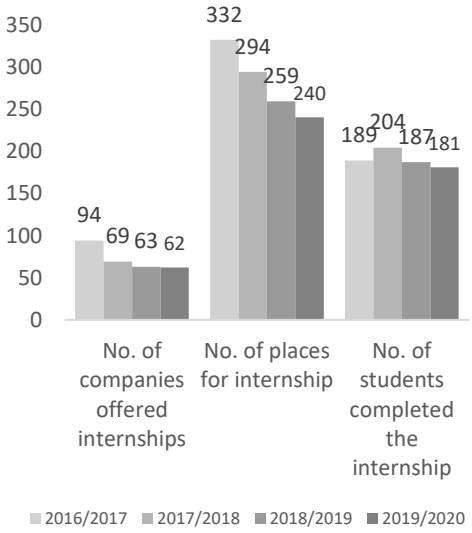


Figure 2. Number of companies and students using the portal



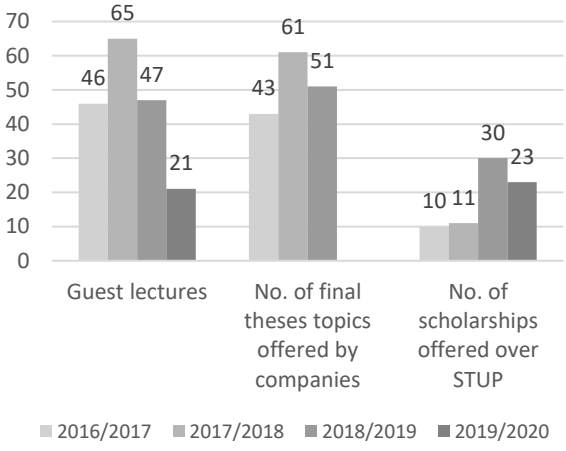


Figure 4. Data about guest lectures, final theses and scholarships announced on STUP

Following these good experiences, within the BENEFIT project the e-platform Industry information portal was created with the same purpose as the portal STUP.

3.2 Industry catalogue and industry information platform – BENEFIT project's e-platforms

In order to improve the cooperation between companies and WB HEIs in the teaching process and to bring companies closer to students, two e-platforms were created within the BENEFIT project, namely *Industry catalogue* and *Industry information portal* (Figures 5, 6 and 7).

The Industry catalogue (Figure 6), contains basic information about registered companies as well as links to their websites and social networking pages. In this way, the catalogue serves as a tool for mapping the economic landscape. The catalogue of ICT companies in WB region makes them more visible and attractive to students and trainees even beyond the region. Currently, 38 companies from Serbia and Bosnia and Herzegovina are registered in this catalogue. Given the good experiences FERIT has with the portal STUP, we expect this number to significantly increase in the future.



Figure 5. BENEFIT project's e-platforms

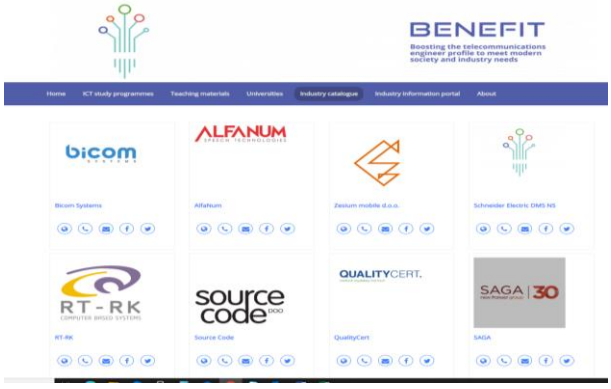


Figure 6. BENEFIT project's Industry catalogue

All companies registered in the Industry catalogue can post notices on the Industry information portal (Figure 7), [8]. The goal of this portal is to offer up-to-date information on the labor needs, vacancies, scholarships, lectures and workshops for students as well as receive feedback from alumni and offer pieces of advice to the student community. Furthermore, this portal is to present information related to the training opportunities and internships as well as graduate theses offered by companies.

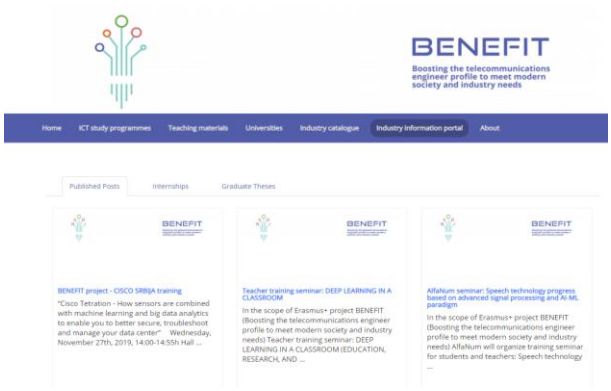


Figure 7. Industry information portal of the BENEFIT project

In order to take full advantage of this portal, the project partners should make additional efforts to promote it to ICT companies in WB region.

4 Conclusions

The experiences of EU universities as well as research conducted within the BENEFIT project show that engineering education requires a higher share of practical work in the curriculum, which can be achieved by greater involvement of companies in the educational process. Industrial practice as well as the preparation of graduate theses in cooperation with companies is of particular importance for the acquisition of practical knowledge and transversal skills.

One way to connect students and companies is through a web portal. One such portal is STUP developed at FERIT for the needs of organizing internships but also for a direct communication between companies and students, which, among other things, helps students find the first jobs. Based on the positive experiences in the use of STUP, the Industry information portal was developed within the BENEFIT project. This portal can be used by all companies registered in the Industry catalogue (also the e-platform of the BENEFIT project) to post information on internships, topics for final theses, vacancies and other activities of interest to students.

Currently, 38 companies are registered in the catalogue and can use the Industry information portal but that number will certainly rise when companies recognize the benefits they could have from using it.

Acknowledgements

This work was carried out within the project *BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs* - 2017-2021 – co-founded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP.

References

[1] R. Immerstein, H. Haselberg and T. Bråthen: Work Placement in Higher Education – Bridging the Gap between Theory and Practice, 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, 2019, pp. 473-477

[2] A. Boyarchuk, V. Kharchenko and V. Sklyar: Models and cases for sustainable university-industry cooperation in IT sector, 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), Kiev, 2018, pp. 667-671

[3] A. Friesel: University-Industry cooperation to empower graduates' preparedness for their future jobs, 2019 29th Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE), Ruse, Bulgaria, 2019, pp. 1-4

[4] A. Afonso, J.J. Ramirez, J. M. Diaz-Puente: University-industry cooperation in the education domain to foster competitiveness and employment, Procedia – Social and Behavioral Sciences, No. 46, 2012, pp. 3947-3953

[5] The BENEFIT project web site, https://www.project-benefit.eu/

[6] Group of authors: D1.1 Consolidated ex-ante analysis and guidelines aimed at boosting the telecommunications engineer profile including a projection of needs for ICT engineers in the future, Erasmus+ project BENEFIT, https://www.project-benefit.eu/wp-content/uploads/2019/02/BENEFIT-D1_1_v16.pdf

[7] STUP Portal, https://stup.ferit.hr/

[8] BENEFIT Industry information portal, https://www.project-benefit.eu/eplatform/?iportal#6

Teaching Methodologies Adopted at the BENEFIT Project for ICT Engineering Education Before and During Covid-19

Vlado Delić[1], Aljo Mujčić[2], Nataša Maleš-Ilić[3], Zdenka Babić[4], Dušanka Bošković[5], Mladen Koprivica[6], Drago Žagar[7], Matej Zajc[8], Nermin Suljanović[2], Andrea Tonello[9]

[1] University of Novi Sad, FTN, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
[2] University of Tuzla, FET, Franjevačka 2, 75000 Tuzla, B&H
[3] University of Niš, FEE, Aleksandra Medvedeva 14, 18000 Niš, Serbia
[4] University of Banja Luka, FEE, Patre 5, 78000 Banja Luka, B&H
[5] University of Sarajevo, FER, Zmaja od Bosne bb, 71000 Sarajevo, B&H
[6] University of Belgrade, ETF, Bulevar kralja Aleksandra 73, Beograd, Serbia
[7] University of Osijek, FERIT, Kneza Trpimira 2b, 31000, Osijek, Croatia
[8] University of Ljubljana, FE, Tržaška 25, SI-1000 Ljubljana, Slovenia
[9] University of Klagenfurt, NES, Lake Side Park B02 – A-9020 Klagenfurt, Austria
E-mail: vlado.delic@uns.ac.rs

Abstract. *Covid-19 pandemic has changed the life in many aspects, including the teaching/learning methodologies. The Erasmus+ KA2 project BENEFIT (2017-21) has been aimed at improving engineering education in the area of ICT, but its goals and teaching methodologies were selected before the pandemic. The last project year covers two semesters of adoption of developed novel and enhanced courses with new labs and teaching methodologies facing new circumstances. The changes in the selection of teaching/learning methodologies and methods for lab sessions and their adoption in modernized ICT engineering courses are analysed in the paper.*

1 Introduction

The project "Boosting the telecommunications engineer profile to meet modern society and industry needs / BENEFIT" (October 2017 – February 2021) is an Erasmus+ KA2 Capacity Building project targeting ICT engineering study programmes in the Western Balkan region [1-4]. The project has been motivated by changes in telecommunication industry and novel technologies such as 5G, IoT and Edge-Fog-Cloud since they are demanding new skills and expertise in the domain of information and communication technologies (ICT) and requiring experts trained to understand and respond to the new information-centric era [1].

The three main goals of the BENEFIT project are:

G1 To modernize teaching methodologies by adopting novel learning/teaching methods,
G2 To upgrade the lab infrastructure through the development of novel thematic industry-academia labs,
G3 To increase the attractiveness of ICT studies through the development of a joint e-platform.

Many activities have been conducted in the first two years of the project timeline and resulted in 64 novel and enhanced courses (21 novel and 43 modernized) at the 1st study cycle level (40 bachelor courses) and the 2nd cycle level (24 master courses) [1], [2, D2.1]. These 64 courses exploit both new teaching/learning methodologies and joint academia-industry labs established at the six Western Balkan universities. The courses have just been accredited, while the joint labs are negotiated by the operational agreements between university and industry partners [1], [2, D3.2]. The teaching/learning methods have been selected by teachers of the 64 novel and modernized courses from a collection of nearly 30 innovative teaching/learning methodologies that are suitable for ICT engineering education [2, D3.1]. The teaching material for most of the 64 courses has been prepared and posted on a common HEI-industry e-platform which is acting as a web repository of class material and also provide many important information related to study programmes and cooperation with industry partners [3]. The project results have been promoted at national and international level, e.g. [4, 5].

However, the BENEFIT project goals, methodology and activities were designed and planned a few years before the Covid-19 pandemic, while the adoption of results, the delivery of new courses based on novel teaching methodologies and the exploitation of the improved lab infrastructure, are facing new circumstances caused by the pandemic. For example, the selection of novel teaching/learning methodologies from the collection has been made without the experience we gathered with the pandemic. This has caused changes in the selection of the teaching/learning methodologies and methods for lab sessions during the Covid-19 pandemic. The changes in teaching methodologies are determined and carefully analyzed in their adoption in the last project year, as described in this paper.

After the introduction which presented the subject of this paper, Section 2 presents the collection of teaching/learning methodologies that was analyzed and considered suitable for ICT engineering education. The initial selection rate of teaching methodologies made by the teachers of the 64 novel and enhanced courses is given in Subsection 2.1, while their plan for changes in the selection and adoption of novel teaching/learning methodologies is presented and analyzed in Subsection 2.2. Section 3 summarizes the discussion and conclusions.

2 Teaching methodologies suitable for ICT engineering education

A set of BENEFIT project activities at WP3 "Modernization of teaching methodologies and infrastructures" is related to the deliverable D3.1 "Development of modernized teaching methodologies". Relationship among WP3 tasks, D3.1 and other deliverables, as well as their impact to the three main project goals are shown in Figure 1. D3.1 is based on the development of innovative training methods involving industry (T3.3) and the development of learning/teaching methods based on student competitions, and development of student challenges and hackathons (T3.4), as well as the adoption of new learning/teaching methods, tools, ICT best practices in teaching (T3.1).



Figure 1. Relations among the project deliverables and WP3 tasks, as well as their contributions to the project goals

After clear identification of limitations of traditional teaching/learning methods on the engineering education (in the preparatory WP1), a collection of nearly 30 novel and innovative teaching methodologies were selected and described as potentially more suitable for ICT engineering education, including usage of e-tools, online courses, social media, cloud-based platforms, etc. [2, D3.1]. Particular attention has been given to the teaching methodologies that involve active participation of industrial partners such as internships, co-advised student projects and theses, that stimulate creativity, innovation and entrepreneurship of students.

Education in engineering disciplines requires a balance between limited time resources and the increased amount of knowledge about a specific topic. Furthermore, the innovations and intellectual property are becoming important as never before. The deliverable D3.1 provides an overview of 30 innovative teaching methodologies that can be adopted as novel/interesting teaching/learning methods for courses in the study programmes for Telecommunications Engineering at the six universities in Serbia and Bosnia and Herzegovina (selected in WP2). The joint result of WP1, WP2 and WP3 is a collection of teaching material for new and modernized courses [2, D3.3] as an impact to the first project goal G1 – to modernize teaching methodologies by adopting novel learning/teaching methods.

The activity to establish of a web servise (T3.5) as a web repository for class and lab sessions material, recorded remote classes, and network of audio libraries [2, D3.4] contributes to the goal G3 – to increase the attractiveness of ICT studies through the development of a joint e-platform. Finaly, the activity to the creation of six joint University-Industry Labs (T3.2) is a direct contribution to the goal G2 – to upgrade the lab infrastructure through the development of novel thematic joint industry-academia labs.

The selected teaching/learning methodologies are divided into three groups oriented to 1) student (S1-S7), 2) technology (T1-T5), or 3) activity (A1-A14), Table 1. More details for each of near 30 teaching/learning methodolgies are given in their references in [2, D3.1].

Table 1. The list of proposed teaching methodologies [2,D3.1] and the number of courses whose teachers have selected them: Before and During the Covid-19 pandemic

Teaching Methodolgies Adopted at the BENEFIT project **1) Student-Oriented Teaching Methodologies:**	**Before**	**During**
S1 The case study method	12	12
S2 Teaching workshops	9	4
S3 Flipped or inverted classroom	7	3
S4 Brainstorming	6	3
S5 Teaching through student competitions	3	0
S6 Student-centered education	2	1
S7 Teaching through debate	1	0
2) Technology-Oriented Teac. Methodologies:	**B%**	**A%**
T1 Teaching support via websites & social media	10	22
T2 Online courses (Video lectures)	6	22
T3 Learning method based on audio library	3	3
T4 Web-based engineering experiments	0	0
T5 Teaching through games (Gamification)	0	0
3) Activity-Oriented Teaching Methodologies:	**B%**	**A%**
A1 Project-based learning	41	41
A2 Active learning	28	18
A3 Research-related teaching	24	25
A4 Z to A approach	10	7
A5 Creative assignments	6	8
A6 Pre-lecture based learning	5	6
A7 Peer group/team (Collaborative) teaching	5	0
A8 Work-based learning	2	0
A9 Self-learning	2	3
A10 Curiosity-driven learning	1	1
A11 Teaching following standardization process	1	0
A12 Block schedule	0	0
A13 Patent base oriented syllabus	0	0
A14 Teaching fundamentals through vocation-related examples	0	0

2.1 Selection of teaching methodologies by teachers of ICT engineering courses at six universities

For example, UNS i.e. the Chair of Telecommunications and Signal Processing (40 members) of the Department of Power, Electronic and Telecommunication Engineering (200) at the Faculty of Technical Sciences (1.200), University of Novi Sad (5.000), provides 98 courses at 7 study programmes, mostly (51) at the Study programme of Power, Electronic and Telecommunications Engineering. Among 49 courses at the Modules of ICT and Signal Processing, there are 21 courses at the 1st study cycle, 15 courses at the 2nd cycle, and 13 at the 3rd cycle (doctoral studies). Out of these 49 courses, 10 courses are created or improved at the project BENEFIT (more than 20%), 6 at the 1st cycle (3 novel and 3 enhanced courses) and 4 at the 2nd cycle (4 novel courses). Most of them are based on new equipment in the Wireless Communications and Information Processing Lab with joint activities and agreements with industry, beginning from RT-RK and Saga companies.

The teaching/learning methodologies have been selected by the teachers of these 10 courses developed at UNS within the BENEFIT project: 8 × A3 Research-related teaching (5/6 at the 1st cycle, 3/4 at the 2nd), 7 × both A1 Project-based learning and A2 Active learning with the same distribution (4/6 at 1st cycle, 3/4 at 2nd), 7 × S1 The case study method (3/6 at 1st, 4/4 at 2nd), 4 × T1 Teaching support via websites and social media, 3 × both S2 Teaching workshops and T2 Online courses (Video lectures), 2 × S6 Student-centred education, while S5 Teaching through student competitions, A4 Z to A approach, and A11 Teaching by following the standardization process (Learning by experiments with open hardware and software testbed) have per 1 choice.

Similar data are gathered for all other Western Balkan universities involved in the BENEFIT project [2].

All 7 **student-oriented teaching methodologies** have been selected by teachers of 64 novel or enhanced courses at the BENEFIT project. S1 The case study method is the most popular in the group (12 out of 64 courses) – equally before and during the pandemic. All other student-oriented teaching methodologies have decreased selection rate during the pandemic: S2 Teaching workshops (9→4 courses), S3 Flipped or inverted classroom (7→3), S4 Brainstorming (6→3) and S5 Teaching through student competitions (3→0 courses).

Five **technology-oriented teaching methodologies** are chosen and described in the collection. T1 Teaching support via websites and social media (10→22 courses) become more and more popular with students from Y and Z generation which lives on social networks. T2 Online courses (Video lectures) (6→22) have the greatest increase in use during the pandemic. Also, T3 Learning/teaching method based on audio library is useful for support since it is easy to recreate new audio-video editions in order to change or improve only some details [5]. Video lectures are recorded as they were performed and it is not simple to make modifications and change some parts since this would require record sections of the lectures.

The largest group of **activity-oriented teaching methodologies** has the most popular ones: A1 Project-based learning (41→41) and A3 Research-related teaching (24→25) which is the most popular at the 2nd level (18 out of 24 master courses). The increase in selection rate have A5 Creative assignments (6→8), A6 Pre-lecture based learning (5→6) and A9 Self-learning (2→3), while several teaching methodologies from the activity-oriented group have great decrease: A2 Active learning (28→18), A7 Peer group/team teaching (Collaborative teaching) (5→0), A4 Z to A approach (10→7) and Work-based learning (2→0).

2.2 Changes in the selection of teaching methodologies due to the pandemic

There are teachers at the 64 new courses who choose the same teaching methodologies before and during the pandemic, but adapted to work remotely and through video platforms that they used for teaching.

Table 1 provides information of changes in the selection rate of teaching/learning methodologies by the teachers of 64 novel and enhanced courses at the project BENEFIT. It is one of the main results of presented research which was conducted during the summer 2020, i.e. between the spring semester with the first experience in the new circumstances in higher education due to the pandemic and the next autumn semester.

After one semester experience with the pandemic, the selection of teaching/learning methodologies has been changed so that T2 Online courses (Video lectures) has become usual for almost all courses, while some of teaching/learning methodologies become more or less frequently selected. Many teachers have removed some teaching methodologies from their plan such as A2 Active learning (-10), A7 Peer group/team teaching (Collaborative teaching) (-5), S2 Teaching workshops (-5), S3 Flipped or inverted classroom (-4), S4 Brainstorming (-3), A8 Work-based learning (-2). On the other side, apart from T2 Online courses (Video lectures) (+16) and T1 Teaching support via websites and social media (+12), other teaching/learning methodologies have become more common: A5 Creative assignments (+2), A6 Pre-lecture based learning (+1), and A9 Self learning (+1), while some methodologies keep a high rate such as A1 Project-based learning (41→41) and A3 Research-related teaching (24→25), Figure 2.

Different e-learning tools and platforms are widely used for distribution of learning materials, conversation with students, as well as their testing and evaluation, like Moodle or Sova. The e-platform of the project BENEFIT has also become popular as a web repository for teaching material as well as the connection between university and ICT industry in the region. The teachers use different platforms for study group support, discussion and opinion pooling, which enables chat, conversations, meetings and video conferencing including online courses and recordings of video lectures (Microsoft Teams, Google Meets, Cisco Webex, Zoom, Skype, Discord). Some of teachers also use web-sites which offers video streaming, like Youtube.

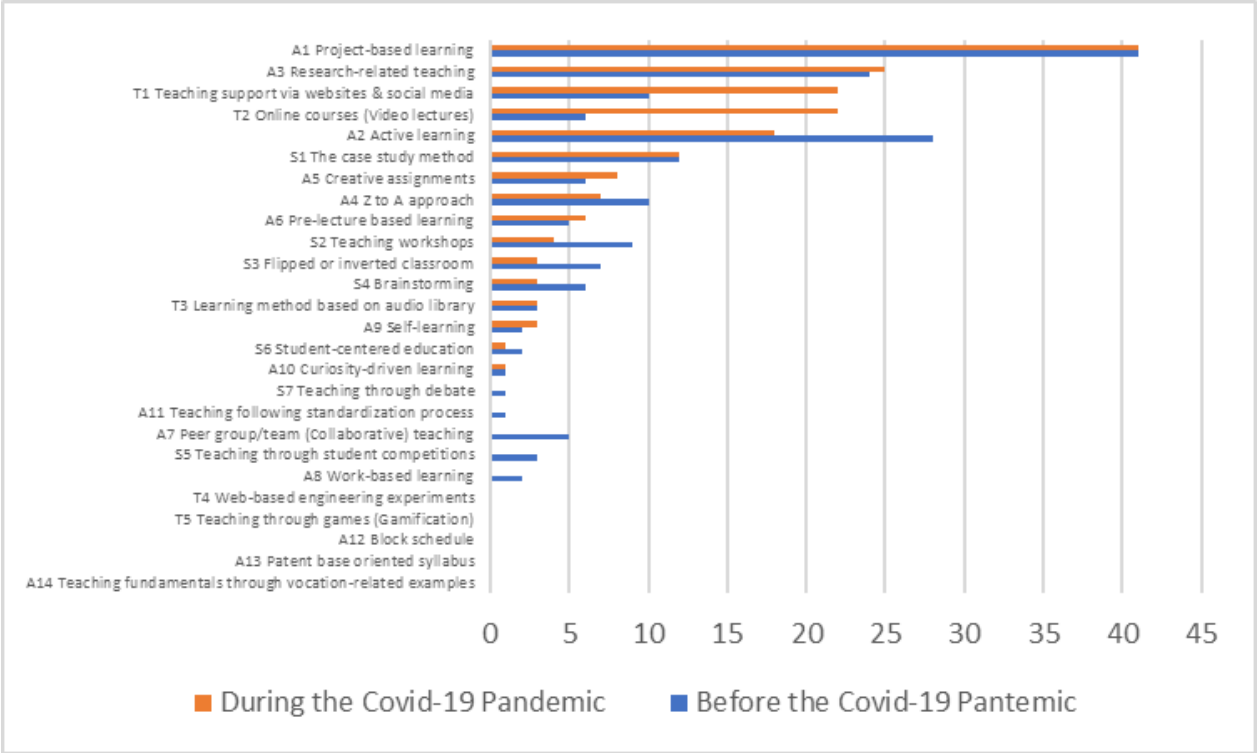


Figure 2. Frequency of selection of the teaching methodologies before and during the Covid-19 pandemic out of 64 novel and enhanced courses which are developed at the BENEFIT project

Regarding the Lab sessions, most of teachers at the project BENEFIT prefer a) Assigning small projects that can be carried out at home, while some of them prefer b) Offering remote access to simulation and HW design tools or c) Operating labs remotely with only the teacher or technician in the lab (most equipment can be operated remotely). Simulation software is also popular in BENEFIT labs, but their use is more effective in live. From the logistics point of view, if teaching will be carried out with physical presence and social distancing stay required, the smaller groups will be necessary with shorter duration labs and multiple sessions.

3 Discussion and Conclusion

Changes in teaching and training due to the Covid-19 pandemic have to be more significant in the engineering education field since it needs a lot of laboratory work and development of practical skills.

A range of innovative teaching methodologies and online/e-learning tools are reselected at the project BENEFIT and presented here as more convenient for exploitation in the new circumstances. The largest changes in the selection are noticed for the teaching methods and e-tools that are closer to Online courses. Due to the pandemic, the project has produced more video lectures than in the initial plan. Students have to work more on Pre-lecture based learning and Self-learning methods, while teachers have to work more as facilitators on the Creative assignments and creation of more interactive online lectures and tests for students. On the other side, Peer group/team teaching, Teaching workshops and Work-based learning have to be postponed and redesigned using virtual teaching platforms.

The selected courses of modernized study programmes in the field of ICT/Telecommunication Engineering will be adopted in the winter semester 2020. Their impact to the industry and society needs will be followed till the end of the project and afterwards.

References

[1] M. Zajc, A. Mujčić, A. Tonello, V. Delić, M. Koprivica, S. Rimac-Drlje, U. Burnik, N. Suljanović: ICT engineering study programmes to meet modern society needs: Erasmus+ project BENEFIT, 29th Int. Electro. & Comp. Sci. Conf., ERK, September 2020, Portorož, Slovenia.

[2] Project Boosting the telecommunications engineer profile to meet modern society and industry needs – BENEFIT, https://www.project-benefit.eu/, accessed August 2020.

[3] E-platform of the project BENEFIT, https://www.project-benefit.eu/eplatform/, accessed August 2020.

[4] V. Delić, D. Vukobratović, M. Sečujski, M. Narandžić, Ž. Bojović, T. Lončar-Turukalo, D. Bajović, N. Jakovljević, V. Šenk: Unapređenje profila inženjera telekomunikacija u skladu s potrebama savremenog društva i industrije, XXV skup trendovi razvoja „Kvalitet visokog obrazovanja“, TREND, February 2019, Kopaonik, Srbija, ISBN: 978-86-6022-157-7, pp. 147-150

[5] V. Delić, D. Mišković: First experiences in application of central audio library of the university of Novi Sad and directions for further development. Int. Conf. Exhibition, INFOTECH, June 2020, Aranđelovac, Serbia.

Acknowledgements

This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP.

Role of M2M communications laboratory formed within BENEFIT project in education processes at FEE-UNI

Dejan Milić[1], Nataša Maleš-Ilić[1], Dejan Ćirić[1], Aleksandar Atanasković[1]

[1]Faculty of Electronic Engineering, University od Niš (FEE-UNI)
E-pošta: dejan.milic@elfak.ni.ac.rs

Abstract. *This paper describes the Machine-to-Machine communications laboratory founded at Faculty of Electronic Engineering, University of Nis, within Erasmus+ project BENEFIT. The laboratory is organized in collaboration with Nis Cluster of Advance Technologies NiCAT that is an industrial partner of UNI at this project. The paper presents the purchased laboratory equipment, some established exercises and measurement setups for student practical work in laboratory and researchers activities at the Faculty of Electronic Engineering.*

1 Introduction

Creation of joint university/industry labs and modernization of the lab infrastructure is the task T3.2 defined in BENEFIT project activities. The University of Klagenfurt (UNI-KLU) is the task leader. Every project partner from Western Balkan High education institutions (WB HEI) are involved in this activity and six joint university-industry labs have been developed during the project:

1. University of Banja Luka: "Signal Processing in Telecommunications Lab" in collaboration with Bicom and AlfaNum;
2. University of Sarajevo: "Telecommunications Lab" in collaboration with BIT Centar;
3. University of Tuzla: "VoIP Services Lab" in collaboration with Bicom and BIT Centar;
4. University of Belgrade: "IoT Networks Lab" in collaboration with CISCO;
5.University of Nis: "Machine-to-Machine Communications Lab" in collaboration with NiCAT;
6. University of Novi Sad: "Wireless Communications and Information Processing Lab" in collaboration with RT-RK and Saga.

The new laboratories equipment and infrastructure established in agreement and cooperation with industry partners should enable students the new trainings and internships. The different teaching methodologies adopted within project BENEFIT, which include student projects, problem related tasks, and competitions, relay on practicing in laboratory. Moreover, collaboration with industry in creation of laboratory challenges and advance solutions will provide new knowledge, skills and job competencies and opportunities.

2 Machine-to-Machine Communications Lab - University of Nis in collaboration with NiCAT

Various kinds of new equipment were purchased in project BENEFIT and have been used for development of new lab sessions for different courses. Laboratory exercises setups have been established and modernized as well as new practicums for lab instructions and video tutorials are under preparation. Operational agreement with NiCAT industrial partner was signed. Due to Covid19 pandemic the join activities implementation in collaboration with industrial partner are impeded but will be continued when the adequate conditions for laboratory development are acquired. The list of equipment in the Machine-to-Machine Communications Lab at the UNI purchased during the project BENEFIT is as follows, [1]:

1. **Software packages**: LabVIEW Full Development System , Matlab for education
2. **Computer equipment**: Blackmagic Design SmartScope Duo 4K + DeckLink Studio 4k, Computer Altos Extreme II, Intel Core i7-8700K/32GB/SSD 480GB/HDD 2TB/nVidia RTX 2080), Laptop HP Envy 13-ah0006nm, 4TT50EA
3. **Software define radio platforms**: NI USRP-2901 , FPGA ZC706 Board + AD-FMCOMMS5-EBZ RF transceiver
4. **Electronic circuit devices and educational kits:** Aaronia low-noise amplifier UBBV 0910 , XBee S2D ZigBee Mesh Kit, NI ELVIS II+ Hardware (For academic use only)
5. **Drone:** DJI Dron Mavic Air Fly More Combo (Arctic White)+Tablet Lenovo IdeaTab4 8 (8504X) + 8.0"HD IPS,QC 1.4GHz/2GB/16GB/FaceUnlock/4GB/Andr 7.1
6. **Measurement equipment:** Vector signal generator VSG25A ,Spectrum analyzer USB-SA44B
7. **Optical equipment:** Mechanism for laser coupling with mono mode fiber MBT612D/M, with Olympus lens RMS20X
8. **Audio equipment:** (Microphon NTi Audio M2230, Sound source NTi DS3, NTi Audio tripod NTI 600 000, Calibrator NTi 94/114 dBSPL)

The listed equipment purchased within the BENEFIT project enables the students and researchers to get acquainted with capabilities and potentials of new trends in the communications technology. The mentioned equipment includes software packages MatLab and LabVIEW as well as important hardware components. This equipment is planned to be used in

laboratory work and research projects present within various courses accredited in UNI. The hardware can be divided in several categories. The items listed in group 3 and 4 belong to the category that will enable students to have practical experiences of working with architectures and programming of up to date hardware for contemporary communication technologies and systems such as M2M in IoT and WSN. Audio equipment, group 8, will broaden communication opportunities by using different sound-based hardware platforms. Additionally, computer equipment and software defined radio platforms, category 1 and 3, can upgrade students and teachers knowledge related to modern technology in monitoring real time video transmission processes as well as signal processing (video, data) required in communication world. The category of measurement laboratory equipment, 6, containing signal generators, measurement instruments, and supporting devices will enable generating various laboratory measurement setups and laboratory exercises. The laboratory work in the domain of optical communication systems will be modernised using the items from group 7.

In the application for BENEFIT project, we planned the equipment item *NI WSN Starter* for education in the field of Wireless Sensor Networks (WSN), but the manufacturer National Instruments did not offer this equipment in the time of equipment ordering procedure. The equipment set that adequately replaces the *NI WSN Starter kit*, and properly complements the purchased equipment is also acquired, and it includes the following items:

1. **Development boards:** Zigbee Development Tools (802.15.4) OpenMote B (Super LOW consumption IoT board 2.5GHZ / SubGHz SMA Antena), Development Boards & Kits - ARM Zedboard + SDSoC
2. **Antennas:** Ant Ext 0.617-3.8GHz SMA, Antennas Ext ISM868/915/2.4 SMA
3. **LoRa system equipment:** RF Modules LoStik – EU, RF Development Tools LoRa(R) Technology Eval Kit – 800, Gateways 8 Channel LoRa Gateway Kit comes with Raspberry Pi LoRa and GPS
4. **RASPBERRY PI Platforms with supporting devices:** Single Board Computers RASPBERRY PI 4 MODEL B, 4GB, , Acrylic Case with Fan, Power Supply 15.3W USB-C with 1.5M Cable 5.1V 3A, 16GB Card with NOOBS 3.1 for Raspberry Pi Computers
5. **Audio equipment:** Audio amplifier, Art Pro Audio SLA1 , Audio card, Focusrite, Scarlett 6i6, 2nd Gen
6. **Additional equipment:** USB Cables / IEEE 1394 Cables USB 2.0 M TO F STRAT 2M CORD, Interface Modules 7 Port USB Hub

Our standpoint is that this equipment will enable us to offer wide and thorough practical education in the area of WSN. The listed equipment will offer students opportunity to form the diverse topologies of WSN supported by contemporary communication technologies and protocols (such as Zigbee, LoRa etc), as well as to track, measure and visualize the network parameters.

BENEFIT team from UNI is engaged in forming laboratory exercises for students and guidelines with instructions for the equipment usage. The installation of all equipment will be finalized after the pandemic.

3 Description of some measurement setups

In this section, some measurements setups formed in M2M communications lab as well as their functions and activities of students and researchers are described.

3.1 Circuit design

NI ELVIS II+ in combination with LabVIEW software makes a unique platform for circuit design, simulation, testing and prototyping for exercises in the field of electrical engineering. Laboratory and workshop exercises at FEE UNI were designed to support theoretical classes in circuit design areas: voltage and current measurement, Ohm's low, voltage and current divider, amplitude and phase response of passive and active filters, etc. Theoretical and practical teaching is realized, as well as the active participation of students in the creation and performance of exercises.

National Instruments provides a software suite for the NI ELVIS II+, which has an instrument panel that includes LabVIEW VI's for an oscilloscope, digital multi-meter (DMM), function generator (FG), bode-plot analyser (BPA) etc., which are all configured to work with the NI ELVIS II+, so that students do not require LabVIEW experience to use it.

3.2 RF power amplifier linearization

In order to perform linearization of RF power amplifier by the method that uses even-order nonlinear digital signal, which has been developed by the researcher group from FEE UNI department of Telecommunications. and to demonstrate the results practically the measurement place on Figure 1 was set up. The fabricated two-way asymmetrical Doherty amplifier was excited the useful 64QAM signal and the required second-order nonlinear signals for linearization, which were generated by using NI USRP platforms.

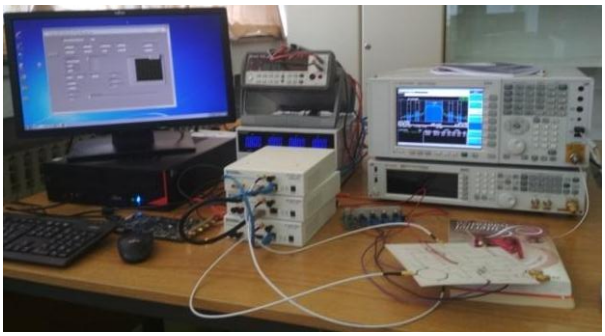


Figure 1. USRPs applied in measurement setup

The USRP platforms were programmed by LabView software to generate the required useful QAM signal,

the signals for linearization, and to control the parameters for the linearization-amplitude and phase. The measurement output spectra for the states before and after the linearization were carried out for different signal power levels.

3.3 Smart Home concept

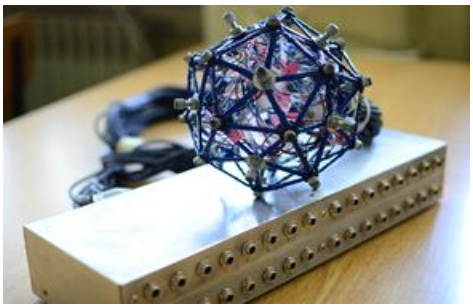
The student bachelor thesis [3] is devoted to the design and realisation of a Smart Home prototype with some basic functions and performances, Figure 2. In addition to theoretical concept of the Smart Home, open hardware platform Arduino UNO and open-source software MQTT Broker and Node-RED are used to monitor and control various devices in the home such as lighting, temperature, movement and some electrical component within the house that can be included in the system. The system is design to operate on a local area network meaning all devices are wired or wirelessly connected to a local WiFi router. MQTT is used as a protocol to communicate between Smart Home devices. Sockets controlled by RF signal and ESP8266 WiFi module are used for switching on and off the device from the power supply. The DHT sensor for temperature and humidity and PIR sensor for movement readings were used. The real-time status of the sensors is displayed on LCD display. The user controls Smart Home system over application on computer or mobile phone. The components applied for the realization of Smart Home concept are procured earlier by some financial supports out of project BENEFIT. It should notice that they are integrated into the laboratory to encompass and complete the concept of M2M communications. In addition, Smart Home example realized in the laboratory is to be further deployed and upgraded by using the equipment purchased in BENEFIT to form another system such as LoRa or by using communications with Zigbee technology.



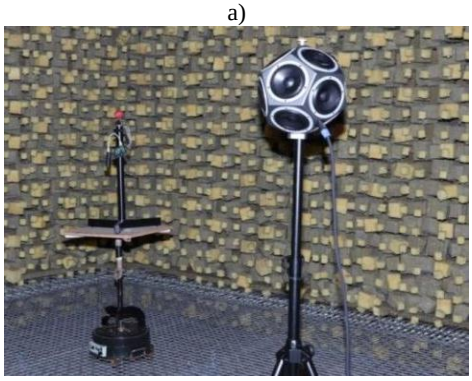
Figure 2. Concept of smart home

3.4 Comprehensive audio system design

Audio technology has recently been passing through a rapid development. It can be considered that important breakthroughs are related to application of digital signal processing, artificial intelligence (AI) and embedded systems. Although classical sound acquisition, processing and reproduction are still widely used, advanced AI-based systems and components have become predominant for a number of tasks. One of such component is a sound acquisition unit based on a spherical microphone array , which are used for spatial sound field analysis.



a)



b)

Figure 3. a) Spherical microphone array in an open configuration, b) omnidirectional sound source applied in measurement of closed spherical microphone array in the anechoic chamber

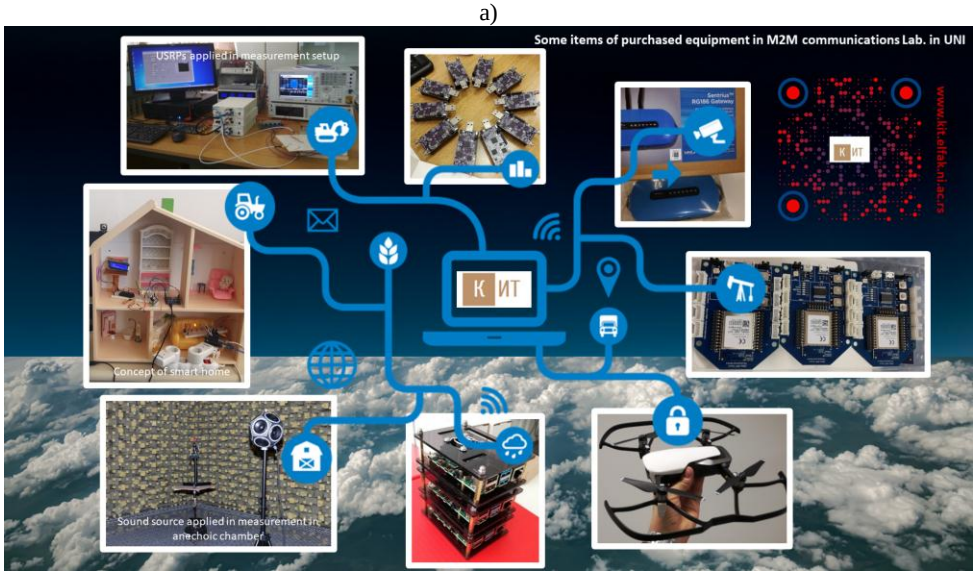
At the laboratory of Acoustics - one of the labs of the department of Telecommunications at FEE UNI, two specific spherical microphones arrays have been realized. The first one is 32 channel spherical microphone array in an open sphere configuration, see Fig. 3(a). The second one, realised by the partner company innSono, is the closed spherical array having 25 microphones. This array in the anechoic chamber of the FEE UNI is presented in Fig. 3(b). One of possible applications of the audio equipment purchased through the BENEFIT project is the measurements of acoustic transducers and systems including spherical microphone arrays. The setup for such a measurement is shown in Fig. 3(b).

3.5 Optical wireless and fiber transmission

Optical wireless represents an indoor alternative to RF transmission which is becoming more congested by the day. Laboratory will make use of this concept to enable student’s hands on experience with free-space optical transmission in M2M environment. In an industrial environment, with lots of noise and interference, signal path may require optical fiber transmission to avoid being contaminated. This is one more topic to be offered at the lab, including exercises with coupling light into optical fiber.



a)



b)

Figure 4. M2M communications laboratory leaflet: a) front page, b) back page

4 M2M communications laboratory leaflet

In order to inform future students of Communications and information technologies module with M2M communication laboratory and its potential, we are proposing in this paper the leaflet represented in Figure 4. It includes fundamental information about M2M laboratory, the equipment, established measurement setups and laboratory exercises. Also, the leaflet includes QR code related to the module enabling the all interested students and others to read more on webpage www.kit.elfak.ni.ac.rs and to find out about our module, study structure and programs, courses, teachers, projects, internships, job offers and opportunities.

5 Conclusion

In this paper we presented the M2M communications laboratory, the purchased equipment, the formed laboratory measurement systems, as well as leaflet that includes relevant information related to the laboratory. Despite the recent global health issues and their impact on education, we have successfully set up the proposed M2M lab in anticipation of crisis resolution. The laboratory will make another useful resource that is going to strengthen the educational direction towards communications, in accordance with project's goals.

References

[1] Erasmus+ project BENEFIT D3.2 Deliverable- Creation of 6 joint University-Industry Labs.

[2] A. Djorić, A. Atanasković, N. Maleš-Ilić, M. Živanović: "Linearization of RF PA by Even-order Nonlinear Baseband Signal Processed in Digital Domain", International Journal of Electronics, Taylor & Francis, Vol.106, Issue No.12, pp.1904-1918, December 2019.

[3] M. Gojković, Diplomski rad-Koncept pametne kućeUNI-FEE, 2020 Niš

Experimental Performance Evaluation of IEEE 802.15.4g Applications Using OpenMote-B Devices

Milica Lekić, Gordana Gardašević

Faculty of Electrical Engineering, University of Banja Luka
E-mail: {milica.lekic, gordana.gardasevic}@etf.unibl.org

Abstract. *The IEEE 802.15.4 standard has become the backbone of numerous low-power and low cost wireless applications. The IEEE 802.15.4g amendment in 2012 has been created particularly for Smart Utility Network (SUN), as well as for Industrial Internet of Things (IIoT) and machine-to-machine applications. The standard aims to provide interoperability of diverse wireless low-power networks and to enable extended coverage. This paper provides the comparison of the IEEE 802.15.4g physical layer performances in three different test scenarios, operating in sub-GHz and 2.4 GHz frequency bands. For measurement purposes, the OpenMote-B hardware platform implementing the IEEE 802.15.4g and equipped with an Atmel AT86RF215 radio transceiver has been used in experimental testing.*

1 Introduction

Low-Rate Wireless Personal Area Networks (LR-WPANs) are being rapidly developed owing to the growing demands for efficient energy consumption and reducing operational costs, thus having a wide range of applications [1]. The IEEE 802.15.4 standard [2] defines a physical layer (PHY) and a Medium Access Control (MAC) layer of Open Systems Interconnection (OSI) model for LR-WPANs. Due to its simplicity and low-cost, this standard is the basis for multiple low-power wireless communication technologies. Widely usage of the IEEE 802.15.4 among standardization bodies and technologies has encouraged its further upgrades. For example, the IEEE 802.15.4e standard [3] defines the MAC layer based on efficient Time Slotted Channel Hopping (TSCH) mechanism. Recently, Low-Power Wide-Area (LPWA) technologies evolve to support long-range, low-cost, and low-power communications, thus extending the functionalities of LR-WPANs. Similarly, the IEEE 802.15.4g standard [4] has developed a new set of PHYs based on three modulation techniques operating in sub-GHz and 2.4 GHz bands and has been designed for outdoor low data rate wireless SUN applications [5].

Given the raising interest in low-rate low-power wireless networks usage in the industrial domain, its range extension and reliability improvement, in this paper we explore the robustness of the different modulation types. In particular, we focus on experimental testing of the IEEE 802.15.4g based applications in three test scenarios using the OpenMote-B hardware and RIOT software platforms. The overall dataset helps in choosing a proper PHY depending on the application requirements.

The paper is structured as follows: Section II provides an overview of the PHYs defined in the IEEE 802.15.4g standard that are evaluated in this paper. Section III presents the methodology and the setup used to conduct the experimental testing of the IEEE 802.15.4g based applications. Section IV summarizes the results obtained from the measurements. Finally, Section V concludes this paper.

2 Overview of IEEE 802.15.4g

The IEEE 802.15.4g standard [4] initially was an amendment to the IEEE 802.15.4 and has been standardized in 2015. The main features of this standard are [4]: operates in free 700-1000 MHz and 2.4 GHz bands; provides data rates from 40 kb/s to 800 kb/s; maximum length of Physical Service Data Unit (PSDU) is 2047 bytes (B) and a complete IPv6 packet can be transmitted without fragmentation; coexistence with other systems operating in the same band (IEEE 802.11, 802.15 and 802.16). The IEEE 802.15.4g defines three PHYs: Multi-rate Multi-regional Offset Quadrature Phase Shift Keying (MR-O-QPSK), Multi-rate Multi-regional Frequency Shift Keying (MR-FSK) and Multi-rate Multi-regional Orthogonal Frequency Division Multiplexing (MR-OFDM) [4]. In this way, the standard provides communication in multiple bands and use of multiple data rates. MR-FSK is a mandatory PHY, the most common in USA in 902-928 MHz band together with Frequency Hoping Spread Spectrum (FHSS) technique [1]. MR-O-QPSK shares the characteristics of IEEE 802.15.4 modulation and there are already O-QPSK devices commercially available. MR-OFDM provides high data rates, spectrum efficiency and robustness against multi-path fading and external interference in environments with frequency selective fading. This standard aims to enable interoperability of networks with different capabilities and capacities by changing modulation and/or data rate on packet-by-packet basis.

2.1 MR-FSK

Any IEEE 802.15.4g device has to support MR-FSK (2-FSK & 50 kb/s) based on Gaussian FSK (GFSK) modulation with 2 or 4 levels providing constant amplitude

of modulated signal [4]. This standard has introduced the novel Mode Switch (MS) mechanism that enables devices to change data rate and/or PHY packet-by-packet basis using MR-FSK PHY. Using MS mechanism, temporary PHY can be changed only for one packet. When communicate, both transceivers have to support desired PHY configuration. MR-FSK supports two Physical Protocol Data Unit (PPDU) formats depending whether MS mechanism is enabled [4]. Additional enhancement is the generic MR-FSK mechanism that provides support to the existing commercial PHY solutions and an adoption of new PHY solutions as a consequence of technological progress or a regulatory change. In other words, this mechanism enables the adoption of changes without the need for standardization.

2.2 MR-O-QPSK

MR-O-QPSK PHY provides multiple data rates by utilization of the Forward Error Correction (FEC), interleaving and frequency spread spectrum technique. There are two frequency spread spectrum techniques depending on the operating band provided: Direct Sequence Spread Spectrum (DSSS) and Multiplexed Direct Sequence Spread Spectrum (MDSSS) [4]. Data transmission with legacy devices is ensured in the following bands: 780, 915, 917, and 2450 [MHz]. O-QPSK-DSSS modulation divides band into 16 orthogonal channels separated by 5 MHz and channel width of 2 MHz. Every symbol presents 4 data bits, so there are 16 possible symbols and each of them is modulated by Pseudo-Random Noise (PRN) chip sequences, which are mutually orthogonal. Data and chip rates provided by MR-O-QPSK PHY are presented in Table 1 [4].

Chip rate [kchip/s]	Rate Mode	PSDU data rate [kb/s]
100	0/1/2/3	6.25/12.5/25/50
200	0/1/2/3	12.5/25/50/100
1000	0/1/2/3	31.25/125/250/500
2000	0/1/2/3/4	31.25/125/250/500/1000

Table 1: MR-O-QPSK data rates.

2.3 MR-OFDM

MR-FSK and MR-O-QPSK are frequently used PHYs in low-power LR-WPANs thanks to its simplicity, low-cost and good performances. Contrary, OFDM modulation is commonly used in systems with strong requirements for signal processing, memory and energy consumption, like xDSL, Long Term Evolution (LTE), WiMAX, Power Line Communications (PLC), and Wi-Fi [6]. OFDM typical application is in security and monitoring systems. MR-OFDM is based on parallel data transmission with orthogonal sub-carriers, each transporting one part of the information in narrow-band channel. This data transmission approach ensures better robustness against multi-path propagation, external interference and improves spectrum efficacy [5]. MR-OFDM provides higher data rates 50-800 kb/s, and the maximum PSDU length of 2047 B [6]. Sub-carrier spacing is constant and equals (31250/3) Hz, and a symbol rate is (23/3) ksymbol/s. OFDM symbol (120 μs) consists of a base symbol (96 μs) and a cyclic prefix (CP). CP presents replication of the last 24 μs of the base symbol and is positioned in front of the base symbol. Cycle feature and the long duration of an OFDM symbol make MR-OFDM PHY more robust against multi-path propagation which can cause Inter-Symbol Interference (ISI) [6]. This PHY offers 4 options (numbered from 1 to 4) characterized by the number of active tones (sub-carriers), signal bandwidth, channel spacing and the number of channels. The Modulation Coding Scheme (MCS) parameter, numbered from 0 to 6, specifies the following: sub-carrier modulation scheme (BPSK, QPSK, 16-QAM), the FEC coding rate (1/2 or 3/4), data rate and whether frequency repetition is applied [4]. The FEC is mandatory in MR-OFDM. Frequency repetition is a technique where more than one sub-carrier (2 or 4) transport the same information. Even though this technique reduces effective data rate, it improves robustness against multi-path fading. The MR-OFDM header is transmitted using the lowest MCS level provided by the option, which reduces energy consumption. MR-OFDM PHY provides extended data rates for options 1 and 2: 1200, 1600 and 2400 kb/s [4].

3 Experimental Setup

Experimental testing of the IEEE 802.15.4g based applications has been conducted based on three test scenarios using the OpenMote-B hardware and RIOT software platforms. The setup consists of two OpenMote-B devices, a transmitter (TX) and a receiver (RX). The traffic analysis has been performed using the following metrics: packet loss [%], average Received Signal Strength Indicator (RSSI) [dBm], min/avg/max Round Trip Time (RTT) [ms] and PHY configuration. The overall acquired dataset is not presented in the paper due to page limitation. The first scenario has been conducted in a controlled environment, i.e. in a Faraday cage which has ensured idealized conditions. The communication performance has been tested in both frequency bands (FBs): sub-GHz (863-870 MHz) and 2.4 GHz. The purpose of this scenario has been to verify accuracy of applied hardware platform. The second scenario has been carried out in sub-GHz FB and in a real-world environment (inside the building in two neighboring rooms) with 10 m device spacing. The goal of this setup is to analyze PHY configuration performances in real conditions. The last scenario has been tested in sub-GHz FB aiming to assess the PHY resistance to the influence of noise. The experiments in this scenario have been conducted using RF coaxial cables and a coupler. The noise has been generated by the Agilent 33500B signal generator. The test considers the following configurations:

- 16 PHY configurations in sub-GHz FB, namely: [MR-O-QPSK, chip rate = 100 kchip/s, RM (Rate Mode) = 0;1;2;3], [MR-FSK, data rate = 50 kb/s; 100 kb/s, modulation = 2-FSK, modulation index = 1, channel spacing = 200 kHz; 400 kHz], [MR-OFDM, option = 1, MCS = 0;1;2;3], and [MR-OFDM, option = 2, MCS = 0;1;2;3;4;5],

- 28 PHY configurations in 2.4 GHz FB: [MR-O-QPSK, chip rate = 2000 kchip/s, RM = 0;1;2;3], [MR-FSK, data rate = 50 kb/s; 150 kb/s; 200 kb/s, modulation = 2-FSK, modulation index = 1;0.5;0.5, channel spacing = 200 kHz; 400 kHz; 400 kHz], [MR-OFDM, option = 1, MCS = 0;1;2;3], [MR-OFDM, option = 2, MCS = 0;1;2;3;4;5], [MR-OFDM, option = 3, MCS = 1;2;3;4;5;6] and [MR-OFDM, option = 4, MCS = 2;3;4;5;6].

3.1 Hardware platform

The OpenMote-B is an open-hardware platform for the IIoT applications in the field of next generation low-power long-range wireless networks based on IPv6 protocol stack. The second version of this platform, depicted in Fig. 1, was released in March 2018 and consists of a Texas Instruments (TI) CC2538 System on Chip (SoC) and an Atmel AT86RF215 radio transceiver. The CC2538 includes an ARM Cortex-M3 micro-controller (32 MHz, 32 kB RAM, 512 kB Flash) and a radio transceiver compatible with the IEEE 802.15.4-2006 standard. The AT86RF215 completely supports the IEEE 802.15.4g standard and provides data transmission in sub-GHz and 2.4 GHz.



Figure 1: The OpenMote-B hardware platform.

3.2 Software platform

RIOT is a free open-source operating system (OS) for memory constrained systems with focus on the wireless low-power IoT devices [7]. Memory size is around 10 KB and it is based on micro-kernel and modular architecture (8, 16, 32 -bit). RIOT OS provides support to multiple protocol stacks as IPv6, 6LoWPAN, and standard protocols: RPL, UDP, TCP and CoAP. Platforms that RIOT OS supports are: TI MSP430, ARM7, ARM Cortex-M0-M0+-M3-M4, AVR micro-controllers and MIPS32r2 [8]. RIOT OS source code is available on GitHub repository. This OS uses broadly spread IT tools: C and C++ programming, gcc, gdb, valgrind tools, minimal code-hardware dependency, development for Mac and Linux OSs. RIOT aims to implement open standards for the IoT systems featured by connection, security, durability and privacy.

4 Results

4.1 I test scenario

To determine the communication link metrics for a given PHY configuration, the TX sends 50 frames of 28 B to the RX with 1 ms inter-packet delay. The transmit power of both devices has been set to 0 dBm. Overall, this scenario enables 44 experiments. These experiments have confirmed that with increased data rate, the RTT parameter becomes smaller. MR-OFDM PHY is the fastest one, which has met the expectations due to its highest data rates. MR-O-QPSK PHY has up to 10 time larger RTT parameter in comparison with other PHYs. This implies that MR-FSK and MR-OFDM PHYs provide better performances in sub-GHz FB. This result confirms that MR-FSK is the most robust PHY in both FBs based on the RSSI value analysis. Another interesting observation is that the difference in the RSSI value between MR-O-QPSK and MR-OFDM2 is less than 2 dB, -34.25 dBm and -36.5 dBm, respectively. Taking into consideration the signal bandwidth of these PHYs - 5 MHz (MR-O-QPSK) and 0.8 MHz (MR-OFDM2), these results prompt the MR-OFDM usage in low-power LR-WPANs.

4.2 II test scenario

Three measurements have been done in this test scenario. In each of them, the TX sends 100 frames of 88 B to the RX with 1 s inter-packet delay. An 80 dB attenuator has been coupled with the RX's antenna interface, and at the TX it has been used attenuator with selective attenuation. In the first measurement, the TX's attenuation has been set to 0 dB, and in the second one to 27 dB. In the first case, the communication has been established for all PHY configurations (16). Contrary, in the second case, the communication link has been established only for MR-O-QPSK RMs 0 and 1 (6.25 kb/s and 12.5 kb/s). During these measurements, enormous human impact on communication link performances was noticed where packet losses have been varying from 20% to 90%. Obtained results are as expected: the bigger the data rate is, with constant power and position of devices, the greater decline in the communication link quality is. Overall results conclude that MR-O-QPSK provides the longest range, i.e. the most robust propagation waves.

In the third measurement, the relative attenuation at the TX has been measured when the communication link is interrupted [packet losses $> 90\%$]. The attenuation is relative because the communication is full-duplex. The results show that MR-O-QPSK PHY supports the lowest transmit power which is in correspondence with the previous conclusions related to this scenario. Fig. 2 summarizes the estimated relative attenuation at the TX, for all tested PHY configurations. The largest value of transmit power is needed for MR-FSK, MR-OFDM1/MCS3, and MR-OFDM2/MCS5 PHYs. MR-O-QPSK is the most robust PHY, while MR-OFDM and MR-FSK PHYs provide similar results. As expected, the higher the data rate is, the larger transmit power is needed. A large decrease of attenuation is observed for MR-OFDM1 MCS2 and MCS3. This can be partly attributed to the fact that MCS3 data rate (800 kb/s) is reduced by half of MCS2 data rate (400 kb/s). The other reason is the usage of frequency repetition technique in MCS2 configuration which provides extra dB of protection. This result clearly shows the benefit of frequency repetition. The same observation can be noticed between MR-OFDM2 MCS2 and MCS3 PHYs, where MCS2 uses frequency repetition and MCS3 does not. Another interesting result comparing the estimated relative attenuation at the TX of MCS1 and MCS2 for both MR-OFDM options is that it is the same and equals 3 dB. These results show that BPSK modulation

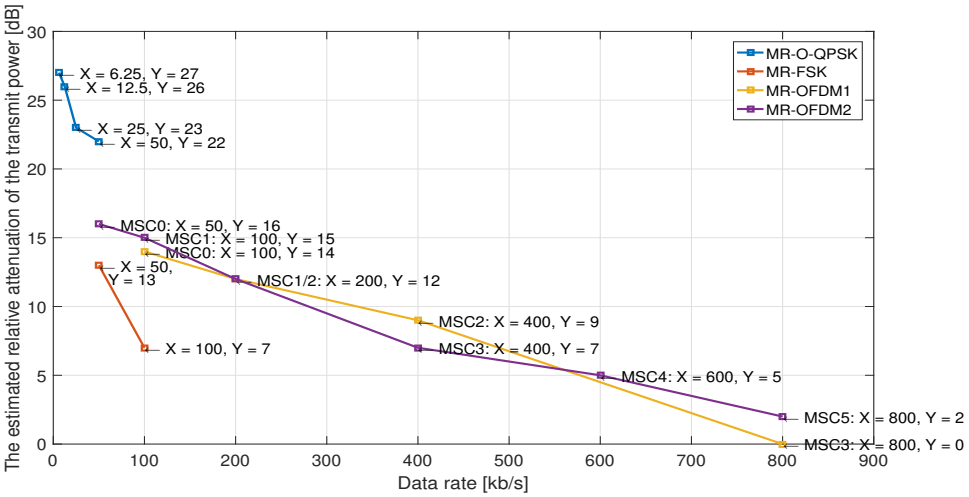


Figure 2: The estimated relative attenuation of TX power when the connection is broken.

provides extra 3 dB over QPSK modulation. Comparing the results of MR-OFDM options, it is concluded that a reduction of the signal's bandwidth makes PHY less robust. This can be attributed to the fact that the Power Spectral Density of the Interference (PSDI) increases as the signal bandwidth is reduced, leading to the increase of Bit Error Rate (BER) (if the signal bandwidth is reduced by half, the PSID increases for 3 dB).

4.3 III test scenario

In this test scenario, the power level of signal and noise have been measured when the connection is interrupted. These values have been measured at the output of a coupler using spectrum analyzer. To determine the communication link metrics for a given PHY configuration, the TX sends 100 frames of 88 B to the RX with 1 s interpacket delay. The results show that the difference between the signal and the noise level is about 15dB in case of MR-O-QPSK PHY RMs 0 and 1, and this difference is reduced to 5dB for RMs 2 and 3. These results imply that as data rate increases, the resistance of PHY to the noise influence reduces. MR-FSK provides noise resistance about 1 dB. MR-OFDM1 configurations have the same value, while MR-OFDM2/MCS0/1/2 provide the difference value of 2 dB, and MCS3/4/5 1 dB. The overall conclusion is that MR-O-QPSK PHY is the least resistant to the noise impact.

5 Conclusions

This paper evaluates the performances and robustness of the PHYs defined in the IEEE 802.15.4g standard, specifically in the sub-GHz FB (863-870 MHz). The experiments were conducted in a controlled setup to evaluate the minimum required TX power by PHY configurations to establish and maintain communication. The results show that MR-O-QPSK provides the longest range in the real conditions, while MR-FSK provides the weakest performances. An interesting observation made from the results is that MR-OFDM PHY is the most resistant to the noise impact. This conclusion is crucial because O-QPSK modulation is widespread in the industry applications, especially when MR-OFDM advantages are taken into account: high data rates, spectral efficiency and an increased level of robustness.

Acknowledgment

This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP.

References

[1] Kuor-Hsin Chang and Bob Mason, "The IEEE 802.15.4g Standard for Smart Metering Utility Networks," *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, pp. 476-480, 2012.

[2] "IEEE Standard for Low-Rate Wireless Personal Area Networks (WPANs)," *IEEE Std. 802.15.4-2015*, 2015.

[3] "IEEE Standard for Local and metropolitan area networks-Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC sublayer," *IEEE Std.802.15.4e-2012*, 2012.

[4] "Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 3: Physical Layer (PHY) Specifications for Low-Data-Rate, Wireless, Smart Metering Utility Networks," *IEEE Std. 802.15.4g-2012*, 2012.

[5] J. Muñoz, E. Riou, X. Vilajosana, P. Muhlethaler and T. Watteyne, "Overview of IEEE802.15.4g OFDM and its applicability to smart building applications," 2018 Wireless Days (WD), Dubai, pp. 123-130, 2018.

[6] P. Tuset-Peiro *et al.*, "Experimental Interference Robustness Evaluation of IEEE 802.15.4-2015 O-QPSK-DSSS and SUN-OFDM Physical Layers for Industrial Communications," *Special Issue on Emerging Trends in Industrial Communications", MDPI Electronics*, 2019.

[7] RIOT OS, "The friendly Operating System for the Internet of Things." [Online]. Available: `https://riot-os.org` [Accessed: July 3, 2020].

[8] E. Baccelli *et al.*, "RIOT: An Open Source Operating System for Low-End Embedded Devices in the IoT," in *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4428-4440, Dec. 2018.

Stimulating Entrepreneurship in Teaching Human Computer Interaction

Dusanka Boskovic[1], Nihad Borovina[2], Merima Zukic[3]

[1]University of Sarajevo, Faculty of Electrical Engineering
[2]BH Telecom dd
[3] University of Sarajevo, Faculty of Philosophy
E-mail: dboskovic@etf.unsa.ba

Abstract. *Software development requires understanding of users, user needs, user tasks and context in which they are operating. These skills are familiar to entrepreneurs, product managers, and marketing experts. However, our teaching experience suggests that students generally find these topics less attractive as they perceive them to be far too theoretical and thus, not as useful. During the years of teaching the Human Computer Interaction course we have noticed students' preferences for learning technology oriented methods, or what we refer to topics belonging to solution domain. The changes in the modernized HCI course introduced Product Market Fit canvas in order to bridge the gap between "theoretical" and "practical" part of the course.*

1 Introduction

The Human Computer Interaction (HCI) course is focused on designing software/system interaction as an approach to develop usable software/system. The HCI course is taught at the master cycle as an elective for students studying Control and Electronics, Informatics and Telecommunications. Students gain knowledge on designing interaction for complex systems such as: SCADA systems, systems for safety and signalization, monitoring and supervision of networks, etc. The process of designing user interaction places a user in the research focus, and user research draws upon good practices from the product development [1, 2]. Interesting example of documenting and managing product requirements based on the user needs is Product Market Fit (PMF) canvas [2]. It enables a developer to directly link problem domain concepts to solution domain.

The Erasmus+ KA2 project BENEFIT (2017-21) [3] has been aimed at improving engineering education in the area of ICT and our course was included in the modernization. In the academic year 2018/2019 we have introduced the PMF canvas, with adjustments for developing software product/service [4]. With this improvement we wanted to stimulate a range of entrepreneurial skills that successful interaction designer requires: critical thinking, multidisciplinary orientation, creative problem solving.

The rest of this paper is organized as follows. Section 2 presents a brief overview of competences related to the entrepreneurship, and appropriate delivery and assessment methods. Section 3 is addressing students' preferences towards learning technology oriented methods, the difference in students' attitude towards topics belonging to solution domain opposite to topics belonging to problem domain. Section 4 describes Product Market Fit canvas, a tool used for documenting and managing product requirements based on user needs. Section 5 briefly describes the experience and knowledge gained from the Teachers Training workshop organized within the framework of the BENFIT project. Section 6 summarizes the discussion and conclusions.

2 What didactics teaches us?

Bellotti et al. state that entrepreneurship is a personal skill and motivation which draws a person to engage his/her abilities and efforts in the creation of a new product and services, with a business value given by the ability to match the market demand [5].

We can use the previous statement to interpret entrepreneurship as a generic skill, since it is not connected to specific field of study. Also, we can think about entrepreneurship as a transferable skill that can be used in all of domains of human functioning. Having said this, market demand and business value of activity depend on the specific context in which person exists. It is not necessary to generate large income or own a company for one to be considered a good entrepreneur.

In regards to entrepreneurship education, it was said that it could not be performed in HEI optimal way, since planning is incompatible with entrepreneur - friendly society [6]. However, as the lifelong learning and personal development are becoming increasingly more relevant and important issues, the entrepreneurship education is gaining popularity, not only with HEI, but also within the society, in general. It is becoming increasingly clear that entrepreneurship should be integrated through the entire education system.

There are a variety of difficulties when it comes to integrating entrepreneurship into standard curricula. It is a transferable and an interdisciplinary skill that needs to take into account pedagogical and psychological knowledge and strategies. At most universities, a primary focus is placed on vocational competency, while entrepreneurial skills are seen to be of secondary importance. There is an issue regarding competences of academic staff for teaching problem solving and critical thinking, and these skills are essential for entrepreneurship. There is sometimes a conceptual ambiguity both on how to teach and how to evaluate development of entrepreneurial skills.

Entrepreneurial approach supports a range of different skills and abilities that are related to creativity, problem solving, positive "to do" attitude, bravery, innovation. It has a multi–disciplinary orientation.

Development of entrepreneurial skills needs to be defined in learning outcomes, reflected on teaching techniques and planned in evaluation. In didactics we call this constructive alignment,

There are different teaching strategies that can be used to foster entrepreneurial skills, but what is important is to plan lessons with this goal in mind and try to engage students in problem solving and critical thinking, no matter the specific topic and strategy.

Assessment tools need to be in-line with learning goals and teaching techniques. Sometimes, classic assessment strategies are not appropriate to measure the level of entrepreneurial skills acquisition.

	Excellent	Competent	Need improvements
Task description			
Dimensions			
Scales			
Description of each level			

Figure 1. Rubric example

Thus, a special attention needs to be devoted to assessment techniques. For this purpose, we have selected Rubrics [7]. This is a special scoring tool that lays out specific expectations for the assignment. The Rubrics are very useful when we want students to develop the ability to appropriately reflect on ill-structured problems. Rubrics divide an assignment into a number of distinct parts and provide a detailed description of what constitutes an acceptable or unacceptable level of performance for each of those parts.

Rubrics include a task description (the assignment), a scale of some sort (levels of achievement, possibly in the form of grades), the dimensions of the assignment (a breakdown of the skills/knowledge involved in the assignment), and descriptions of what constitutes each level of performance (specific feedback) all set out on a grid. In fig. 1, an example of a rubric used during the BENEFIT project teachers training is presented. Participants were asked to develop a rubric as a grading tool for an assignment from their course, and the assignment expectations were defined as a rubric.

3 What we have learned from experience?

During the years of teaching the Human Computer Interaction course we have conducted student evaluation surveys to use data for continuous improvement. The survey is addressing the importance of topics covered in the course and students reflection on their achieved results. We have notices students' preferences for learning technology oriented methods, and topics belonging to solution domain. The results suggest that students find topics belonging to problem domain more theoretical and less attractive and this is related to their perception of usefulness of these topics.

HCI syllabus includes the following topics: usability principles, methods and tools for designing interaction and user experience, prototyping techniques, user interaction design patterns, and evaluation of usability. Traditional lectures covering theoretical principles are followed by the lab exercises, homework assignments and a final project. There is no written nor oral exam. Students are engaged in a project to design and implement an interactive application by using software development technology of their own choice.

Our objective was to find an appropriate learning activity that is able to bridge the gap between "theoretical" and "practical" part of the course. Developing a successful product requires understanding of user needs, user tasks, as well as the context in which they are operating. These skills are familiar to product managers, business analysts and marketing experts. Product Market Fit (PMF) canvas is an interesting example of documenting and managing product requirements based on user needs. It directly links problem domain concepts to solution domain. In this academic year we have introduced this tool, with adjustments for developing software product/service.

After the course completion, we organized the survey again, with two questions for each topic, assessing how useful and how fun it was for students (5 point Liker scale). The topics assessed: Cognitive concepts and interaction design principles, User and user needs analysis methods, Task analysis methods (HTA), UX –Product Market Fit canvas, Prototyping – paper and mockup tools, Interaction design patterns, Visualization, Usability and user experience evaluations. In the end of the survey we asked students to pick up the three concepts they have found the most important for Interaction Design and also to evaluate the level of their competences.

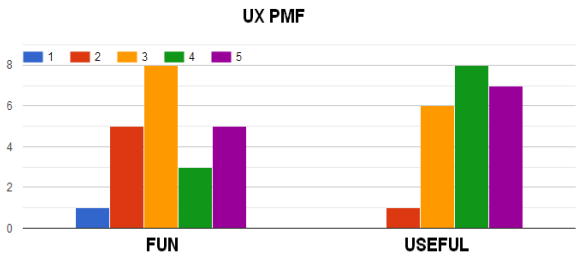


Figure 2. Evaluation of UX PMF Fun vs Useful: Mean values (3.45) vs (4.09)

The analysis of the survey results was conducted. It shows that students have generally found problem domain topics interesting, but less useful. The results for solution domain topics are opposite. The least fun topics were: UX –Product Market Fit canvas (3.45), HTA

(3.81), Prototyping (3.91) and Interaction design patterns (3.91), but Design patterns were graded as the most important (4.55), with Mockup tools following (4.45). Importance of Cognitive concepts and interaction design principles is graded (4.09) and UX – Product Market Fit canvas (4.09).

When selecting the most important topics 80% of students selected Interaction Design Patterns, followed by Error prevention (60%) and Iterative development (50%), User involvement (40%), Responsive design (40%) and Minimalism (20%). Interesting result is that no one selected Visual design and User help.

In assessing the level of their competences, we have offered a scale corresponding to simplified version of the revised Bloom taxonomy [8]. Students evaluated their competences after the course completion on the higher levels: at the level of application (half of the students) and at the level of evaluation (other half).

4 Product-market Fit (PMF) concept adapted for teaching HCI

Interaction design is part of product development and teaching HCI has a lot in common with teaching product development, especially in the world of digital products, smartphone applications and web portals. Efficient and simple interaction make these product successful on the market.

One of the most important goals of product development is to achieve product-market fit, meaning being in a good market with a product that can satisfy that market. Product Market Fit pyramid [9] is actionable framework for product development which decompose the process into five components: (1) define target customers, (2) identify underserved needs, (3) build value proposition, (4) specify feature set, (5) do UX design.

Product development is a team activity, usually with many participants. One of the challenges in such process is to ensure that all participants understand the objectives and the business model in the same way. Nowadays, characterized by fast market changes, a large number of innovations, demanding users and high competition, visual tools are increasingly used in order to represent the key aspects of product development process in a simple, concise and clear manner.

Product Market Fit canvas is a tool that supports product-market fit pyramid. Around 150 Lean product development specialists participated in the development of this tool [9]. The most important parts of the Business Model Canvas [2] are the relationship between the Value Proposition (what we are building) and the Customer Segment (what users need). These two components of the business model are so important that can be isolated as a separate canvas with its own name: "Product/Market Fit." [10]

The left side of the PMF canvas is dedicated to the problem domain, while the right side is designated to the solution domain. Each field in the right column is related to a field in the same row in the left column. The version adapted for describing the software application is presented in Fig. 3. Customer is User, Product is Application, and we discard the distribution channel.

The central topic of our HCI course is a fully functional application development, where a particular emphasis is placed on interaction design. Students work in teams and work through all steps of product development. The first step to propose an idea that is realistically applicable in the market. After preliminary approval, the idea is subsequently presented in more details and the final project topic is negotiated. During the development, students receive feedback during regular consultations. Final detailed evaluation and grading is also referring to the PMF.

In that sense, Product Market Fit canvas is intended:

1. To prevent misunderstandings in defining scope and goals of assignments
2. To improve control of the process of application development
3. To emphasize the importance of dealing with the problem space
4. To focus on WHY and WHAT in application development

This approach provides a sense of realism. However, it is also has some limitations. There is not much room for introducing a new topic in the tight course schedule, so the PMF is covered within one week. That is why it is important to have a simple and usable tool for a quick introduction to the topic.

If more time could be allocated to the problem domain, further research of user needs would be conducted using field surveys, contextual inquiries, etc.

The primary goal of using Product Market Fit canvas was to engage students with the problem. However, in practice we have noticed the benefit of its application in several phases of students' work. The main guidelines for Product Market Fit canvas use within the HCI course are:

A. Left-hand side of canvas is used to submit the project proposal

Students are proposing the topic of their project at the beginning of the semester. The fields on the left side of the PMF canvas describe users, their needs, require-ments, and expected outcomes, which is quite enough to encourage students to analyze how much their idea may fit the market needs.

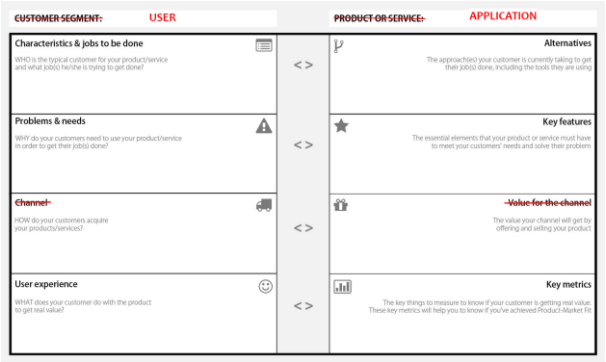


Figure 3. PMF template adapted for describing students' project

B. Full canvas is used for discussion and project approval process

This is followed by an approval process which is activity of detailed presentation of the full canvas to professor, assistants and colleagues a few weeks later. Each field of the canvas is presented in detail. Due the simplicity and the clarity of this form, this type of presentation can foster a very productive and meaningful discussion. This activity is already included in the grading scheme.

C. Canvas is used during the consultancies and for the final project evaluation

Once the topic is approved, the canvas is archived and later used as a reminder of the objectives during the student consultation hours. Finally, canvas is also used in the evaluation of the final project and to assessing how well the final results have met the goals set at the beginning.

Documenting achieved level of engagement is not an easy task, but level of project completion was raised after introduction of this approach. Percentage of students completing the project during the first exam term raised to 44% (2018/2019) from 37% (2017/2018) and 38% (2016/2017).

5 Teachers training experience

As a part of the BENEFIT project [3], teachers training "Designing a Course for Stimulating Entrepreneurship in Higher Education" was organized by the University of Sarajevo, where we shared our experiences with the project partners. It was a two-day training with the first day dedicated to didactics and the second day to teaching product development using the PMF. There were 50 participants on each day. Event evaluations were positive with assessment grades in range 4.3 – 4.8.

Traditional education does not correlate well with the development of entrepreneurial skills, and the topic is rarely adequately addressed in the engineering study programs. Entrepreneurship is quite often considered only in the narrow, commercial form, but it could also be social, public and corporate. The latest form is focused on internal innovation, endeavor and strategic renewal [11] what is all important for a modern engineer.

6 Conclusions

The results presented indicate that stimulating the entrepreneurial skills is an appropriate approach for the HCI course. Product Market Fit canvas is a tool we used to help students bridge the gap between "WHY" and "HOW," between the problem domain and the solution domain. The survey results provided the evidence that students recognized importance and benefits of this tool. The improvement in number of timely completion and submission of the students' projects is backing up this conclusion.

This topic was included in the teachers training within the framework of the BENEFIT project, with relevant didactic topics included. The participants' feedback was positive and indicating broader interest in introducing activities that will stimulate entrepreneurial attitude in engineering students.

Acknowledgements

"This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP".

References

[1] J. Gothelf, *Lean UX: Applying Lean Principles to Improve User Experience*, J. Seiden Ed., O'Reilly Media; 1st edition, 2013.

[2] A. Osterwalder, Y. Pigneur, G. Bernarda and A. Smith, *Value Proposition Design*, John Wiley & Sons, 2015.

[3] ERASMUS+ CBHE Project "*Boosting the telecommunications engineer profile to meet modern society and industry needs – BENEFIT*", [online] Available https://www.project-benefit.eu/, [accessed August 2020].

[4] E-platform of the project BENEFIT, *Human Computer Interaction Course (MSc)*, [online] Available https://www.project-benefit.eu/eplatform/?courses=72#view,[accessed August 2020].

[5] F. Bellotti, R. Berta, A. De Gloria, E. Lavagnino, F. Dagnino, M. Ott, and I.S. Mayer, "Designing a course for stimulating entrepreneurship in higher education through serious games," *Procedia Computer Science*, 15, 2012, pp 174-186

[6] M. Hoppe, M. Westerberg and E. Leffler, "Educational approaches to entrepreneurship in higher education," *Education+ Training*, 2017.

[7] D.D. Stevens, and A.J. Levi, *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*, Stylus Publishing, LLC. 2013.

[8] D.R Krathwohl, "A revision of Bloom's taxonomy: An overview", *Theory into Practice*, Vol. 41, Number 4, 2002, pp. 212-218

[9] D. Olsen, *The Lean Product Playbook: How to Innovate with Minimum Viable Products and Rapid Customer Feedback*, 1st Edition , John Wiley & Sons, 2015.

[10] S. Blank, "The Business Model Canvas Gets Even Better - Value Proposition Design," Steve Blank Blog, 2014, [online] Available https://steveblank.com/2014/10/24/17577/ [accessed August 2020].

[11] D.F. Kuratko, "Corporate Entrepreneurship & Innovation," in *Handbook of Entrepreneurship*, G. Ahmetoglu, T. Chamorro-Premuzic, B. Klinger and T. Karcisky eds. John Wiley & Sons, 2017.

An Overview of Open Source Environments for 5G Internet of Things Experimentation

Srdjan Sobot, Nikola Gavric, Zivko Bojovic, Dejan Vukobratovic
University of Novi Sad,
Faculty of Technical Sciences,
Trg Dositeja Obradovica 6, Novi Sad, Serbia
Emails: {srdjansobot, gavric.de1.2019, zivko, dejanv}@uns.ac.rs

***Abstract*—Recent years brought proliferation of solutions that enable researchers to set up a small-scale testbed environment for 3GPP-compliant cellular technologies. Using inexpensive software-defined radio (SDR) modules and general-purpose processing (GPP) computers, the main elements of the 3GPP radio access network (RAN) and evolved packet core (EPC) architecture can be deployed in the lab. In this paper, we review the most popular open-source tools for 3GPP network experimentation. We focus on a possible experimentation and testbed development for cellular Internet of Things (IoT)/Cloud integration. Besides a general overview, we provide some practical considerations based on our own experience with setting up and exploiting open-source 3GPP testbed environment, and point the reader to appropriate sources for detailed step-by-step descriptions and respective documentation.**

***Index Terms*—3GPP 4G LTE, 3GPP 5G NR, 3GPP NB-IoT, OpenAirInterface, srsLTE**

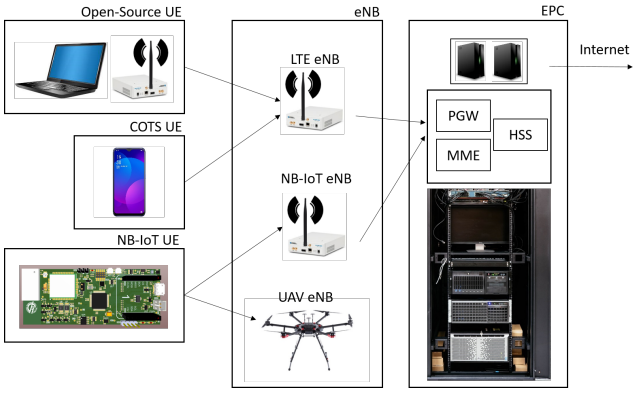


Fig. 1. 3GPP architecture implemented using open-source SDR/GPP nodes.

I. Introduction

Deploying 3GPP-compliant network architecture with main Radio Access Network (RAN) and Evolved Packet Core (EPC) elements in the lab environment using standard Software Defined Radios (SDRs) and General Purpose Processing (GPP) servers is now a reality. Even though softwarization of baseband signal processing parts is not a new concept, and the projects that target implementation of the software-based networks date back to 2G Global System for Mobile (GSM) [1] and 3G Universal Mobile Terrestrial Systems (UMTS) technologies [2], these trends are accelerating due to ever-growing accessibility, affordability and compactness of SDRs and GPP units. For 4G Long Term Evolution (LTE) technology, two open-source 3GPP network solutions stand out as the most popular among researchers in the field: OpenAirInterface [3] and srsLTE [4]. Both feature open-source implementation of User Equipment (UE), base stations (eNode - eNB) and the main elements of Evolved Packet Core (EPC), developed to a varying level of standard features (mostly up to Release 10). OAI continued with development of 5G New Radio (NR) software and is currently the only solution that provides basic 5G NR features as an open-source implementation [5].

In this paper, we review and share our first-hand experience in establishing an educational and research-oriented lab experimentation environment based on two major open-source 3GPP software: OAI and srsLTE. In addition to basic network setup, we discuss extensions of open-source platforms to three additional scenarios:

- Cellular IoT scenario by integrating 3GPP NarrowBand IoT (NB-IoT) into the existing sofware architecture;
- UAV-assisted cellular IoT scenario using drone-mounted SDR/GPP equipment;
- EPC core extension to support Cloud services and network slicing and virtualization support.

Both in LTE as well as in NB-IoT setup, besides open-source UE running on GPP-based lap-top or similar device, one can attach commercial off-the-shelf (COTS) mobile phones and NB-IoT devices to the OAI/srsLTE-based eNB.

The paper is divided into the three main sections (Sections 2, 3 and 4) covering open-source software for Evolved Universal Terrestrial Radio Access (E-UTRAN), EPC and Cloud/SDN domain, respectively. E-UTRAN consists of base stations or eNodeB's (eNBs), including Unmanned Aerial Vehicle (UAV) mounted eNBs. Each eNB serves a population of User Equipments (UEs). EPC contains main data and control entities such as Serving and Packet Gateways (SGW/PGW), Mobility Management Entity (MME) and Home Subscriber Server (HSS). Overall 3GPP architecture is illustrated in Figure 1 represented by usual SDR/GPP configurations used in the lab. In each of the three sections, we first describe general information about open-source implementations in a given domain, and then provide some practical considerations and suggestions based on our own experience.

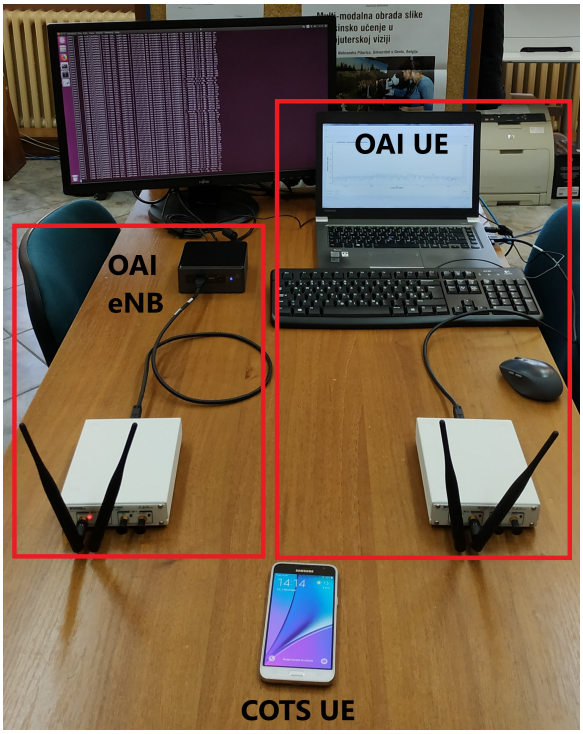


Fig. 2. OAI setup where OAI and COTS UE are connected to OAI eNB.

II. Open Source 3GPP Radio Access Network

A. 3GPP LTE/NR RAN Software

General Description: OAI is an open and flexible 4G/5G experimentation platform created by the Mobile Communications Department at EURECOM [6]. OAI offers an open-source implementation of the LTE system spanning the full protocol stack of 3GPP standard both in E-UTRAN and EPC, and lately, it is oriented towards 5G NR technology. Additionally, OAI has launched a new project to develop a software-defined radio-based NB-IoT eNB. OAI is written in standard C for several real-time Linux variants and released as free software under the OAI License Model.

srsLTE is a highly modular LTE library with minimal inter-module or external dependencies, used to build LTE-based applications such as a complete eNodeB or UE [7]. srsLTE library is written in C and available under both commercial and open-source licenses. srsUE & srsENB are written in C++ and build upon the srsLTE library.

In both platforms, the transceiver functionality is realized via an SDR front-end connected to a host computer for processing and both platforms provide interfaces to the Universal Hardware Driver (UHD), giving support to the Ettus USRP family of devices.

Practical Considerations: OAI platform is divided into two separate projects: OAI eNB&UE [18] and OAI EPC [19]. Both projects provide several tutorials that allow users to set up the software in real-time/emulation or simulation mode. Before starting with instructions on how to run the software, a few steps are required to successfully install OpenAirInterface:

- Hardware Requirements - OAI requires Intel architecture-based PC, also it is useful to check out requirements such as supported RF, UE/Sim Card Requirements, etc [17].
- Kernel Requirements - it is recommended to use Linux, i.e., Ubuntu 16.04 with low-latency kernel 4.4.0 or higher, furthermore it is recommended to disable C-states from BIOS and CPU frequency scaling. OAI should run also on other Linux distributions, but the build scripts need to be adopted [20].
- Cloning source code from their Git repository and reading instruction and tutorials from the Tutorials page [21].

From our experience, the aforementioned steps are the most important prerequisite to successfully build OAI, especially, the kernel requirements need to be met. Besides, the best way to get a global picture of using OAI is to follow the basic ”first-time-user” tutorial, i.e., connecting OAI eNB (USRP B210) with COTS UE [22]. By following this tutorial, a user will go through the entire procedure of building a complete LTE system, which is divided into three stages: i) installing and configuring OAI eNB, EPC, and HSS on a single or different hosts; ii) configuring UE (e.g., smartphone service mode, USIM card) and register the user on the HSS database; and iii) configuring networking settings between eNB and UE.

On the other hand, srsLTE is simpler to install and does not require any special configuration of PC nor Linux distribution. They provide a list of required libraries, and in a few commands in the terminal whole srsLTE system can be installed. Also, configuring eNB, EPC or UE are very straightforward. Useful documentation can be found at [23].

Figure 2 shows one possible use case where we connected OAI UE (USRP B210) and commercial smartphone to OAI eNB. During the connection procedure, MME assigned IP addresses to both devices. In the case of OAI UE, the host PC had an Internet connection through a so-called OIP interface. With the smartphone, we were able to browse the Internet from another room and see how the Channel Quality Indicator (CQI) gradually decreased towards the ”cell-edge”. Also, the srsENB application runs in the console and offers rich set of metrics (rnti, cqi, snr, bitrate, etc.) that are provided on a per-UE basis for the downlink and uplink respectively.

B. 3GPP NB-IoT Software

General Description: OAI and srsLTE are in the process of extension to support the NB-IoT radio access technology. All downlink/uplink channels and signals defined in the standard are developed taking into account the NB-IoT requirements (repetitions, 15 kHz/3.75 kHz sub-carrier spacing, etc.).

Practical Considerations: Currently, with OAI there is a possibility to connect COTS UE to the NB-IoT eNB and ping an external server via Nokia EPC software. Furthermore, AT commands can be used to send a UDP packet to the LTEBox. Even though the uplink data transmission is achieved and the server can successfully receive the UDP packet, transmission of a data packet longer than 4 bytes may not be successfully decoded at the PHY layer of eNB. This is a known issue, and

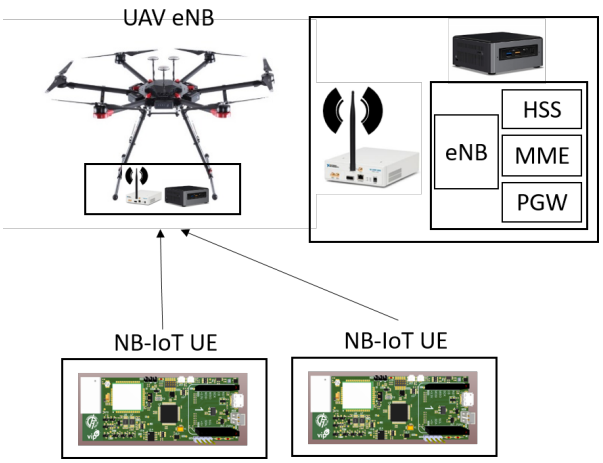


Fig. 3. srsLTE/OAI-based UAV eNB setup [13].

will probably be resolved soon. A comprehensive tutorial can be found at [9].

srsLTE provides two example applications that act as PHY-only NB-IoT eNodeB and UE. The *npdsch_enodeb* example acts as an eNodeB by accepting data traffic on a TCP socket and transmitting it using the NB-IoT downlink physical channels. The *npdsch_ue* example acts as the counterpart and provides a PHY-only UE that receives and decodes the DL transmission of the eNB [10]. Guides for specific srsLTE use-cases can be found on their documentation page under the Application notes section [23].

C. Extensions to Mobile UAV-based eNB

General Description: Unmanned aerial vehicles (UAV) have been used with SDR devices to build aerial base stations and relays to extend an existing terrestrial infrastructure boosting the LTE connectivity to the users [11], or to compensate for temporary cell overload or outage [12]. Both srsLTE and OAI can be adopted to provide cellular connectivity to ground users [13]. Fig. 3 shows one possible use case where UAV base station is utilized to extend terrestrial infrastructure and provide connectivity to unreachable NB-IoT devices.

Practical Considerations: We are in the process of integrating UAV-based eNB using DJI Matrice 600 Pro drone. This hexacopter has total weight of 9.5kg and may carry additional load of 6kg, making it suitable for the load containing Intel NUC Mini PC, USRP B210, and appropriate battery-pack to provide power supply for Mini-PC (and USB-attached USRP). With standard 6-pack of batteries, and maximum load of 6kg, Matrice 600 Pro is able to provide hovering time of 16 mins, which gradually increases by reducing load to the maximum of 32 mins for an unloaded drone. In our setup, backhaul connection to external networks is achieved using 4G LTE USB modem attached to the Mini PC. In such a configuration, UAV eNB can serve as a relay with reliable LTE-based link towards infrastructure, and enhanced direct links to LTE or NB-IoT UEs on the ground that are set in challenging radio conditions to connect to 4G network directly.

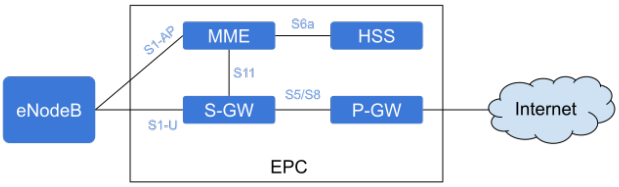


Fig. 4. srsLTE/OAI EPC overall architecture [23].

III. Open Source 3GPP Evolved Packet Core

General Description: OAI and srsLTE offer the implementation of key components of LTE core network (EPC), namely, Home Subscriber Server (HSS), Mobility Management Entity (MME), Service Gateway (S-GW), and Packet Data Network Gateway (P-GW) [6]–[8]. As can be seen in Figure 4, OAI and srsLTE EPC share the same overall architecture which is defined by 3GPP standardization.

In the case of OAI, each EPC component can be installed on a different host, however, they are usually installed on a single host due to simplicity. S-GW and P-GW are merged, so there are no S5/S8 interfaces between these two entities. Also, it is not recommended to deploy EPC on the same host as eNB, because eNB consumes a lot of PC resources. The OAI EPC software has been tested on Ubuntu 14.04x64 and requires a kernel version greater or equal to 4.7 [19].

srsEPC application runs as a single binary and provides all key EPC components, and can be installed on any PC with a GNU/Linux based operating system [23].

In some scenarios, third-party EPC software may be needed. For example, in OAI-based NB-IoT scenario, software-based EPC called LTEBox developed by Nokia Bell Labs is used, and may be obtained from Nokia under specific terms for academic use [9].

Practical Considerations: From our experience, OAI EPC is more complicated to install and configure. Also, we have faced an issue a few times, EPC crashed during establishing the connection to a commercial smartphone. But generally speaking, the OAI EPC is stable. srsLTE EPC is very simple to install and configure and there is a simple CSV-based user database. A very helpful tutorial on how to program a sim card to register user to a database can be found at [24], [25].

IV. Flexible and Programmable Network Services

The advent of SDR-based software platforms, such as OAI and srsLTE, has enabled the establishment of a complete mobile network with adequate performance and compatibility with commercial equipment. The availability of these platforms has enabled part of the industry and the academic community to work on the development of a new 5G networking solutions and to conduct tests under approximately realistic conditions.

The test results show that future solutions within 5G networks will be diverse, primarily due to cloud infrastructure use. The management of a wide range of services in such a virtual infrastructure depends primarily on the existence

of highly scalable orchestrators. However, the support that orchestration software brings in the network domain is still under massive development.

The architecture of the mobile network with network slicing capabilities requires the usage of several software platforms to make a rough comparison with traditional network domains (RAN, EPC, and Management and Orchestration component - MANO). Regarding the RAN and EPC domain, the most relevant solutions are OAI and srsLTE with basic and extended EPC functionalities (mobility management entity - MME, home subscriber server - HSS, and packet and serving gateways). MANO is an ETSI-hosted project, whose goal is to develop an Open Source MANO software stack aligned with ETSI NFV. One of the leading solutions for implementing a fully-fledged mobile network orchestrator is Open Source MANO (OSM). Its key components are the NFV Orchestrator and VNF Manager (NFV MANO) and service Orchestrator, which should enable specific NFV services. For the implementation of an appropriate testbed environment, it is necessary to provide efficient orchestration of resources. One part of the academic community uses OpenStack as a virtual infrastructure manager (while Amazon AWS EC2 and Kubernetes are also eligible), whose role is to manage the local Network Function Virtualization Infrastructure (NFVI) deployment [14].

Mosaic5G is a project created as an ambition of individual representatives of industry and academia to encourage software development and implement 4G/5G service platforms. In fact, it is an ecosystem of open source platforms and use cases that relies on SDN, NFV, MEC and AI technology in the implementation of 4G/5G systems. Mosaic5G consists of five software components [15]:

- JOX – an event-driven service orchestrator
- Store – a collection of SDKs, applications and datasets
- LL-MEC – SDN controller for core network
- FlexRAN RAN – SDN controller for RAN
- OAI RAN and OAI CN –FlexRAN and LL-MEC compatible 3GPP-complaint implementation of LTA/LTE-A features

FlexRAN is a programmable platform for SDN-RAN which allows deployment of network applications over the controller, dynamic RAN control and separation of control and data planes. It is an extension to a modified version of the OAI and consists of a controller and agents. It uses FlexRANAgentAPI with 5 API call types for control and action functions. Even though FlexRAN causes a slight increase in memory and CPU usage, FlexRAN does not affect QoE compared to OAI [16].

V. Conclusion

This paper presented a review of popular open access tools for 4G/5G/NB-IoT experimentation including additional cloud-based components for implementation of 5G services. The paper provides general description and personal experience with deployment and usage of given tools. The described system provides excellent support for teaching and research of 3GPP mobile cellular technologies in the lab, and is particularly suitable for project-based learning and demonstrations.

Acknowledgment

This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP.

References

[1] http://openbts.org/
[2] https://github.com/RangeNetworks/OpenBTS-UMTS
[3] https://www.openairinterface.org/
[4] https://www.srslte.com/
[5] https://gitlab.eurecom.fr/oai/openairinterface5g
[6] Nikaein, N., Marina, M.K., Manickam, S., Dawson, A., Knopp, R. and Bonnet, C., 2014. OpenAirInterface: A flexible platform for 5G research. ACM SIGCOMM Computer Communication Review, 44(5), pp.33-38.
[7] Gomez-Miguelez, I., Garcia-Saavedra, A., Sutton, P.D., Serrano, P., Cano, C. and Leith, D.J., 2016, October. srsLTE: an open-source platform for LTE evolution and experimentation. In Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization (pp. 25-32).
[8] Gringoli, F., Patras, P., Donato, C., Serrano, P. and Grunenberger, Y., 2018. Performance assessment of open software platforms for 5G prototyping. IEEE Wireless Communications, 25(5), pp.10-15.
[9] Chen, C.C., Cheng, R.G., Ho, C.Y., Kanj, M., Mongazon-Cazavet, B. and Nikaein, N., 2020. Prototyping of Open Source NB-IoT Network. arXiv preprint arXiv:2006.02729.
[10] Puschmann, A., Sutton, P. and Gomez, I., 2017. Implementing nb-iot in software-experiences using the srslte library. arXiv preprint arXiv:1705.03529.
[11] Gangula, R., Esrafilian, O., Gesbert, D., Roux, C., Kaltenberger, F. and Knopp, R., 2018, June. Flying rebots: First results on an autonomous UAV-based LTE relay using open airinterface. In 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) (pp. 1-5). IEEE.
[12] D'Alterio, F., Ferranti, L., Bonati, L., Cuomo, F. and Melodia, T., 2019, December. Quality Aware Aerial-to-Ground 5G Cells through Open-Source Software. In 2019 IEEE Global Communications Conference (GLOBECOM) (pp. 1-6). IEEE.
[13] Powell, K., Sabri, A., Brennan, D., Marojevic, V., Barts, R.M., Panicker, A., Ozdemir, O. and Guvenc, I., 2020. Software Radios for Unmanned Aerial Systems. arXiv preprint arXiv:2004.01987.
[14] Garcia-Aviles, G., Gramaglia, M., Serrano, P., Gringoli, F., Fuente-Pascual, S. and Pavon, I.L., 2020. Experimenting with open source tools to deploy a multi-service and multi-slice mobile network. Computer Communications, 150, pp.1-12.
[15] Nikaein, N., Chang, C.Y. and Alexandris, K., 2018. Mosaic5G: Agile and flexible service platforms for 5G research. ACM SIGCOMM Computer Communication Review, 48(3), pp.29-34.
[16] Foukas, X., Nikaein, N., Kassem, M.M., Marina, M.K. and Kontovasilis, K., 2016, December. FlexRAN: A flexible and programmable platform for software-defined radio access networks. In Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies (pp. 427-441).
[17] https://gitlab.eurecom.fr/oai/openairinterface5g/-/wikis/OpenAirSystemRequirements
[18] https://gitlab.eurecom.fr/oai/openairinterface5g
[19] https://gitlab.eurecom.fr/oai/openair-cn
[20] https://gitlab.eurecom.fr/oai/openairinterface5g/-/wikis/OpenAirKernelMainSetup
[21] https://gitlab.eurecom.fr/oai/openairinterface5g/-/wikis/OpenAirUsage
[22] https://gitlab.eurecom.fr/oai/openairinterface5g/-/wikis/HowToConnectCOTSUEwithOAIeNBNew
[23] https://docs.srslte.com/en/latest/index.html
[24] https://gitlab.eurecom.fr/oai/openairinterface5g/-/wikis/SimCardConfig
[25] https://philclaridge.com/how-to-program-a-sysmocom-sjs1-sim-for-use-with-oai/

Multi-subject Project-Based Learning: Two Semester Pilot Study

Goran B. Marković[1], Jelena D. Ćertić[1], Marija Božić[2], Mladen Koprivica[1]

[1]*School of Electrical Engineering, University of Belgrade, Serbia*
[2]*Faculty of Engineering, Science and Technology, Bernardo O'Higgins University, Santiago, Chile*
E-mail: certic@etf.bg.ac.rs

Abstract. *In this paper, we present the pilot study of two different implementations of the two-subject project-based learning scheme. Proposed two-subject schemes are aimed to emphasize the importance of the classical digital signal processing DSP courses to the broader filed of the Telecommunications by joint project assignments between a DSP course and the fundamentals of telecommunications course. We explain the motivation for introducing this type of learning methodology, describe the implemented study, and discuss some preliminary findings.*

1 Introduction

The results of the previous research project implemented at the School of Electrical Engineering, University of Belgrade, indicated that the set of generic and professional competences can be significantly fostered by student-centered learning strategies [1]. The project was realized as an internship-like course with a major goal of preparing students for their future real-life work surroundings [2]. Because of that, the course was designed for last year's students of bachelor's degree studies. The learning methodology implemented was based on project and problem-based learning. The internship-like course was chosen, among other reasons, because at that time the internship was the mandatory part of the studying, and introduced internship-like course provide additional internship position options for students.

In the meantime, as a part of the consortium of the ERASMUS+ project BENEFIT-Boosting the telecommunications engineer profile to meet modern society and industry needs [3], we took part in defining guidance for modernization of the study programs in the area of telecommunications and introducing new teaching and training models. One goal of the study program modernisation was a better connection between different knowledge areas relevant to the telecommunications and information technology ICT studies. Encouraged by the very good results of our internship-like course 4-year project [2], [1], we decided to introduce forms of active learning as early as possible into the curriculum of the study program Electrical Engineering and Computing, module Telecommunications and Information technology. Based on different factors: the number of students enrolled, previous results and experiences, limitations in the number of laboratory working places, and equipment, we choose the problem-based learning methodology as appropriate for our pilot study. We designed two different two-subject project-based schemes, one with a joint project between subjects Signal Processing 1 and Telecommunications 2, and the other with a joint project between subjects Signal Processing 2 and Telecommunications 3.

The rest of the paper is organized as follows. In section 2 brief description of the study module Telecommunications and Information Technology is presented and the motivation for introducing a two-subject problem-based learning methodology is explained. Section 3 contains details of the introduced pilot projects. Section 4 concludes the paper.

2 Structure of the study program and study module

Bachelor's degree studies at the School of Electrical Engineer, University of Belgrade last four years (240 ECTS). There are two study programs available:

1. Electrical Engineering and Computing, and
2. Software Engineering.

The new accreditation cycle starts in the school year 2020/2021. The first year is common for all students enrolled in Electrical Engineering and Computing. In the second year, students select one of the six study modules:

1. Electronics and Digital Systems,
2. Power Engineering,
3. Computer Engineering and Informatics,
4. Signals and Systems,
5. Telecommunications and Information Technology,
6. Physical Electronics.

The Telecommunications and Information Technology module consists of the core telecommunications and IT subjects, and elective subjects of related fields. Students can choose one of the three paths of the module (submodules) with slightly different sets of mandatory and elective courses.

There is a strong line of digital signal processing DSP subjects through the whole Telecommunications and Information Technology curriculum. However, the role and importance of digital signal processing in the field of telecommunications are not always clear to the students. Because of that, additional steps were made to incorporate the DSP subjects into the whole curriculum in a better way. The proposed two-subject project-based learning schemes are the results of those efforts. The proposed two-subject project-based learning model is the simplest variant of the multi-course project-based learning [4], [5], [6].

The expected benefits of this approach are:

- Benefits of the project-based learning including learning by doing, real-world problems, collaboration, and team-work, creating and presenting an end product,
- Students are expected to perceive certain topics from different points of view through a multidisciplinary approach,
- Students are encouraged to see a broader picture of the area of their study program or module instead of learning in "subject by subject", or "topic by topic" way.

3 The proposed project-based learning schemes

In this study, two schemes related to problem-based learning were developed. The first scheme (6th-semester two-subject scheme) is, actually, a modification of the project-based learning that has been offered to students through the subject Signal Processing 2 for more than a decade.

The Signal Processing 2 project assignments are mostly based on the engineering and scientific papers from conferences or journals that can be understood by the students of the advanced DSP course. Long-lasting Signal Processing Magazine Tips & Tricks section [7] and the online www.dsprelated.com portal [8] are examples of the collections of appropriate manuscripts. For that reason, this scheme can be considered as a form of research-related learning.

The proposed modification of the current project-based learning scheme incorporates one more subject, Telecommunications 3. By this modification, the project assignments that can bring a broader perspective to real-life challenges in DSP applications in telecommunications are introduced. Besides, there are two more teachers, with their expertise, ideas, and point of view. The proposed modification is expected to help students to perceive a broader picture of the challenges of their future careers. The details of the proposed 6th-semester scheme are presented in subsection 3.1.

The second scheme (5th-semester two-subject scheme) is developed based on the good results obtained by the first one, i.e. 6th-semester case. We recognized that many of our students haven't got the opportunity to practice at least some components of project-based learning until the 6th semester of their studies. For example, many of them told us that during the presentation they experienced nervousness because they hadn't done anything like that before. For that reason, we decided to try to introduce a two-subject project-based learning scheme in the 5th semester. Due to a very large number of students, the current version of the proposed scheme looks more like a standard semester assignment, but we hope to further develop it in the following semester. One possible way of improvement is to develop one large project-based model that lasts two semesters and overlaps with four different subjects. The details of the current implementation of the proposed 5th- semester scheme are given in section 3.2.

3.1 The two-subject project-based learning in the 6th semester

The first pilot scheme of this study introducing project-based learning was realized for the students of their third year of study (out of four), 6th semester, during the spring semester in 2019. The project assignment was related to two subjects, Telecommunications 3 (mandatory subject, lecturers Goran Marković, Kristina Josifović) and Signal Processing 2 (elective subject, lecturers Jelena Ćertić, Miloš Bjelić). The two different project assignments were offered as a two-subject project:

1. The project built on the 5G solution proposed in the conference paper [9],
2. The project related to the simulation of the GSM system with three-step decimation in the receiver.

The theoretical aspects of GSM and OFDM technology are standard topics of the subject Telecommunications 3. Advances in filter design and implementation, including the realization of the Software Defined Radio (SDR) receiver are topics of the subject Signal Processing 2.

The semester project, based on the principles of project-based learning is the obligatory part of the subject Signal processing 2, however, not all the project themes are related to topics covered by subject Telecommunications 3. For that reason, the two-subject project was offered as an option for the students. The project assignment was realized in small teams of students (up to four students). Students had access to laboratory equipment and resources (including access to a selection of scientific and engineering journals and magazines that are available as Kobson service [10]). Consultancies with teachers were organized during laboratory exercises classes or on the students' request.

As the results of their work, each group of students had to prepare a short written project report (including text, results, and code examples of the realization) and 10-15 minutes presentation.

In this pilot study, there were five groups of students that took the multi-subject project. Four groups selected the theme related to 5G, and one group selected the theme related to GSM. At the end of the semester, the project assignments were defended before the subjects' teachers and all students enrolled in the subject Signal processing 2. The examples of the concluding slides of the two sample presentations are given in Fig. 1. All students had the opportunity to ask questions to other teams. Presentations were graded by students from other teams, in the questionnaire defined by teachers. Those grades were scaled as part of the final grade. The four projects were finished and defended in time, and the overall scores for the presentations are given in Table 1 (grade range 1 to 5).

The common question related to project-based learning is how to grade the students' work. In the case

of the subject Signal Processing 2, the problem is minimized due to some favorable circumstances. First of all, the subject Signal Processing 2 is elective. It means that students who choose the subject are highly motivated to the area and for the proposed methodology of the subject. The subject organization is explained to them during the first week of the semester, and they have enough information and time to choose their elective subjects. Second, the whole group is not too large, they know each other and there is a form of self-control by the group. And, finally, due to extensive discussions and consultations, each student is forced to take part in all steps of the project realization. The final grade for the subject Signal Processing 2 is based on the scores for two theoretical tests (16% each), oral exam (30%), and semester project (38%). The grade of the project itself consists of the grade for the developed code, the grade for the written report, and the grade for the presentation. The grades given by other students (Table 1) were included as part of the presentation grade. Our impression was that the presentation grades given by the students are highly correlated with the grades given by the teachers. The oral exam is organized as a part of the project defense.

For subject Telecommunications 3, this type of project was not mandatory, i.e. the other type of projects were offered to students, and the project score can improve the overall grade by 10%.

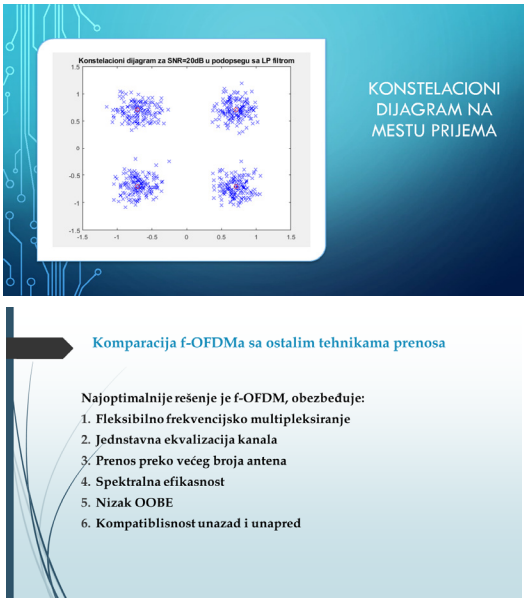


Fig. 1. Examples of the concluding slides of the students' presentations.

Our general impression was that the students enjoyed the challenge of realizing the task that is related to 5G topics. We asked them typical exam questions, but the answers showed a deeper understanding of the matter comparing to an average group. It was interesting that all four groups of students organized their defences in a specific way, i.e. there were no similarities between their performances. The students in the audience were very active during the defences of other teams, they asked questions, and they were free to give their opinions and comments.

Table 1. Team scores.

Team	The goal of the project is clearly stated	Presentation of the main idea of the solution	Developed solution	Presentation of the results	Overall impression
Group 1	4.57	4.79	4.73	4.71	4.86
Group 2	4.73	4.36	4.45	4.55	4.55
Group 3	5.00	4.62	4.60	4.70	4.87
Group 4	4.80	4.87	4.27	4.30	4.60

3.2 The two-subject project-based learning in the 5th semester

The second project was realized for the students of their third year of study (out of four), 5th semester, during the autumn semester in 2019. The project assignment was related to mandatory subjects, Telecommunications 2 (lecturers Goran Marković, Kristina Josifović) and Signal Processing 1 (lecturers Jelena Ćertić, Miloš Bjelić). For subject Signal processing 1, the project could be taken as a replacement for the (obligatory) practical semestral work related to the implementation of digital signal processing algorithms in MATLAB/Octave. For subject Telecommunications 2, this type of project was not mandatory, i.e. the other type of projects were offered to students, and the project score can improve the overall grade by 10%.

For the classical Signal Processing 1 practical work, the well-specified assignment is given, and it looks like the practical part of the exam. Students have assigned time slots in the laboratory, and they write the code, solution to the assignment in a specified time. With this proposed model, we introduced the project assignments that are not well-defined, i.e. many parameters have to be assumed, the structure of the code is left to the students. It is more time consuming for students, but they can do the project anytime during the semester. They can ask for the time slots in the laboratory if they need access to some specific equipment or computer.

For this pilot study, the topic of the project was the simulation of the telecommunication system based on QPSK modulation. The simulation model was fully explained and partly developed by the teachers. The part that was left for the student was, actually, software implementation of the receiver of the system. The students also had the option to implement their solution by using software-defined radio (SDR) platform (ADALM Pluto – donated by Analog Devices) [11], and two project groups opted for that. The project was chosen by about 20% of the students. The project assignment was realized in pairs of students.

As the results of their work, each group of students had to prepare a short presentation, explain the code solution, explain the values chosen for the unspecified parameters, and discuss obtained results by comparing them to theoretical values. As an example, the graph comparing theoretical results with results obtained by the simulation that was presented by one of the project teams is shown in Fig. 2.

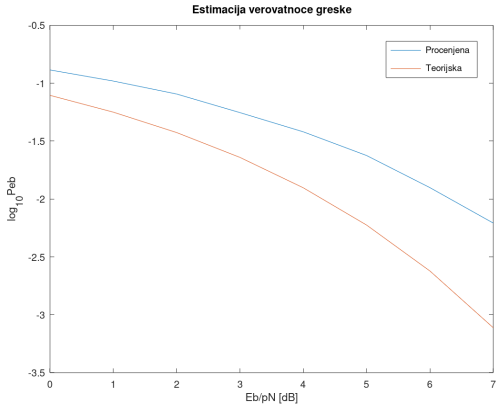


Fig. 2. Example of the presentation of the results.

Our general impression is that the two-subject project helped students to better understand the entire hierarchy of the telecommunications system and the role of the basic signal processing in the telecommunications system. Although the implemented scheme is not a full project-based learning model, based on our previous research [1] we expect that competences such as systems thinking, communication: asking questions, the ability to apply knowledge in practice, presentation and solving ill-structured problems can be improved. We also expect to assess professional competences related to the subjects included in this study. We hope that it will be possible to repeat this experiment during the next semester and perform tests that would provide conclusions about the development of generic and professional competences by this learning scheme.

4 Conclusions

The proposed two-subject project-based learning model has certain challenges that we noticed and that should be mentioned:

- The need for excellent communication and mutual understanding of teachers of different subjects,
- The choice of appropriate project assignment,
- The decision about the appropriate weight of the project assignment when calculating the overall grade of each of the subjects,
- The increased need for consultation hours,
- The support from all teachers involved in the curriculum, especially when comes to a correlation of activities that are time-consuming for students (other semester projects, laboratory exercises, and similar).

At this point, we believe that the careful preparation of a two-subject project-based scheme and choice of an interesting and relevant project topic result in increased students' motivation to do their best. Students' motivation is an important factor for achieving as many benefits of this approach as possible.

The colleagues from the Telecommunications department supported us in the realization of the presented project. As a part of the improvement and modernization of the module Telecommunications and Information Technology and the preparation for the new accreditation cycle, we had many discussions about the subjects' contents as well as the innovative teaching methodologies.

From our pilot study, we can conclude that the multi-subject and multi-semester project-based learning schemes are a suitable model for engineering studies.

Acknowledgements

This work was carried out within the project BENEFIT - Boosting the telecommunications engineer profile to meet modern society and industry needs - 2017-2021 - cofounded by the ERASMUS+ KA2 program – grant agreement n. 585716-EPP-1-2017-1-AT-EPPKA2-CBHE-JP.

References

[1] M. Božić, J. D. Ćertić, M. Vukelić, and S. Čizmić, "New instructional approach for fostering generic and professional competences: Case study of the project and problem based learning engineering practice course," *International Journal of Engineering Education*, vol. 34, no. 5, pp. 1581-1591, 2018.

[2] M. Božić, S. Čizmić, D. Šumarac-Pavlović, and M. T. Escalas-Tramullas: Problem-Based Learning in Telecommunications, "Internship-Like Course Bridging the Gap between the Classroom and Industry," *The International Journal of Electrical Engineering & Education*, vol. 51, no. 2, pp. 110–120, 2014.

[3] "Project-Benefit," [Online]. Available: https://www.project-benefit.eu/. [Accessed 01.08.2020].

[4] C. G. Schaller and R. Hadgraft, "Developing student teamwork and communication skills using multi-course project-based learning," *in Proc. AAEE Conf.*, Dec. 2013, pp. 1–10.

[5] B. Alorda, J. Verd, V. Canals, K. Suenaga, and V. Martinez, "Multi-subject Project Based Learning initiative," *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, Marrakech, 2012, pp. 1-7.

[6] L. Lundheim, T. Ekman, B. Gajic, B. B. Larsen and T. Tybell, "Early innovation projects: first experiences from the electronic engineering ladder at NTNU," *Proceedings of the 12th International CDIO Conference, Turku University of Applied Sciences, Turku, Finland,* June 12-16, 2016.

[7] https://signalprocessingsociety.org/publications-resources/ieee-signal-processing-magazine/category/sp-tips-and-tricks [Accessed 01.08.2020].

[8] https://www.dsprelated.com/ [Accessed 01.08.2020].

[9] X. Zhang, M. Jia, L. Chen, J. Ma, and J. Qiu, "Filtered-OFDM - Enabler for Flexible Waveform in the 5th Generation Cellular Networks," *2015 IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015, pp. 1-6.

[10] https://kobson.nb.rs/kobson.82.html [Accessed 01.08.2020].

[11] ADALM-Pluto, [Online]. Available: https://www.analog.com/en/design-center/evaluation-hardware-and-software/evaluation-boards-kits/adalm-pluto.html#eb-overview [Accessed 01.08.2020].

Sodobne učne strategije in tehnologije: lažna obrnjena učilnica s poizvedovalnim učenjem in 3D tiskom

Špela Kosec, Tanja Rozman, Janez Jamšek

Oddelek za fiziko in tehniko, Univerza v Ljubljani, Pedagoška fakulteta, Kardeljeva pl. 16, 1000 Ljubljana
E-pošta: janez.jamsek@pef.uni-lj.si

Contemporary learning methods and technologies: faux-flipped classroom by inquiry based learning and 3D printing

Abstract. *Paper presents the faux-flipped classroom by inquiry-based learning in a simple and straight forward practical example of technology activity day in the 7th primary school grade. It is a distance learning method starter that develops student's independent learning skills and aims to higher taxonomic levels of knowledge. Blended learning method, i.e., flipped learning is executed in a classroom to surpass information-communication issues. E-learning material was prepared along with the working sheets to preserve sufficient understanding of learning material. Inductive active inquiry learning was used to engage students in problem solving. Functional propeller-driven vehicle was used to engage students. For the crucial part, axle brackets, students were directed to use contemporary technology, 3D printing to make model and print it. 3D printing was used intentional to direct students toward production rather than reproduction. New technology proved to be highly motivational and enabled all students to be problem solving successful and generic. By using pre-tests and post-tests we monitored students' progress. The method used proved to engage high order taxonomic levels very efficiently (up to 40 %). Our main aim is to motive and encourage technology teachers, regardless the education level, to use contemporary approaches for teaching/learning.*

1 Uvod

Najbolj znana in najstarejša metoda učenja/poučevanja je tradicionalni pouk. Pouk poteka v učilnicah, kjer imamo prostorske in časovne omejitve. Kot rešitev se je ob začetku 20. stoletja razvila metoda učenja na daljavo – dopisno izobraževanje (1892, W. R. Harper, Univerza v Chicagu). Pri tem so udeleženci študirali preko poštnega dopisovanja. Prvi poskusi izobraževanja na daljavo so se pojavili na polovici 20. stoletja, ko v Veliki Britaniji ustanovijo prvo »odprto univerzo« (Open University). Ta svoje programe izvaja izključno na daljavo, izobraževanje pa temelji na posnetkih predavanj, ki jih predvajata BBCjeva televizija in radio [1]. Z razvojem in širjenjem interneta so učni materiali prešli na splet. Nova tehnika učenja je bila poimenovana e-učenje/e-izobraževanje. E-učenje ponuja udeležencem samostojno učenje in poteka z uporabo elektronskih medijev in IKT. Udeleženec in učitelj sta krajevno in časovno ločena, vendar med njima obstaja komunikacija [1, 2]. Kljub začetni uspešnosti metode e-izobraževanja, se je izkazalo, da predavanja niso dovolj zanimiva in udeležencev ne motivirajo. Prihodnost e-učenja je v kombiniranem učenju (KU) [2].

Enotnega dogovora o definiciji KU še ni. KU določajo tri temeljna izhodišča: kombinacija učnih modalitet, kombinacija učnih metod in kombinacija različnih načinov e-učenja in tradicionalnega učenja [2, 3]. Bonk in Graham sta zapisala delovno definicijo KU [2]: »Sistemi kombiniranega učenja združujejo tradicionalni pouk z elementi e-izobraževanja.«. Poznamo več modelov KU. Najpogosteje uporabljeni po H. Staker so [4]: iz oči v oči (face to face), rotacijski model (rotation model), prilagodljiv model (flex model), spletni laboratorij (web lab), samostojno kombiniranje (self-blend) in spletni voznik (online driver). KU se najhitreje uveljavlja na visokošolskem nivoju izobraževanja, kjer ga je najenostavnejše vpeljati. Nižje kot gremo po vertikali izobraževanja, bolj dovršena mora biti njegova vpeljava, da nam zagotovi uspeh. Trend motivacije učencev/dijakov/študentov je v upadanju zlasti zaradi ustaljenega, tradicionalnega, podajanja znanja na zalogo, ki je, nižje kot gremo po vertikali, bolj široko je dostopno na svetovnem spletu.

V prispevku se osredotočamo na KU in model obrnjenega učenja na osnovnošolski ravni ciljno za tehniški dan v 7. razredu. Uporabljena je model lažne obrnjene učilnice, ki za razliko od obrnjene učilnice izvaja fazo samostojenega učenja z informacijsko-komunikacijsko tehnologijo (IKT) v učilnici in ne doma. Zaradi visoke stopnje možnosti transferja na drugo ciljno skupino in nujnosti izvajanja (pandemija) in uvajanja učenja na daljavo je namen prispevka motivacijski za aplikacijo KU v učni proces na celotni tehniški vertikali izobraževanja.

2 Metoda obrnjenega učenja z učnim modelom poizvedovalnega učenja

Metoda obrnjenega učenja (flipped classroom) je ena izmed metod rotacijskega modela [4]. Rotacijski model je definiran kot pouk, pri katerem učenci rotirajo iz ene učne modalitete v drugo, od katerih je vsaj ena spletno učenje. Rotacije so v pouk vključene že več let, vključitev komponente spletnega učenja v rotacijski model pa omogoči, da je rotacijski model postal eden izmed modelov KU [4].

Pri obrnjenem učenju je bistvo samostojna obravnava učne vsebine s spletnim gradivom, običajno doma. V kolikor to ni mogoče, se poslužimo metode

lažne učilnice. Pri tem modelu učenci snov, ki bi jo pri klasičnem obrnjenem učenju predelali doma, predelajo v učilnici, vendar so pri delu samostojni. Uporabljajo v naprej pripravljeno spletno gradivo ali drugo e-gradivo, ki ga pogledajo na računalnikih. Gradivo pred tem pripravi učitelj, ki med izvedbo lažne učilnice učence le opazuje in usmerja. Za večji učinek lahko spletni ogled predavanj izvedemo v računalniški učilnici, ki bo v tem primeru lažna učilnica, praktično delo pa v delavnici.

Metoda obrnjenega učenja učitelju omogoča, da čas pouka uporabi efektivno in s tem zagotovi premostitev in okrepitev področij, v katerih je znanje učencev pomanjkljivo. Tak pristop je zelo pomemben, saj učenci prenehajo biti le pasivni poslušalci in postanejo aktivno vključeni v učenje [3, 4]. Raziskave kažejo, da pozornost učencev po prvih 10 minutah pouka znatno upada. Čeprav se pozornost učencev v zadnjih nekaj minutah pouka vrne, si zapomnijo le 20 % vsebine, ki je bila predstavljena v času pouka. Pouk, pri katerem so učenci aktivni, povečuje učne rezultate učencev ter izboljša njihovo motivacijo in celo odnos do predmeta, pouka in učenja, spodbuja razmišljanje in znanje na višjih taksonomskih stopnjah, reševanje problemov, kritično razmišljanje, hkrati pa zagotavlja povratno informacijo tako učencu kot učitelju [5]. Pri metodi obrnjenega učenja je učenje osredotočeno na učenca. Učenec nosi odgovornost, da pride k pouku z osnovnim razumevanjem e-gradiv, ki je potrebno, da lahko sodeluje pri pouku. Pridobljeni čas lahko učitelj porabi za učenje, ki je čim bolj osredotočeno (aktivno) na učenca, kot so projektno, problemsko, poizvedovalno učenje (PU), učenje z odkrivanjem, … [5-7].

Za aktivni del obrnjenega učenja izhajamo iz PU. Izbrana strategija je priporočena za uvajanje učencev v induktivne oblike dela [7]. Navajamo jih na znanstveno metodo dela, raziskovanje in jim omogočimo samostojno (nad)gradnjo znanja. Informacije z vpeljevanjem IKT v šolski prostor postajajo dostopnejše učencem. Računalniki omogočajo učencem lažje, samostojnejše iskanje informacij in hkrati lažji nadzor učitelja nad njihovim delom. Proces učenja je strukturiran po fazah, ki imajo poudarek na metodologiji znanstvenega raziskovanja [8]. Izmed številnih učnih učnih modelov PU smo izbrali model 5E, ki se je izkazal za uspešnega tekom izvajanja pedagoške prakse za bodoče učitelje tehnike na Pedagoški fakulteti v Ljubljani. Model 5E zajema faze [9]: vključitev, odkrivanje, pojasnjevanje, izdelovanje in evalviranje. V raziskavah [8], so dokazali, da PU pozitivno vpliva na zmožnost predvidevanja in preizkušanja rešitev problema. S kritičnim mišljenjem učenci vrednotijo ideje, iščejo vzročno posledične povezave in v skladu z rezultati popravijo svoje rešitve. Njihova identifikacija problema je hitrejša, prav tako postavljanje hipotez in načrtovanje poskusov za preizkus rešitev.

3 Primer lažne obrnjene učilnice

Kombinirano učenje z metodo lažne obrnjene učilnice (LOU) z integriranim učnim modelom 5E PU na stopnji strukturiranega vodenja podajamo na primeru izvedbe tehniškega dne za učence 7. razredov osnovne šole.

Učna priprava za izvedbo tehniškega dneva temelji na dvanajstih učnih ciljih (I1-12). Učenec: I1: našteje sestavne dele 3D tiskalnika; I2: pozna delovanje 3D tiskalnika; I3: spozna delovni prostor programa Sketch up; I4: izdela tehnološki list; I5: nariše shemo električnega vezja; I6: pravilno in varno izvaja postopek lotanja; I7: v programu SketchUp nariše model nosilca za kolesa; I8: izdela sestavne dele vozila in jih sestavi v vozilo; I9: v hitrostni dirki preizkusi funkcionalnost in uporabnost vozila; I10: ugotovi, da so lažja vozila hitrejša; I11: analizira, kako popraviti nefunkcionalno vozilo; I12: napiše seznam izboljšav. Vrednotenje I ciljev, po revidirani Bloomovi taksonomski lestvici (RBT), je prikazano tabeli 1. V krepkem so navedeni I cilji, ki so zastavljeni, da jih učenci dosežejo doma. Ti cilji so nižjih RBT (1,2). Cilji, ki jih dosežejo učenci v šoli, so nastavljeni ciljno z izbrano induktivno metodo in posegajo po višjih RBT (3-6).

V prvem delu LOU, ki je vezan na cilje I1-3, se je tehniški dan izvajal tako, da so učenci v dvojicah ogledali e-gradivo na temo električnega kroga in umetnih snovi. Po ogledu so rešili delovne liste, preko katerih so bili vodeni v samostojno ponovitev vsebin. Prvi del LOU se je končal s preverjanjem rešitev na delovnih listih. Večino vprašanj so učenci pravilno rešili. Nerazumevanje smo zasledili pri risanju električnega kroga. Slednjega učenci še niso obravnavali v sklopu elektrike pri pouku Tehnike in tehnologije Tematika jim je poznana iz vsebin predhodnega predmeta Naravoslovje in tehnika.

V drugem delu LOU je tehniški dan potekal po modelu PU 5E (faze 1-5). Model 5E je moderna metoda poučevanja pri kateri učenci usvajajo novo znanje z izkušnjami. V prvi fazi, **vključitev**, je bila učencem podana problemska naloga kot izhodiščna motivacija učencev – nefunkcionalno vozilo na propeler. Izdelki vozil so zelo priljubljeni izdelki v okviru obveznega osnovnošolskega predmeta Tehnika in tehnologija. Večina vozil je tipskih primerov izdelkov brez ažuriranja na sedanjost, da bi lahko vzdrževala motivacijo učencev. Prav slednje je bilo ciljno pri

Tabela 1. Preglednica učnih ciljev razvrščenih po RBT, kjer pomenijo P- pomnjenje, R - razumevanje, Up- uporaba, A- analiza, E- evalvacija, Us- ustvarjanje, Fa - faktografsko, Ko - konceptualno in Pr - procedualno. V krepkem so označeni učni cilji, ki jih učenci dosežejo doma.

Znanje/ proces	1 - P	2 - R	3 - Up	4 - A	5 - E	6 - Us
A - Fa	**I1**	**I2, I3**				A6
B - Ko			I5	I11	I9, I10	
C - Pr			I4, I6, I7, I8			

razvijanju modela vozila kot sredstva za doseganje učnih ciljev. Vozilo na propeler, gnan z enosmernim elektromotorjem na baterije, je bil ciljno razvit za aplikacijo v problemsko izhodišče PU kot aktivni delobrnjenega učenja. S pomočjo metode pogovora so bili učenci usmerjeni v razumevanje, kaj sploh je problem in kakšna je njihova naloga - povrniti osnovno funkcijo vozila (vožnja naravnost) in izboljšati vozilo v razpoložljivem času. Učenci so ugotovili, da lahko nosilce natisnejo s pomočjo 3D tiskalnika. V tej fazi smo z učenci skupaj določili še kriterije, ki jih je smiselno in potrebno upoštevati pri izdelavi vozila.

V drugi fazi, **odkrivanje**, smo učence razdelili v 6. skupin s po tremi člani (omejitev v št. 3D tiskalnikov). Vsaka skupina je najprej pogledala e-gradivo na temo 3D tiskalnika z namenom dopolnjevanja zaznane potrebe po novem znanju. Za večjo strukturiranost novega znanja so tekom ogleda e-gradiva izpolnjevali še delovni list. Pridobljeno znanje so uporabili in samostojno po tiskanih navodilih izdelali model enostavnega ravnila in ga natisnili. Nekateri učenci so potrebovali veliko dodatnega usmerjanja in pomoči. Največ težav so jim povzročale različne verzije SketchUp-a (nekateri računalniki so bili posodobljeni). Na koncu faze so učenci dobili tehniško dokumentacijo vozila. Le to so ustrezno popravili na mestih, kjer so načrtovali izboljšave vozila. Svoj nosilec za kolesa so podali v obliki osnutka z okvirnimi dimenzijami.

Tretja faza, **pojasnitev**, je predvidevala predstavitev sprememb/izboljšav načrtovanih vozil sošolcem. Zaradi nastalih težav v 2. fazi in s tem preseganja predvidenega časovnega okvirja, je bila izvedena reducirana oblika predstavitev učencev (le učiteljem).

V četrti fazi, **izdelovanje**, je bilo načrtovano izdelovanje tehnoloških in operacijskih listov za določitev izvajanja tehnoloških postopkov v delavnici in porazdelitve dela v skupini. Učenci so izdelali le okrnjeno verzijo listov. Za varno in pravilno delo je bilo zato potrebno sprotno opozarjanje. Skupine so si porazdelile delo tako, da je večina učencev izdelovala sestavne dele vozila v delavnici in le en učenec je izdeloval model nosilca. Ker sva tehniški dan izvajali dve študentki, je vsaka nadzirala eno skupino učencev. Nekatere sestavne dele vozila kot so kolesa, podstavek in cev so učenci dobili že izdelane.

Vse skupine so naredile manjše vozilo, v primerjavi s prikazanim, pri čemer so upoštevali hipotezo, da je lažje vozilo hitrejše. Dve (od 6) skupin sta izdelali po obliki precej drugačno vozilo. Eno je imelo štiri kolesa, ostala po tri, kot vzorčno vozilo. Največ težav so imeli učenci pri izdelovanju modela nosilca. Pri delu v delavnici, smo zaradi težav z delovanjem spajkalnika učence razbremenili spajkanja vodnikov na elektromotor in stikalo.

Zadnja, peta faza, **evalvacija,** je obsegala tekmovanje izdelanih vozil. Vsaka skupina je svoje vozilo vrednotila in podala možne izboljšave. Pred tekmovanjem so učenci preizkusili funkcionalnost vozil in jih popravljali do te mere, da so jo zagotovili. S tekmovanjem, nad katerim so bili učenci navdušeni, so ugotovili, da so med vozili razlike. Tekmovanje je bilo na izločanje. Vozila smo naključno razdelili v dve skupini s po tremi vozili. Zmagovalca obeh skupin sta se pomerila v finalu. Podajanje možnih izboljšav smo skrčili na verbalno poročanje skupin. Pred zaključkom tehniškega dneva smo izkoristili čas še za izvedbo post-testa. Pred-test smo izvedli pred začetkom LOU.

Tabela 2. Povprečno število doseženih točk na pred (T_1)- in post-testu (T_2), kjer pomeni P- postavka, RBT - revidirana Bloomova taksonomska lestvica, $\bar{x}$ - aritmetična sredina in s_x - standardni odklon.

P	RBT	T_1: $\bar{x}$ (s_x)	T_2: $\bar{x}$ (s_x)	T_2-T_1/%
1	2A	55,6 (51,1)	70,6 (45,0)	15,0
2	1A	61,0 (50,2)	70,6 (45,0)	9,5
3	2B	27,8 (46,1)	58,8 (50,7)	31,1
4	2A	58,3 (42,9)	61,8 (21,9)	3,4
5	1B	63,9 (33,5)	73,5 (31,2)	9,6
6	4A	30,6 (25,1)	70,6 (30,9)	40,0
7	3B	38,9 (36,6)	52,9 (27,8)	14,1

Pred-/post testa sta zajemala sedem postavk izbirnega tipa. Postavke so ciljale na nižje RBT stopnje v področju faktografskega in konceptualnega znanja, tabela 2. Postavke pokrivajo zlasti zastavljene I cilje, ki jih učenci ne dosežejo s praktičnim delom. Nanašajo se na uporabljeno novo tehnologijo 3D tiskanja. Postavki 6 in 7 ugotavljata v kakšni meri se učencem poveča stopnja razumevanja funkcionalnosti in uporabnosti izdelanih vozil na podlagi aktivne praktične izkušnje, zato sta zastavljeni višje po RBT. Postavke (P) 4-7 so imele 2 pravilna odgovora, zato so bile točkovane z 2. točkami, preostale pa z 1.

Na predtest, ki so ga učenci reševali pred pričetkom tehniškega dneva, je odgovarjalo 18 učencev. Na posttest, po izvedbi tehniškega dneva, jih je odgovarjalo 17. Iz rezultatov predtesta lahko razberemo, da so učenci že bili seznanjeni s 3D tiskom. Najslabše so odgovarjali na 6. (31 %) in 7. postavko (39 %), ki sta višjih taksonomskih stopenj ter postavko 3 (28 %), ki se nanaša na vzrok za uporabo 3D tiskanja. V povprečju so dosegli učenci na predtestu 48 % vseh možnih točk. Postest rešujejo učenci bolje pri vseh postavkah. Največji prirast znanja izkazujejo učenci pri dveh najslabše reševanih postavkah (3 in 6) na predtestu. Procentualni delež razlike teh dveh postavk med post in predtestom je kar 31 % oziroma 40 %. V povprečju so dosegli učenci na postestu 66 % vseh možnih točk, kar je za 18 % bolje kakor na predtestu.

4 Diskusija

Z izvedbo tehniškega dneva po principu induktivnih metod (LOU, PU) ter uporabo nove tehnologije smo ugotovili, da imajo učenci veliko interesa pri preizkušanju nečesa novega, v tem primeru tehnologije 3D tiska. Kot slaba stran se je izkazala potrata časa. Induktivne metode zahtevajo precej več časa kot običajne tradicionalne metode poučevanja. Po drugi strani se je izkazalo, da so učenci usvojili precej novega znanja, kar zagovarjajo rezultati pred/post-testov. Ogled e-gradiva v dvojicah v učilnici, namesto posamezno in

Tabela 3: Informacije o nosilcih po skupinah, kjer pomenijo, S- skupina, ČT - čas tiskanja, STP - stopnja pomoči pri modeliranju (1- brez pomoči, 2 usmerjevalna, 3 - obsežna)

S	ČT /min	Primernost: 1-zadostna, 2- dobra, 3- odlična, 4- izvirna	STP
1	120	1; zmanjšati debelino nosilca	3
2	10	3; hitro tiskanje, majhna masa.	2
3	2	3; hitro tiskanje, majhna masa.	1
4	45	2-4; počasno tiskanje.	1
5	20	2; hitro tiskanje, poznana rešitev.	2
6	90	1; zmanjšati debelino nosilca	3

doma, je bilo glede na reševanje delovnih listov zelo uspešno.

Ključna faza tehniškega dneva je bila v 2. delu LOU, četrti fazi PU. Vse skupine so uspešno zasnovale nosilec za kolesa in upoštevala kriterije. Nosilce so si učenci zamislili sami. Nobena rešitev nosilca ni bila identična rešitvi druge skupine. Nekaj rešitev si je bilo podobnih med seboj (1 - 6 in 2 - 3). Rešitev prve in šeste skupine je bil kvader z izvrtino za os v širini podvozja. Rešitev 2. in 3. skupine sta bili dve ožji kocki oziroma kvadra. Skupina 4 si je zamislila pol valj z izvrtino. Rešitev skupine 5 je bila najbližje predvideni in nakazani rešitvi z vzorčnim vozilom - L profil. V tabeli 3 so vrednoteni nosilci po časovni racionalnosti, primernosti in samostojnosti učencev pri izdelavi modela nosilca in tiskanju, kjer najbolj izstopa skupina 3.

Vse skupine so po uspešno sestavljenih vozilih ugotovile, da vozila še niso funkcionalna. Vozilo ene skupine je viselo preveč naprej (težišče), pri drugem so bili nosilci premajhni (trenje), pri tretji so se kolesa zatikala ob podvozje (nenatančna namestitev)... V kratkem preostalem času do tekmovanja je prišlo do opaznega dviga sodelovanja, spodbujanja in pomoči, da bi dosegli funkcionalnost vozila. Po nekaj neuspešnih popravkih so skupine hitro konvergirale proti pravim rešitvam. Manjše nenatančnosti so reševali z brušenjem, med nosilec in kolo so dodali distančnik... Nekateri učenci pa so ideje za rešitve iskali pri vzorčnem vozilu.

V fazi evalvacije, na tekmovanju, se je v načrtovanju najbolj zavzeta skupina 1, uvrstila tudi v finale. Skupina 5 je imela najbolj izvirno rešitev. Imeli so eno kolo več (4 kolesa), podvozje so integrirali z nosilcem za elektromotor in vpeli propeler na svojsten način, vendar ob hkratni omejitvi pretoka zraka in s tem znižali uporabnost vozila. V finalu sta bili vozili skupin, ki odražata najvišjo in najnižjo stopnjo pomoči (1, 3), tabela 3, kjer je bila zmagovalna skupina 3.

5 Sklep

Uporaba metode lažne obrnjene učilnice kot modela obrnjenega učenja s strategijo poizvedovalnega učenja se je izkazala za zelo uspešno. Motiviranost učencev je bila zelo visoka in kljub večurni izvedbi konstantna. Problemsko izhodišče je bilo strukturirano kot tudi stopnja vodenja poizvedovalnega učenja, kar se je izkazalo za ključno odločitev. Posledično smo dosegli, da so bili vsi učenci uspešni pri reševanju osrednjega problema z izdelkom kot sredstvom za dosego učnih ciljev. S spodbujanjem učencev v razumevanje bistvenih parametrov, ki vplivajo na funkcionalnost vozila, smo dosegli, da so bili motivirani v preseganje za tehniko obvezne funkcionalnosti. Uporaba sodobnih tehnologij (3D tisk) je visoko motivacijska, vendar se motivacija lahko zelo hitro obrne ob pojavljanju težav pri uporabi tehnologije (izgubljanje povezave s programskim orodjem, 3D tiskalnikom, lomljenje filamenta...). Vpeljava tekmovanja med različnimi rešitvami jih je dodatno spodbudila k lastnim in izvirnim rešitvam ter odpravljanju številnih tekočih težav pri doseganju in preseganju funkcionalnosti izdelka. Iz rezultatov izstopa najvišji napredek v znanju učencev, kjer smo predvideli, da bodo najšibkejši - razumevanje parametrov vpliva na funkcionalnost. S podanim primerom KU želimo spodbuditi učitelje v vpeljevanje učencev/dijakov/študentov v oblike učenja na daljavo. Prilagoditve tekom izvajanja, ki jih ne moremo predvideti, so žal neobhodne, vendar kljub temu ob višji stopnji vodenja metoda zagotavlja uspeh.

Literatura

[1] E. K. Kahiigi, Exploring the e-Learning State of Art, *The Electronic Journal of e-Learning*, 6 (2), str. 77 - 88, 2008.

[2] C. J. Bonk in C.R. Graham. *The handbook of blended learning: Global perspectives, local designs* (New Jersey, John Wiley & Sons, 2005).

[3] M. G. Moore , *Handbook of distance education* (New York, Routledge, 2013).

[4] H. Staker, *The rise of K–12 blended learning - profiles of emerging models* (Lexington, Innosight institute, 2011).

[5] J. E. McLaughlin in ostali, The Flipped Classroom: A Course Redesign to Foster Learning and Engagement in a Health Professions School, *Academic Medicine*, 89 (2), str. 236–243, 2014.

[6] N. Lokar, J. Jamšek. Sodobne metode poučevanja v tehniškem izobraževanju: kombinirano učenje pri predmetu tehnika in tehnologija, V: Zbornik ERK 2017, 25. – 26. september 2017, Portorož, Slovenija, B. Zajc, ur., A. Trost, ur., Ljubljana: IEEE Region 8, Slovenska sekcija IEEE, str. 551-554, 2017.

[7] M. J. Prince, R. M. Felder, Inductive Teaching and Learning Methods: Definitions, Comparisons, and Research Bases, *Journal of Engineering Education*, 95 (2), str. 123-138, 2006.

[8] M. Pedaste in ostali, Phases of inquiry-based learning: Definitions and the inquiry cycle, *Educational research review*, 14, 47-61, 2015.

[9] R.W. Bybee, The BSCS SE Instructional Model: Personal Reflections and Contemporary Implications, *Science and Children*, 51 (8), str. 10-13, 2014.

Virtualni laboratorij za poučevanje digitalnih sistemov

Janez Perš, Andrej Trost

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, 1000 Ljubljana
E-pošta: andrej.trost@fe.uni-lj.si

Virtual laboratory for Digital Systems Education

The paper describes a digital system teaching methodology, based on online software tools and virtual laboratory. The tools are freely available and enable effective remote learning in a course with a limited time frame. A teaching and student experience with the proposed methodology implemented in an introductory electrical engineering course Programmable Digital Systems is presented.

1 Uvod

Obravnava digitalnih sistemov je pomembna za študente elektrotehnike, ki se bodo v praksi srečevali s snovanjem ali programiranjem digitalnih naprav [1]. Praktično znanje pridobivajo z načrtovanjem vezij na razvojih ploščah s programirljivimi vezji FPGA [2]. Pri delu potrebujemo industrijska orodja proizvajalcev FPGA, ki so optimizirana za razvoj kompleksnih vezij, obravnava manjših, učnih primerov vezij pa je relativno zahtevna in zamudna. Za razliko od programske opreme, smo pri izbiri orodij za razvoj vezij FPGA zelo omejeni [3]. Na laboratorijskih vajah je potrebno poenostaviti postopek načrtovanja za študente z osnovnim poznavanjem računalniških orodij in omejenim časom izvedbe naloge.

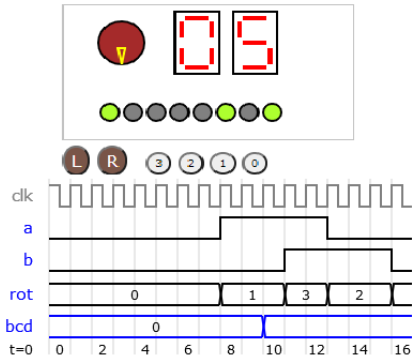
Vaje za študente prvih letnikov smo do sedaj izvajali z vnaprej pripravljenimi projekti in poudarkom na shematskem načrtovanju vezij. Vaje smo prenovili z vključitvijo spletnega orodja in virtualnega laboratorija, ki omogoča opravljanje poskusov iz domačega računalnika.

Digitalne sisteme načrtujemo shematsko ali pa v standardnem jeziku VHDL oz. Verilog [4]. Razlaga jezika, kot je VHDL, zaradi posebnosti sintakse ni primerna za izbirni predmet z manjšim obsegom pedagoških ur. Rešitev vidimo v poenostavljenem jeziku, ki vključuje le osnovne koncepte in vodi načrtovalca digitalnega vezja k pravilni uporabi stavkov pri opisu vezja [5]. Poenostavljen jezik s spletnim prevajalnikom in simulatorjem omogoča utrjevanje snovi in delo od doma. Virtualni laboratorij pa dopolnjuje ali celo nadomešča izvedbo vaj in poskusov v pravem laboratoriju [6]. V prispevku bomo predstavili: posodobitev poučevanja digitalnih sistemov s spletnimi orodji, vzpostavitev virtualnega laboratorija in izkušnje pedagogov in študentov pri izvedbi laboratorijskih vaj. Obravnavan pristop smo uporabili pri predmetu Programirljivi digitalni sistemi (PDS) v prvem letniku študijske smeri Aplikativna Elektrotehnika.

2 Orodja za načrtovanje vezij

Na laboratorijskih vajah predmeta PDS utrjujemo predavano snov z izvedbo praktičnih primerov kombinacijskih in sekvenčnih vezij, sekvenčnih strojev in digitalnih sistemov s procesorjem. Primere preizkušamo na razvojnih ploščah Altera DE0 Nano [7]. Digitalna vezja načrtujemo in prevajamo z orodjem Intel/Altera Quartus Prime, ki je na voljo brezplačno za nekomercialne namene. Orodje omogoča risanje sheme, prevajanje vezja iz jezikovnega opisa, upravljanje razvojnih plošč in simulacijo z zunanjim programom, kot je npr. Modelsim.

Industrijsko orodje zahteva skrbno pripravo navodil študentom in vodenje ali demonstracijo uporabe. Izdelavo vezja z risanjem sheme študenti usvojijo hitro, vendar je praktično uporabna le za nekatera osnovna vezja ali izdelavo sistemov iz vnaprej pripravljenih komponent (npr. vmesniki, jedra CPU).



Slika 1: Simulirana razvojna plošča v spletnem orodju SHDL in potek signalov pri izvedbi vaje elektronska ključavnica.

Strojno-opisni jeziki omogočajo bolj učinkovito načrtovanje, zahtevajo pa obsežnejšo razlago sintakse in veliko praktičnih vaj. Pri predmetu prvega letnika smo zato uvedli poenostavljen jezik SHDL, ki omogoča načrtovanje v spletnem orodju in povezavo z industrijskimi orodji za preizkus na vezjih FPGA [5].

Spletno orodje omogoča izdelavo in simulacijo modelov vezij brez zamudnega nameščanja zahtevne programske opreme. Kombinacijska vezja opišemo s prireditvenimi stavki, ki vsebujejo logične in aritmetične izraze in pogojnimi stavki. Sintaksa operatorjev je povzeta po jeziku VHDL, podatkovni tipi pa so le enobitni ali celoštevilski vektorji. Deklaracije priključkov in notranjih signalov nadomestimo z vnosom signalov v tabelo. Opis sinhronih sekvenčnih gradnikov (flip-flopi, registri) naredimo s posebnim prireditvenim operatorjem. Sintaktična pravila so zaradi omejenega

nabora konstruktov jezika SHDL izjemno poenostavljena in omogočajo hitro izdelavo modelov kombinacijskih ter sekvenčnih vezij, pri katerih so načrtovalci osredotočeni na pomen (semantiko) stavkov strojno-opisnega jezika. Simulacijo modela naredimo v spletnem orodju tako, da določimo število simulacijskih ciklov, v vsakem ciklu grafično nastavimo vrednosti vhodov, izvedemo simulacijo in pregledamo rezultat.

Spletno orodje smo nadgradili še z grafičnim prikazom razvojne plošče na kateri lahko pritiskajo tipke, vrtijo kodirnik in opazujejo izhod na 7-segmentnih prikazovalnikih (slika 1). S spletnim orodjem študenti izvajajo eksperimente in so sposobni samostojno rešiti vajo načrtovanja digitalnega vezja, saj poenostavljen jezik ne zahteva veliko razlage in vodenja.

3 Virtualni laboratorij

V pravi učilnici, v kateri se izvajajo vaje s programirljivimi vezji, je za študente na voljo več kot 20 računalnikov, in 20 prilagojenih razvojnih plošč DE0-Nano, zgrajenih okrog programirljivega vezja Altera/Intel Cyclone IV. Vaje izvajamo v skupinah, ki štejejo od 15 do 20 študentov. V času, ko izvajanje vaj v laboratoriju ni bilo mogoče, smo analizirali več možnosti izvajanja vaj na daljavo, in sicer:

- **Oddaljeni dostop** do obstoječih računalnikov zahteva najmanj tehničnih posegov, vendar pa ostane problem povratne informacije – študenti ne morejo videti rezultatov svojega dela na razvojnih ploščah. Rešitev tega problema bi zahtevala najmanj 15 spletnih kamer, za vsako delovno mesto eno, poleg tega pa bi bilo treba konfigurirati oddaljen dostop za vse računalnike v učilnici.
- **Delo na domu** z razvojnimi ploščami je bilo nemogoče izvesti, saj je na voljo 20 plošč ob 45 vpisanih študentih, vaje pa so bile načrtovane v treh ciklih.
- **Virtualni strežnik** je ponujal največ prednosti – ob uporabi dovolj zmogljive strojne opreme je bilo treba konfigurirati samo en fizični računalnik, ter ga kombinirati z eno samo kamero visoke ločljivosti. Ves virtualni laboratorij je tako zavzemal le 2 m^2 prostora in je prikazan na sliki 2.

3.1 Virtualni strežnik

Uporabljen je bil visoko zmogljiv osebni računalnik v naslednji konfiguraciji: 64 GB RAM, 6-jedrni procesor Intel(R) Core(TM) i7-8700K z delovno frekvenco 3.70 GHz, SSD disk 500 GB ter dva magnetna diska 2 TB. Dve grafični kartici NVIDIA RTX-2080Ti v okviru virtualnega laboratorija nista bili uporabljeni.

Na strežniku je tekel OS Ubuntu Linux 18.04 LTS, ki smo ga uposobili za virtualizacijo s pomočjo emulatorja QEMU (angl. Quick Emulator, [8]) in modula jedra KVM (angl. kernel virtual machine, [9]), obe v standardnih verzijah za Ubuntu Linux 18.04 LTS. QEMU



Slika 2: Postavitev virtualnega laboratorija. Zmogljiv strežnik je pod mizo. Vidnih je 15 razvojnih plošč in IP kamera visoke ločljivosti na stojalu. V času delovanja virtualnega laboratroija so bile okenske žaluzije zaprte in prižgane stropne fluorescentne luči, da smo zagotovili čimbolj konstantne svetlobne razmere, ki bi drugače vplivale na vidnost prikazovalnikov na razvojnih ploščah.

je sicer emulator za vrsto različnih arhitektur, vendar pa lahko na platformah z arhitekturo x86/x64 uporablja KVM in s tem deluje kot pravi hipervizor. V nadaljevanju članka pri uporabi termina ”hipervizor” torej mislimo na kombinacijo OS Ubuntu ter orodij QEMU in KVM.

3.2 Virtualizacija študentskih računalnikov

Na strežniku je s pomočjo KVM teklo 15 virtualnih računalnikov z OS Windows 7. Vsak virtualni računalnik je imel na razpolago 3 GB RAM, 120 GB magnetnega diska (za sistemsko particijo) in 5 GB diska SSD (za namizje, ki je služilo kot delovni prostor za študente).

V vsakem virtualnem računalniku je bila nameščena programska oprema Altera/Intel Quartus, različica 18, skupaj z USB gonilniki za razvojne plošče DE0-Nano.

V virtualnih računalnikih so bili nameščeni optimizirani gonilniki, ki jih ponuja standard VirtIO [11], torej gonilnik za virtualni grafični vmesnik (”grafično kartico”), gonilnik za virtualni omrežni vmesnik, gonilnik za virtualni disk, ter gonilnik, ki omogoča hipervizorju, da virtualnemu računalniku *odvzame* neuporabljen, a dodeljen pomnilnik (angl. VirtIO Baloon driver). Tako hipervizorju, kot tudi vsakemu virtualnemu računalniku je bil dodeljen IP naslov v naslovnem prostoru, dostopnem študentom in asistentom preko povezave VPN (angl. virtual private network).

Dostop do virtualnih računalnikov je bil možen na tri načine:

- Preko specializiranega vmesnika SPICE [10], ki omogoča optimalno uporabniško izkušnjo. Neposredno je povezan z hipervizorjevim delom virtualizirane grafične kartice (pri povezavi navedemo IP naslov hipervizorja, ne virtualnega računalnika, vsak virtualni računalnik pa zahteva svoja IP vrata na hipervizorju), kar omogoča zelo optimiziran prenos grafičnega prikaza tudi preko omrežnih povezav, vendar le enemu uporabniku (v tem primeru študentu). Takšen način povezave recimo omogoča tudi, da se ob spremembi velikosti okna klienta SPICE (na študentovem računalniku) odrazi v ločljivosti, ki jo grafični gonilnik v Windows 7 sporoča aplikacijam, kar zelo poveča ergonomičnost dela. Protokol SPICE uporablja tudi video kodiranje slike za dele ekrana, na katerih se dogajajo hitre spremembe.
- Deljena mapa pod OS Windows, v namen prenašanja datotek. Poleg tega tudi protokol SPICE omogoča prenos datotek po sistemu ”povleci in spusti” (angl. drag and drop), ki so ga študenti tudi najpogosteje uporabljali.
- Preko nameščenega strežnika VNC, ki je bil namenjen asistentom.

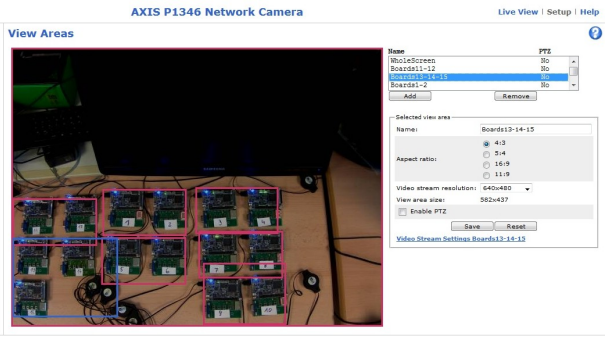
Kombinacija vmesnikov SPICE in VNC je omogočala, da so imeli študenti neposreden in hiter grafični dostop, ki je omogočal tudi načrtovanje (”risanje”) vezij, asistent pa se je lahko v primeru vprašanj povezal na njihov virtualni računalnik, ne da bi vmesnik SPICE prekinil povezavo s študentom.

3.3 Povezava razvojnih plošč v virtualiziran sistem

Razvojne plošče DE0-Nano imajo USB vmesnik, ki rabi ustrezen gonilnik na strani operacijskega sistema (v našem primeru Windows 7). Hipervizor omogoča neposredno povezavo USB (angl. USB passthrough) fizične naprave z virtualnim računalnikom, pri čemer gonilniki za hipervizor niso potrebni.

Fizični priklop 15 razvojnih plošč je zahteval dva razdelilca USB (angl. USB hub) z lastnim napajanjem.

Problem, ki se pojavi pri veliko USB napravah je, da so vse naprave fizično, električno in programsko povsem enake, zato hipervizor ne more določiti, katera razvojna plošča (to je, katero študentsko delovno mesto) je bilo ob katerem trenutku priklopljeno v USB vmesnik. Delo s takšnim sistemom je vseeno možno, če konfiguracijo virtualnih računalnikov (konfiguracijske datoteke za QEMU, KVM in povezane knjižnice) dokončno nastavimo po priklopu vseh plošč. V primeru izklopa računalnika, ponovnega zagona hipervizorja, ali izpada elektrike pa je vrstni red dodeljenih številk povsem naključen, kar zahteva ponovno rekonfiguracijo. Problem se da nasloviti, če imamo možnost nastaviti unikatno serijsko številko vsaki od razvojnih plošč, česar pa se zaradi časovne stiske nismo lotili. V dveh mesecih delovanja virtualnega laboratorija v našem primeru ni bilo dogodkov, ki bi povzročili potrebo po rekonfiguriranju.



Slika 3: Nastavljanje virtualnih pogledov v spletnem vmesniku kamere. Ker kamera omogoča največ 8 pogledov, smo v en pogled zajeli dve ali tri ploščice, kar še vedno pomeni radikalno zmanjšanje potrebne pasovne širine za prenos videa do študentov.

Drugih nastavitev, kot le ID številke USB naprave za vsak virtualni računalnik hipervizor ne rabi. V virtualnem računalniku je tako priključena razvojna plošča razpoznana tako, kot če bi jo priklopili v fizični računalnik brez posredovanja hipervizorja.

3.4 Kamera za vizualno povratno informacijo študentom

Bistven del virtualnega laboratorija je možnost, da študenti vidijo, kako se razvojna plošča odzove na naloženo vezje. Pri vajah v pravi učilnici smo uporabljali 8 LED diod in matrični LED prikazovalnik, zato je bilo treba zagotoviti enako funkcionalnost tudi v virtualnem laboratoriju.

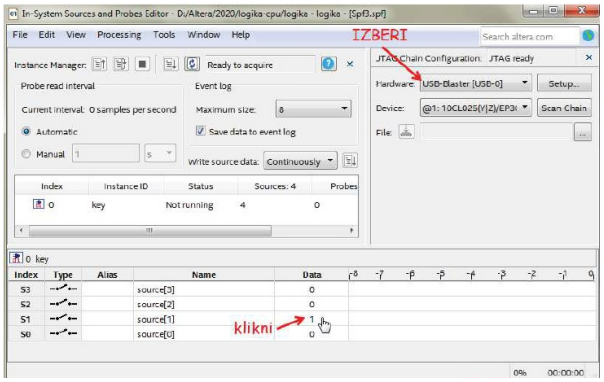
To smo storili s pomočjo ene same IP kamere visoke ločljivosti (Axis P1346), ki je delovala v polni ločljivosti 5 megapikslov. Kamera v svoji vgrajeni programski opremi omogoča določanje do osmih virtualnih pogledov, s katerimi dosežemo, da ni treba vsakemu od študentov prenašati videa visoke ločljivosti, ampak le manjši del, ki prikazuje tisto razvojno ploščo s katero dela – vsak pogled je namreč dostopen preko unikatnega URL naslova video toka. Definicija pogledov za naš primer s 15 razvojnimi ploščami je prikazan na sliki 3.

3.5 Krmiljenje virtualnih vhodnih signalov

Pri delu na daljavo so študentom nedostopne tudi tipke na razvojni plošči, zato je potrebna drugačna rešitev. Orodje Quartus ponuja ”In system Sources and Probes Editor”, ki ga prikazuje slika 4. S pomočjo tega vmesnika v orodju Quartus so lahko študenti spreminjali vrednost vhodnih signalov tudi na daljavo ob minimalnih spremembah v shemi vezja.

4 Izvedba laboratorijskih vaj

Vaje v virtualnem laboratoriju so potekale tako, da so se študenti zbrali v ”virtualni učilnici”, ki jo ponuja orodje za učenje na daljavo `e.fe`, sicer namenjeni predavanjem, in tam dobili vsa potrebna navodila. Ves čas vaj sta bila prisotna tako asistent kot laborant. Izkaže



Slika 4: Vmesnik v orodju Quartus, ki nadomesti pritiskanje fizičnih tipk na razvojni plošči.

se namreč, da delo v virtualnem svetu zahteva več naporov s strani pedagoškega osebja. Poleg pomoči študentom pri sami izvedbi vaj je laborant skrbel tudi za administrativna opravila, ki smo jih zbrali v spletni tabeli. V njej je laborant beležil prisotnost, dodeljevanje virtualnih računalnikov študentom ter evidentiral ocene, ki jih je sporočal asistent. Asistent se je tako lahko povsem posvetil študentom brez pogostega brskanja po dokumentaciji, kar lahko zaradi pomanjkanja prostora na zaslonu zelo upočasni delovni proces.

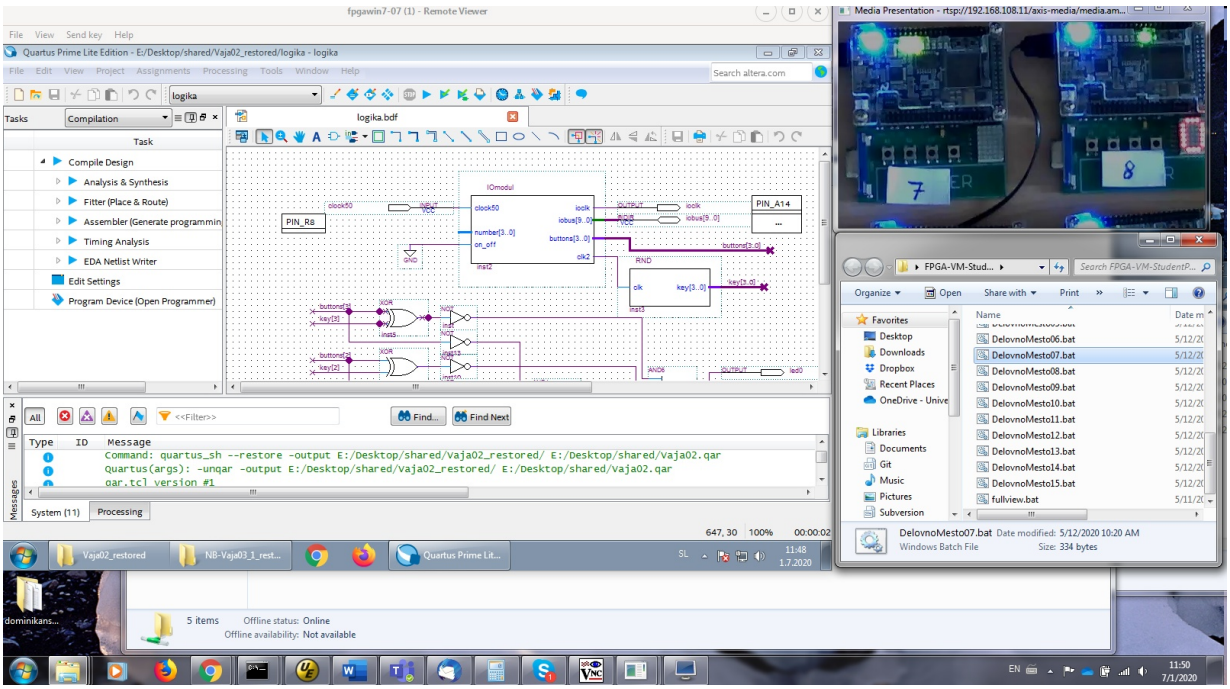
Pred uporabo virtualne učilnice smo naredili poskusno vajo z nekaj študenti - prostovoljci, s katerimi smo preverili delovanje sistema. Prostovoljci so skupaj z dodatnim pedagoškim osebjem pomagali pri prvi vaji v virtualnem laboratoriju, kjer smo predvsem reševali težave z nastavitvami programske opreme na njihovih računalnikih. Z uporabo deljenja zaslona na spletni učilnici smo začetne težave uspešno rešili, ob zaključku vaje pa so študenti učilnico uporabili tudi za kratek zagovor vsake opravljene vaje.

5 Zaključek

Konfiguracija virtualnega laboratorija se je izkazala za primerno, in je omogočala udobno delo, če so študenti prevanjanje vezij izvedli na svojem računalniku, in uporabili virtualne računalnike le za nalaganje prevedenih datotek na razvojne plošče. Sočasno prevajanje na več kot 5 delovnih mestih lahko strežnik popolnoma upočasni, po naši analizi pa so ozko grlo vhodno/izhodne operacije orodja Quartus. Takšen laboratorij bi v idealnem primeru deloval na 3-5 SSD diskih, pri čemer bi na vsakem bili nameščeni virtualni diski 3-5 virtualnih računalnikov.

Laboratorijske vaje je v začetku semestra obiskovalo 45 študentov. Uvodne tri vaje smo izvedli na Fakulteti, sedem pa v virtualnem laboratoriju. Vse vaje je uspešno opravilo 39 študentov (87%), kar je podoben delež kot v letih, ko smo imeli običajno izvedbo vaj.

Po izvedenih laboratorijskih vajah smo naredili anketo med študenti pri kateri smo prejeli 21 odgovorov. Študenti so zelo dobro ocenili delo v virtualnem laboratoriju (ocena 4.4 od 5), možnost dela doma izven terminov vaj (4.1) in spletno orodje (4.5). Delo s programi virtualnega laboratorija in spletnim orodjem so ocenili kot enostavno, uporabo orodja Quartus pa zahtevnejše. V odzivu so posebej pohvalili možnost dela od doma v naročenem terminu.



Slika 5: Pogled na virtualno učilnico, kot ga vidi študent. Levo je preko odjemalca (prikazovalnika) SPICE prikazano namizje virtualnega računalnika, pod njim pa deljena mapa z virtualnim računalnikom.Desno zgoraj je okno, kjer se prikazuje video iz učilnice v realnem času – vidni sta dve razvojni plošči. Desno spodaj je nabor skriptnih datotek, ki priklopijo študenta na izbrano delovno mesto v učilnici.

Literatura

[1] M. M. R. Mano, M. D. Ciletti, “Digital Design, Global Edition”, Pearson Education Limited, Harlow, UK, 2018.

[2] D. Zhang, P. Qian, L. Wang, Y. Guo, C. Sun, Experimental Case Design of Digital Logic Based on Through-Type Teaching, 13th International Conference on Computer Science & Education (ICCSE), Colombo, 2018, pp. 552-555

[3] K. Morris, The FPGA Tool Problem Why Not Open Source?, Electronic Engineering Journal, October 6, 2016

[4] H. A. Ochoa, M. V. Shirvaikar, A Survey of Digital Systems Curriculum and Pedagogy in Electrical and Computer Engineering Programs, ASEE Gulf-Southwest Section Annual Conference, Austin, 2018

[5] A. Trost, A. Žemva, Poenostavljen strojno-opisni jezik, Zbornik osemindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2019, 339-342, September 2019, Portorož, Slovenija

[6] N. Fujii and N. Koike, "A time-sharing remote laboratory for hardware design and experiment with shared resources and service management,"2005 6th International Conference on Information Technology Based Higher Education and Training, Santo Domingo, 2005, pp. T2B/5-T2B10

[7] Altera Programirljiva vezja in razvojne plošče, FE, 2020, https://lniv.fe.uni-lj.si/altera/

[8] F. Bellard, QEMU, a Fast and Portable Dynamic Translator, Proceedings of the USENIX Annual Technical Conference 2005, Anaheim, CA, 2005, str. 41.

[9] Linux KVM, `https://www.linux-kvm.org/page/Main_Page`. Dostopano 25.7.2020.

[10] SPICE project, `https://www.spice-space.org/index.html`. Dostopano 25.7.2020.

[11] Libvirt virtualization API: virtIO, `https://wiki.libvirt.org/page/Virtio`. Dostopano 25.7.2020.

Učenje na daljavo: spletno učenje v osnovnošolskem tehniškem izobraževanju - pedagoška praksa

Bernarda Urankar, Janez Jamšek

Oddelek za fiziko in tehniko, Univerza v Ljubljani, Pedagoška fakulteta, Kardeljeva pl. 16, 1000 Ljubljana
E-pošta: bernarda.urankar@pef.uni-lj.si, janez.jamsek@pef.uni-lj.si

Distance learning: online learning in primary school technology education - teaching practice

Abstract. *During introduction of pandemic the education process had to be transformed to meet the current needs. Paper presents distance learning model that took place as primary schools were lock down. Tools that enable learning model are classified and presented. Online learning for D&T primary school subject executed by technology students on teaching practice is evaluated regarding used programme/web tools. Arnes web class is the most often used as E-learning platform. Multiple tools are typically used for designing teaching material. First, tools for sound and video files (Audacity, OSB, Adobe premiere) are used. Processed files are incorporated in PowerPoint document and used as animation/video. Some technology tools can be used for field specific topics (technical drawings, electric circuits). The prepared learning material was often published on YouTube. Pupil's knowledge assessment was most often obtained by using Google forms, Kahoot and 1ka. For clarifying the conceptual leaning material misunderstanding mentors/students used online video meeting such as MS teams or Zoom. Although online learning was new for pupils, mentors and students all of them benefited from this learning/teaching experience. Whereas pupils met learning customization possibilities mentors/students gained awareness of distance learning needs to meet contemporary learning technology.*

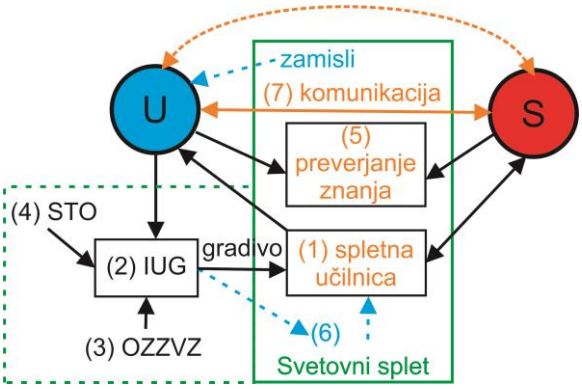
1 Uvod

Tehniško izobraževanje na celotni vertikali od osnovnošolske do univerzitetne stopnje se večinsko izvaja z uporabo najstarejše metode učenja oziroma poučevanja, s tradicionalnim poukom. Zanj je značilna izvedba v učilnicah, kjer obstajajo prostorske in časovne omejitve. S tehnološkim razvojem se pojavi zahteva za preseganje časovno/prostorskih omejitev. V začetku 20. stoletja se pojavi dopisno izobraževanje, ki ga štejemo kot začetke učenja na daljavo [1]. Pospešen razvoj tehnologije radia in televizije, še bolj pa z razvojem in širjenjem interneta so se učni materiali preselili na splet. Nov način učenja je bila poimenovan kot e-učenje/e-izobraževanje in je doživelo svoj vrhunec v letih 1997-1999. Učenje na daljavo in e-učenje se v nekaterih primerih prekrivata, vendar nista enaka [1, 2]. E-učenje je kakršno koli učenje, ki vključuje tehnologijo kot pomoč pri učenju. Izvesti ga je možno tako v razredu kakor tudi, kadar sta učitelj in učeči ločena, kar ne drži za učenje na daljavo [3]. Skupaj z naglim razvojem novih tehnologij se vzporedno razvija tudi e-učenje od prvega pojava integracije osebnega računalnika in s tem računalniško podprtega poučevanja (CAI) do danes aktualnih množičnih odprtih tečajev na spletu in njihovih različic (MOOC, c-MOOC, x-MOOC, LOOC,...) [4]. E-učenje ima zelo veliko različnih poimenovanj, ki se vežejo na uporabljeno tehnologijo, naprimer spletno, virtualno, mrežno, distribuirano, WEB poučevanje... Danes bolj razširjen termin za e-učenje je spletno učenje (online learning). Opredeljeno je kot učna izkušnja v sinhronih ali asinhronih okoljih z uporabo različnih naprav (npr. mobilnih telefonov, prenosnikov itd.) z dostopom do interneta. V teh okoljih so učenci lahko kjer koli (krajevno neodvisni) in se učijo in sodelujejo z inštruktorji in drugimi študenti [5]. Kljub začetni uspešnosti metode e-izobraževanja, se je v nadaljevanju izkazalo, da strukturirano učno okolje in ciljno izdelana e-gradiva udeležencev ne motivirajo zadostno, da bi vztrajali do konca. Nadgradnja e-učenja je kombinirano učenje [2].

S pojavom pandemije zaradi razširjenja virusa covid-19 so bili učitelji na celotni tehniški vertikali (in tudi vsi drugi) prisiljeni čez noč preiti iz tradicionalnega poučevanja na spletno učenje. Medtem, ko so bodoči učitelji tehnike deležni ustreznih znanj na podiplomskem študiju in so zato usposobljeni za takšen način uvajanja in učenja učencev na osnovni/ srednješolski ravni, pa teh znanj še nimajo bodoči učitelji na dodiplomski stopnji. Tekom študija imajo v 3. in 4. letniku obvezno pedagoško prakso. V prispevku podajamo smernice za spletno učenje in rezultate izvedbe pedagoške prakse bodočih učiteljev tehnike na osnovnih šolah v situaciji za katero niso bili pripravljeni ne učitelji mentorji, ne bodoči učitelji-študenti in ne učenci v šoli. Zaradi možnosti ponovitve razmer, ki pogojujejo izvajanje spletnega učenja oziroma učenja na daljavo na celotni tehniški vertikali izobraževanja, je namen prispevka usmerjevalen, na primeru.

2 Spletna orodja za učenje na daljavo

Pri razglasitvi pandemije so bili izvajalci izobraževanja v različnih državah deležni zelo različnega usmerjanja in vodenja. 15. 03. 2020 sta UNESCO INRULED in Inštitut za pametno učenje (SLIBNU) izdala priročnik o spodbujanju prožnega učenja med izobraževalnimi motnjami v katerem so podane smernice in vodila učiteljem v novo nastali situaciji [6]. V Sloveniji ob



Slika 1. Prikaz intuitivnega modela spletnega učenja kjer pomenijo U - učitelj, S - učenec, IUG - izdelava učnih gradiv, OZZVZ - orodja za zajem, obdelavo video in zvočnih posnetkov, STO - specifična tehniška orodja in (6) - objava na svetovnem spletu.

uvedbi pandemije ni bilo vodil učiteljem, kako naj transformirajo učni proces. Zaradi nepoznavanja učnih modelov in orodij bomo izhajali iz intuitivnega modela spletnega učenja, slika 1, [7]. Učitelji (U) teoretični del predavanj prelevijo v podajo učnih gradiv učencem (S) preko spletnih učilnic (1), ki jih pridobijo na svetovnem spletu ali jih deloma/v celoti izdelajo sami s pomočjo razpoložljivih spletnih orodij (na spletu ali na računalniku) (2-4). Stopnjo usvojenega znanja iz posredovanih e-gradiv preverjajo z uporabo orodij za preverjanje znanja (5) (vprašalniki, kvizi, naloge,...). V primeru ugotovljenega nerazumevanja z učenci izvedejo zvočno in/ali video komunikacijo (7) preko spletnih orodij ali telekomunikacijskih naprav. Praktični del pouka se prelevi v podajanje in evalviranje že izvedenih primerov.

V nadaljevanju izhajamo iz predpostavljenega znanja tipičnega učitelja (naprimer obvladovanje Microsoft office orodij) in najbolj razširjenih spletnih orodij s katerimi se je že spoznal ali jih uporabil (naprimer youtube). Orodja podajamo po smiselnih sklopih (1-8) glede na prikazani model na sliki 1: (1) - orodja za spletne učilnice, (2) - orodja izdelovanje učnih gradiv, (3) - orodja za zajem, obdelavo video (V) in zvočnih (Z) posnetkov, (4) specifična tehniška orodja, (5) orodja za preverjanje znanja, (6) orodja za objavo učnih vsebin na spletu, (7) orodja za video spletna srečanja (7) in (8) ostala orodja.

(1) Orodja za spletne učilnice izhajajo iz sistemov za upravljanje učenja iz začetkov e-učenja in se danes razvija v učne platforme. Le te v osnovi najmanj omogočajo ustvarjanje in organiziranje/posredovanje nalog in drugega učnega gradiva, nudenje povratne informacije in enostavno komunikacijo z učenci. Pri nas je najbolj razširjena spletna učilnica Arnes. Mednarodno brezplačne učilnice z daljšim pojavom Edmodo, Beenpod, Goclass in novejše popularne alternative kot Schoology, Canvas, Google Classroom [8].

(2) Orodja za izdelovanje učnih gradiv. Najbolj osnovni orodji izhajata iz Microsoft office orodij. To sta Word in PowerPoint (PPT). Z Wordom lahko enostavno generiramo *.pdf dokumente. PPT je veliko primernejši za pripravo učnih gradiv. Poleg besedila, slik in grafov omogoča vstavljanje video in zvočnih posnetkov ter izdelavo animacij in interaktivnih možnosti z uporabnikom. S časovno zaporedno vezanimi animacijami predvajanja slik, video posnetkov ali zvočnih razlag, je podajanje učnega gradiva podobno postopni razlagi v razredu. S tem lahko dosežemo višjo motivacijo učencev za ogled vsebin. Učenci lahko pregledujejo posamezne prosojnice v svojem tempu in imajo ob tem čas za razmislek, pisanje zapiskov.... Prednosti video posnetkov, dostopnih na YouTube portalu, v primerjavi s PPT je večja dostopnost z različnimi napravami (računalniki, telefoni, tablicami...), ki imajo dostop do interneta. Pri tem učenci ne potrebujejo PPT pregledovalnika. Učenci lahko video predvajanje ustavijo, ali ga predvajajo v počasnejšem oz. hitrejšem tempu, vendar to ni primerljivo s pregledovanjem PPT prosojnic v lastnem tempu.

(3) Orodja za zajem, obdelavo video (V) in zvočnih (Z) posnetkov. Med ogledom video posnetka je primerneje, da učenci poslušajo razlago in pri tem opazujejo dogajanje, kot da berejo podnapise. V nadaljevanju podajamo nekaj najbolj med študenti razširjenih in prosto dostopnih orodij (OBS, Geforce experience, Bandicam, Speechnotes, Adobe Premiere, Audacity) [9]. OBS, Geforce experience Bandicam omogočajo visoko zmogljivo snemanje s kamer in drugih zunanjih enot kot tudi dogajanja na zaslonu (predavanja, seminarje, video konferenc). V orodje Speechnotes se lahko vpišemo z Google računom. Govorjeno besedilo z video posnetka lahko zapiše v pisno obliko. Transkript govora v npr. angleškem jeziku kopiramo v spletni prevajalnik in ga prevedemo. Besedilo preberemo in prepišemo zvočni zapis izvornega video posnetka z naprimer Youtube portala. Navedena orodja omogočajo tudi osnovno video obdelavo, zmogljivejše orodje pa je Adobe Premiere. Izdelane video posnetke montiramo v učno video gradivo. Ciljno orodje za snemanje in urejanje zvočnih zapisov iz različnih zunanjih naprav in drugih medijev je Audacity. Je enostavno za uporabo, omogoča izrez nepotrebnih zvočnih zapisov in odstranjevanje šumov. Omogoča izvoz datotek v različnih formatih. Datoteke z zvočnimi posnetki, npr. razlage dogajanja v pretočni hidroelektrarni, lahko vstavimo v PPT predstavitev.

(4) Specifična tehniška orodja se nanašajo na strokovna področja strojništva in elektrotehnike. Za osnovnošolski nivo podajamo za prvo področje nekaj orodij, ki omogočajo tehniško dokumentacijo (CiciCAD, Qcad, freeCAD, Google sketchup in Solid Edge) in nekaj orodij v povezavi z električnim tokokrogom (Edison, Crocodile clips, Yenka in Phet) [10]. Ciljno za osnovnosolski nivo je bilo razvito orodje CiciCAD, ki se še dan danes uporablja, čeprav ne zadostuje več potrebam sodobne dokumentacije. Zaradi enostavnosti je zelo razširjen Google Sketchup, ki pa ne omogoča zahtevam standardov tehniškega risanja. V

slovenščino preveden je QCAD, ki pa je brezplačen le za čas preskušanja. Brezplačno orodje je FreeCAD, medtem ko je profesionalno orodje Solide Edge brezplačno za izobraževalne namene brez okrnjenih funkcij. Orodje Edison je brezplačen za čas preskušanja. Njegova značilnost sta dve delovni površini, ena z modeli realnih gradnikov električnih vezij na drugi pa el. simboli, ki jih učenec povezuje v shemo. Dodaja lahko številne merilnike. V slovenščini obstaja okrnjena verzija, kjer vezje ni prikazano tudi s simboli in ima le omejeno število komponent/elementov. Crocodile clips je ciljno orodje za osnovnošolske učence. Omogoča izdelovanje simulacij delovanja električnih vezij (tudi nekaterih strojnih elementov). Uporaba programa je enostavna in omogoča grafičen izris opazovanih parametrov. Yenka je nadgradnja Crocodile clips simulacij. Je brezplačno programsko orodje brez časovnih omejitev. Primerna je za poučevanje na daljavo, kot tudi za samostojno učenje v varnem virtualnem okolju. Ponujeno je sestavljanje električnih vezij s pravimi 3D modeli električnih komponent ali s simboli posameznih električnih elementov in shranitev delovanja sestavljenega vezja kot delujoče simulacije. Phet orodje ponuja on-line simulacije zlasti za naravoslovje in ponuja virtualni laboratorij. Je prosto dostopen in preveden v slovenski jezik. V odvisnosti od pogojev električnih parametrov v sklenjenem električnem krogu se prikaže tudi npr. različna svetilnost žarnice in ne le stanje vklopa/izklopa. Orodje je zelo uporabno zlasti zaradi svoje nazornosti.

(5) Orodja za preverjanje znanja. Obstoječih je zelo veliko najrazličnejših orodij. V nadaljevanju podajamo le tri najbolj pogosto uporabljena, brezplačna orodja z različno domeno uporabnosti (Google forms, Kahoot, Quizizz) [11]. Google forms (obrazci) je najbolj razširjeno orodje, razvito za več operacijskih sistemov, za mobilne telefone in tablice. Z ustvarjenimi anketami, kvizi, nalogami lahko zelo hitro pridobimo vpogled v razumevanje in znanje učencev/dijakov/ študentov. Generiramo lahko vprašana izbirnega tipa, spustnih seznamov in linearnih lestvic. Obrazce lahko oblikujemo z dodajanjem slik, video posnetkov, logo tipi, kopiramo, premikamo, ustvarjamo odstavke, omogočimo logično preskakovanje vprašanj. Odgovori na izpolnjene obrazce se sproti, pregledno in samodejno zbirajo v obrazcih v obliki podatkov o odzivih in grafikonih. Zbrane podatke lahko dodatno analiziramo. Kahoot je orodje za sestavljanje kvizov. Najbolj je primeren za preverjanje razumevanja znanja med rednim poukom v šoli, ko učenci lahko med seboj tekmujejo v številu doseženih pravilnih odgovorov na novo spoznanih učnih vsebin in hitrosti odgovarjanja. Močna stran orodja je takojšnja povratna informacija učitelju. Za utrjevanje znanja so bolj primerni kvizi v drugih aplikacijah, npr. 1ka ali kvizi v Google obrazcih, kjer je več možnosti za izdelavo različnih vrst vprašanj in jih učenci rešujejo posamično. Quizizz omogoča enostavno sestavljanje različnih vrst vprašanj. Na voljo so obsežni arhivi primerov kvizov v angleškem jeziku, primerni za različna predmetna področja in različne stopnje izobraževanja. Izdelan kviz lahko delimo z učenci s kodo ali preko portalov kot: Google clasroom, Remind, Canvas, Schoology, MS Teams, Twiter...

(6) Orodja za objavo učnih vsebin na spletu. Za zbiranje in deljenje ustvarjenih gradiv lahko uporabimo različne spletne portale. Najbolj poznan je Youtube, po katerem je nedavno nastal tudi Arnes Video za izobraževalne namene [12]. Na tem mestu opišemo le orodje Padlet, ki omogoča »skupno rabo« npr. skupini učencev, ki lahko med seboj delijo zbrana gradiva (npr.: slike izdelkov, teste, video prispevke...). Učenci lahko »všečkajo« ali komentirajo deljena gradiva.

(7) Orodja za video spletna srečanja. Za izobraževalne namene sta se uveljavila zlasti dve orodji, MS Teams in Zoom [13], ki sta brezplačno razpoložljivi za izobraževalne namene v okrnjeni verziji. MS Teams je razvit za računalnike in za mobilne naprave. Udeleženci lahko izvajajo spletna srečanja, video konference (do 20), video in glasovne klice, klepetajo, imajo skupno rabo zaslona, izvajajo skupno rabo datotek, neposredno sporočanje, nastavljajo ozadja, možnost dodajanja ogromnega števila različnih aplikacij, sporočanje in sprejemanje nalog, ocenjevanja. Omogoča snemanje srečanj, pisanje po beli tabli, dodajanje različnih spletnih aplikacij in tudi druge možnosti. Orodje podira učenje na daljavo in ga lahko uporabimo tudi kot spletno učilnico. Drugo orodje je Zoom. Brezplačna verzija ima časovno omejitev video konferenčnega srečanja na 40 minut in omejitev števila udeležencev srečanja do 100. Plačljiva verzija omogočajo več funkcij kakor MS Teams, za bolj prijetno in učinkovito učenje na daljavo.

(8) Ostala orodja. Pod to skupino zajemimo še tri orodja. Prvo, Pinterest [14], je namenjeno iskanju idej za izdelavo učnih gradiv. Drugo, že omenjeno, Youtube, je orodje, ki ga učitelji radi uporabljajo za javno objavo svojih učnih gradiv. Zadnje orodje, We-transfer [15], je namenjeno prenosu večjih datotek. To so zlasti daljše video datoteke višje ločljivosti.

3 Pedagoška praksa na daljavo

V študijskem letu 2019/2020 smo izvedli v obliki učenja na daljavo pedagoško prakso za študente 3. (6. -10. 4. 2020) in 4. (14. -20. 4. 2020) letnikov vezave tehnika, smeri dvopredmetni učitelj na pedagoški fakulteti, Univerza v Ljubljani. Pedagoška praksa se je izvajala za skupaj 11 študentov na 11. različnih osnovnih šolah v različnih krajih Slovenije (Ljubljana, Prevalje, Braslovče, Bohinjska Bistrica, Novo mesto, Ajdovščina). V nadaljevanju podajamo analizo izvedbe pedagoške prakse v kontekstu spletnega učenja in uporabljenih orodij v navezavi na učni model na sliki 1 po posameznih sestavinah (1-8).

(1) Orodja za spletne učilnice. V devetih od skupaj 11 osnovnih šol, kjer so študentje opravljali pedagoško prakso na daljavo, so študenti pričakovano večinsko (82 %) uporabljajali Arnes učilnico, ki temelji na spletnem sistemu Moodle (AAI račun ustvarjen preko portala SIO.MDM) v preostalih dveh šolah pa so uporabili lastne spletne učilnice, tabela 1.

(2) Orodja za izdelovanje učnih gradiv. Za izvajanje učenja na daljavo so študenti izdelovali učna

Tabela 1. Preglednica uporabe orodij za izvajanje spletnega učenja, kjer pomeni V - video in Z - zvočno.

Orodje	Naziv/ (delež uporabe / %)
(1) Orodja za spletne učilnice	Arnes učilnica (82); lastne spletne učilnice (18)
(2) Orodja za izdelovanje učnih gradiv	PPT (82); Word (36)
(3) Orodja za zajem, obdelavo video (V) in zvočnih (Z) posnetkov	OBS studio (27); ShareX (18); Nvidia GeForce Experience (18); Bandicam (9); Loom (9); pametni telefon (9); Speechnotes - transkrip (9); Adobe Audition (Z) (9); Audacity (Z) (9); Diktafon (Z) (9); Adobe Premiere (V) (36); ShotCut (V) (9); Gimp (V) (9)
(4) Specifična tehniška orodja	Solid Edge (teh. dokumet.) (18); Solidworks (teh. dokumet.) (18); Edison (el. vezja) (27); Yenka (el. vezja) (18); Phet (el. vezja) (9); Crocodile clips (el. vezja) (9)
(5) Orodja za preverjanje znanja	Google forms (64); Kahoot (36); 1ka (9); Quizizz (9); Word (36)
(6) Orodja za objavo učnih vsebin na spletu	YouTube (83); Arnes Video (27); Padlet (18)
(7) Orodja za video spletna srečanja	Zoom (36); MS Teams (18)
(8) Druga orodja	We-transfer (prenos datotek) (9); Youtube (iskanje vsebin/idej) (64); Pinterest (iskanje vsebin/idej) (9)

gradiva za učence. Način izdelave in vrsta uporabljenih orodij je bila odvisna od teme. Največkrat (82 %) so uporabili orodje MS PPT v katero so namestili razne tekstovne dele, prilepili slike, dodajali video in zvočne posnetke, pripravili animacije in vse skupaj didaktično smiselno interaktivno povezali v celoto. Tako pripravljene predstavitve so večinoma posredovali učencem, ki so lahko samostojno, z lastnim tempom pregledovali nove učne vsebine. V 46 % so izdelano PPT predstavitev izvozili kot video posnetek in slednjega posredovali učencem. Tudi video posnetek so lahko učenci po potrebi ustavljali/ponovno ogledovali. Manj študentov, 36 %, je izdelovalo učne vsebine z Word-om. Največkrat v primeru izdelovanja vsebinskih povzetkov učne snovi in delovnih listov za določanje poteka uvajanja/usvajanja novih učnih vsebin.

(3) Orodja za zajem, obdelavo video (V) in zvočnih (Z) posnetkov. Za izdelovanje učnih gradiv so se študenti posluževali številnih prosto dostopnih orodij, ki so jih spoznali tekom študija, ali pa so jim jih svetovali mentorji. Največji delež študentov, 36 %, je za snemanje zvoka (in zajem zaslona) uporabilo OBS studio. Ena študentka je posnela zvok s pametnim telefonom in ga naknadno obdelala še z orodjem Audacity. En študent je posnel govor z diktafonom in ga naknadno obdelal z Adobe Audition. Za snemanje

Tabela 2. Preglednica uporabe spletnih orodij s strani učiteljev na Irskem, kjer pomeni V - video in Z - zvočno [16].

Orodje	Naziv/ (delež uporabe / %)
(1) Orodja za spletne učilnice	SeeSaw (30); ClassDojo (41); Schoology (1); Google Classroom (18); VS Ware (1)
(2) Oro. za izd. uč. gradiv	PPT (49);
(3) Orodja za zajem, obdelavo V/Z posnetkov	Screencastify (1)
(6) Orodja za objavo učnih vsebin na spletu	SeomraRanga (49); Twinkl (90); ScoilNet (71);
(7) Oro. za vid. srečanja	MS Teams (5,8); Zoom (15);
(8) Druga orodja	One Note (2)

celotnega ali delnega zaslona je 18 % študentov uporabilo orodje ShareX in Nvidia GeForce Experience Posamezni študenti so video posnetke pripravljali s programom Bandicam in Loom. En študent je originalni video posnetek v angleškem jeziku s transkriptom in prevodom (Google translate) posnel v slovenskem jeziku. Slaba polovica študentov, 45 %, je zajete posnetke nadalje obdelovala, večinoma v Adobe Premiere in eden študent v ShotCut-u.

(4) Specifična tehniška orodja. V primeru učnih tem iz tehniške dokumentacije (2 študenta) sta oba študenta uporabila dve različni orodji, Solid Edge, ki so ga študentje spoznali v času študija in Solidworks (samostojno ali po priporočilu mentorja). V primeru učnih tem vezanih na električna vezja (električni krog, zaporedna in vzporedna vezava stikal, elektrotehniške sheme) je največ študentov, 27 %, uporabilo Edison, in Yenk-o, 18 %. En študent je uporabil Phet on-line simulacijo enosmernega električnega toka – virtualni laboratorij, drugi študent pa orodje Crocodile clips.

(5) Orodja za preverjanje znanja. Za preverjanje usvojenega znanja in razumevanja iz podanih učnih gradiv je 64 % študentov izdelovalo kvize/vprašalnike z Google obrazci. 36 % študentov je izdelalo kviz za tekmovanje učencev v znanju z orodjem Kahoot. Po en študent je izdelal vprašalnik z 1ka in s Quizizz-om.

(6) Orodja za objavo učnih vsebin na spletu. Večina študentov, 73 %, je objavilo izdelana učna gradiva na spletnem portalu YouTube. 27 % študentom so mentorji omogočili objavo na spletnem portalu slovenskih izobraževalnih video vsebin, Arnes Video. En študent je uporabil orodje Padlet.

(7) Orodja za video spletna srečanja. Možnost srečanja študentov z učenci preko video konferenc so imeli le nekateri študenti, od teh jih je 36 % uporabilo Zoom in 18 % MS Teams.

(8) Ostala orodja. Izmed preostalih orodij se pojavi orodje za prenos večjih datotek, We-transfer, 1 študent. Za primere in zamisli učnih vsebin je večina študentov iskalo na Youtube, 83 % študentov, preko portala Pinterest pa le en študent.

4 Diskusija in zaključek

Z uvedbo pandemije, 16. 03. 2020, v Sloveniji ni bilo podanih nobenih jasnih smernic in vodil učiteljem za izvajanje učenja na daljavo (na vseh stopnjah). Učenje na daljavo je bilo za vse udeležene osnovnošolske učence in učitelje mentorje novo. Vsi mentorji so izvajali pouk tehniških predmetov po učnem modelu na sliki 1. Kljub prvemu primeru takšnega izvajanja pouka so spodbujali izvedbo pedagoške prakse.

Pri učno manj zmožni učencih in učencih s posebni potrebami so zaznali težje sledenje spletnemu učenju. To se je odražalo v manjši stopnji motivacije za učenje. Pogrešali so vodenje in dodatno razlago. Le nekateri od teh učencev so imeli podporo s strani družine. Medtem ko je študentom samostojen študij iz literature samoumeven, je bilo to za učence nekaj novega. Učno zmožnejši učenci so imeli pri tem manj težav. Učenje na daljavo za njih ni predstavljalo takšne ovire, kot drugim učencem, ki že v normalnih okoliščinah potrebujejo več razlage, zlasti pa učencem s posebnimi potrebami. Vsi učenci niso imeli enakega dostopa do informacijsko-komunikacijskih tehnologij. Nekateri učeneci so bili prepuščeni sami sebi, nekateri so imeli pomoč družine. Za pomembno se je izkazalo, da so učna gradiva za učence podana v slovenskem jeziku. Pri preverjanju znanja so bila vprašanja/naloge s slikovno podporo boljše razumljena in rešena.

Tako pri predmetu tehnika in tehnologija kot tudi pri skoraj vseh drugih predmetih so učenci dobivali domače naloge. Pri oddajanju nalog je bila zaznana precejšnja zmeda. Ker je potekala pedagoška praksa v začetnem delu izvedbe učenja na daljavo je bilo pri izvajanju še veliko sprememb. Učenci so morali sprva pošiljati domače naloge preko e-pošte, nato so jo oddajali v šolsko spletno učilnico, spet drugič v Google učilnico. Pri tem se je pojavljalo, da učenci niso bili gotovi, kaj morajo oddati in ali so sploh uspešno oddali.

V primeru praktičnega dela, ki je specifično za tehniko, so le tega prelevili v izvajanje delovnih nalog za izdelavo izdelkov glede na razpoložljivost gradiv in orodij, s katerimi so lahko razpolagali učenci na svojem domu. Objava in delitev slik oz. video posnetkov izdelanih izdelkov je opazno dvigovala motivacijo učencev. Pri učencih, ki so razpolagali z zelo dobro opremljeno domačo delavnico je prišlo do sodelovanja s starši in bistveno boljših izdelkov.

Slovenski učitelji so bili bistveno bolj prisiljeni v izdelavo učnigh gradiv, kakor naprimer angleško govoreči učitelji na Irskem, tabela 2. Ti so si večinsko pomagali z obsežno bazo že izdelanih gradiv. Prav tako so imeli že po enem mesecu uvedbe pandemije (11. 3. 2020) obsežne delovne smernice [16].

Z vidika udeleženih študentov so mnenja, da pouk na daljavo nikoli ne bo nadomestil pristnega stika z učenci, hkrati pa so spoznali dejansko pomembnost in uporabnost tehnike in tehnologije na vseh možnih področjih v naših življenjih, katere se po njihovem mnenju ne zavedamo dovolj. V okviru pedagoške prakse smo prvič zaznali nastanek sodelovanja med študenti in mentorji, saj so bili študenti tehnološko in IKT bolj pismeni. Posledica takšnega dela se je izkazala v višji kvaliteti izdelanega učnega gradiva, ki je imelo v nastali situaciji ključno vlogo.

Mentorji in študenti so se z izvajanjem učenja na daljavo zavedli prednosti, ki jih prinaša takšen način, hkrati pa tudi nujnosti v posodabljanju tradicionalnega poučevanja. Le s sodobnimi učnimi strategijami dobi sodobna izobraževalna tehnologija smisel in neizmeren potencial motivacije učencev/dijakov/študentov. Slednje želimo, da je vodilo sodobnemu učitelju ne glede na raven tehniškega izobraževanja.

Literatura

[1] E. K. Kahiigi, Exploring the e-Learning State of Art, *The Electronic Journal of e-Learning*, 6 (2), str. 77 - 88, 2008; M. G. Moore , *Handbook of distance education* (New York, Routledge, 2013).

[2] C. J. Bonk in C.R. Graham. *The handbook of blended learning: Global perspectives, local designs* (New Jersey, John Wiley & Sons, 2005).

[3] J. Kemeny in T. Kurtz, *The Dartmouth Time-Sharing System* (Washington, National Science Foundation, 1967).

[4] M. Deimann in N. Friesen, Exploring the Educational Potential of Open Educational Resources, *E-Learning and Digital Media*, 10 (2), 112-115, 2013.

[5] V. Singh in A.Thurman, How Many Ways Can We Define Online Learning? A Systematic Literature Review of Definitions of Online Learning (1988-2018). *American Journal of Distance Education* 33 (4), str. 289-306, 2019.

[6] Huang, R.H. in ostali, *Handbook on facilitating flexible lear. during edu. disruption: The Chinese experience in main. undisrupted lear. in COV.-19 outbreak* (Beijing: SLIBNU, 2020).

[7] D. Rihtaršič in J. Jamšek, *Izzivi in priložnosti uporabe IKT v ped. proc. na področju naravoslovja, tehnol. in mat.: Razvoj gradiva za kombinirano učenje pri izvajanju predavanj in kliničnih vaj pri predmetu odprti Odprti učni sistemi v tehniki* (Ljubljana, Univerza v Ljubljani, 2019).

[8] new.edmodo. com; beenpod. com; www.goclass. com/; www.schoology. com; learn.canvas. net; classroom.google.com.

[9] obsproject. com; www. nvidia. com; www. bandicam. com; speechnotes. co; www. adobe. com; www.audacityteam.org.

[10] www.cicicad.si; qcad.org; www.freecadweb.org; www.sketchup.com; solidedge.siemens.com; http://www.edisonlab.com; crocodile-clips1. software.informer.com/3.5/, www.yenka.com; phet.colorado.edu.

[11] docs.google.com; kahoot.com; quizizz.com.

[12] www.youtube.com, video.arnes.si.

[13] teamsdemo.office.com, zoom.us.

[14] www.pinterest.com.

[15] wetransfer.com.

[16] J. Burke in M. Dempsey, Technical report: *Covid-19 Practice in Primary Schools in Ireland Report* (Dublin, Maynooth University, 2020), [www.researchgate.net/publication/340528449].

Študentski članki

Student papers

Time series classification using time-frequency analysis and Convolutional Neural Networks

Domen Kavran

Faculty of Electrical Engineering and Computer Science, University of Maribor
Koroška cesta 46, 2000 Maribor, Slovenia
E-mail: domen.kavran@student.um.si

Abstract

Technological advances in various industries have led to an increase in the amount of data in which patterns, trends and useful informations are hidden. Data are quantitatively large, complex and generated at high speed, which makes them impossible to process with traditional methods. A common kind of sequential data are time series, where time order is important. Regression applications are often based on time series, but they can also be used for classification purposes. In recent years, deep learning has proven to be very successful in classifying complex time series data. In this paper, a new method for time series classification is presented using deep learning. Different time-frequency analysis methods are applied on time series in order to obtain time-frequency representations, which serve as input data to convolutional neural networks. By combining various time-frequency representations, the proposed method achieved average classification accuracy of 92,28 %.

1 Introduction

Data collection is being introduced into different areas of industries where this has not been done before. Multitudes of different sensors with high sampling frequencies gather complex data. Such an environment produces large volumes of varied data that are obtained at high velocity. This kind of data is named "big data". Important types of big data are time series, which represent values of a given quantity over time. Physical quantities, such as electric current, temperature, and velocity are observed most commonly [1]. Time series are obtained from different sources, ranging from mobile phones and Internet of Things (IoT) devices to industrial machines and medical equipment [2]. Acquired time series can be noisy and may have information gaps. Sampled time series is treated as a signal consisting of complex sinusoids. These oscillate with different frequencies. Each sinusoid reaches a certain amplitude and has an offset, so-called phase. Sinusoids are present at different time points in a signal. Its amplitude, phase and energy density in time and frequency are analyzed using time-frequency analysis methods, which are presented in the following Sections.

Various methods calculate time-frequency representations, which are limited by the ratio of time and frequency resolution. Time series are, thus, represented as two-dimensional images in the time-frequency plane, where both axes represent time t and frequency f, with signal properties being expressed in grayscale shades. Time-frequency representations are used as input data to convolutional neural networks to perform classification. Most of the time-frequency representation algorithms have long been developed, but they have recently been introduced to the field of Deep Learning as part of the data preparation process [3]. The challenges of time series classification are present in many research areas - from identification of anomalies in financial markets to the automated recognition of brain and cardiovascular diseases [2].

Feature selection is a challenging process, as domain knowledge is often needed. An alternative approach is offered by Convolutional Neural Networks (CNNs), which acquire information-rich features from multidimensional input data using an array of adaptive convolutional filters, called kernels [4]. Many methods for time series classification using CNNs have been proposed throughout the years. Residual neural network (ResNet) is currently considered to be the state-of-the-art deep neural network architecture for time series classification [5]. A recently developed ensemble of five deep CNNs, named InceptionTime, has proven to be very competitive with the state-of-the-art algorithm HIVE-COTE [5].

The proposed CNN architectures use two-dimensional and three-dimensional convolutional kernels. Two-dimensional kernels calculate features in time-frequency space. By combining various time-frequency representations, different local characteristics of the signal are used for three-dimensional feature extraction.

This paper consists of five sections. The next section presents data preparation using Hilbert transform and time-frequency analysis methods. The third section presents CNNs and proposed architectures for time series classification. Classification results are presented in the fourth section. The conclusions are given in the last section.

2 Time series analysis

The analysis of a time series, treated as a signal x, is performed using mutually independent methods and mathematical operations performed in a sequence called a pipeline. Each pipeline consists of several steps. The output of the i-th step is the input to the i+1-th step. The last step is an exception, as its output is the result of the pipeline. The proposed pipeline has the following steps:

1. Calculation of the analytic signal x_a using the Hilbert transform,
2. Calculation of phase or energy density by using one of the following time-frequency analysis methods:
   - Short-Time Fourier Transform,
   - Smoothed Pseudo Wigner-Ville Distribution,
   - Continuous Wavelet Transform,
3. Normalization of values to the interval $[0, 1]$.

Use of an analytic signal makes a frequency spectrum more interpretable, but its use is also mandatory for some time-frequency analysis methods [6]. Phase and energy density are unique signal characteristics in time and frequency. Normalization, as a final step, ensures CNNs learn faster, with input values having equal scales without distortions. Methods in pipeline are described in the following subsections.

2.1 Hilbert transform

Hilbert transform performs phase manipulation without affecting the amplitude spectrum of the signal. The Hilbert transform $\mathcal{H}$ of a signal x is defined by (1), where $h = \frac{1}{\pi t}$ and the Cauchy principal value is denoted by P [6]. The transform is used to calculate an analytic signal that is complex and does not contain negative frequency content. The analytic signal is written by (2) [6].

$$x_H(t) = \mathcal{H}[x(t)] = x(t) * h(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau \quad (1)$$

$$x_a(t) = x(t) + i x_H(t) \quad (2)$$

2.2 Short-Time Fourier Transform

Short-Time Fourier Transform (STFT) is a time-frequency analysis method which calculates Fourier transform on segments of a non-stationary signal. Multiplying segments with window function w reduces spectral leakage. Short-Time Fourier Transform has a constant frequency and time resolution that depends on the length of the segments. Computing squared amplitude is represented by a spectrogram, which shows energy density in time and frequency. Short-Time Fourier Transform is defined by (3) [7].

$$STFT_x^w(t, f) = \int_{-\infty}^{\infty} x(\tau) w^*(\tau - t) e^{-i2\pi f \tau} d\tau \quad (3)$$

Short-Time Fourier Transform was demonstrated on a recording of a bird chirp, that was sampled at 22050 Hz. Hann window function was applied on signal segments, which overlapped by half of their length. The spectrogram of the bird chirp is shown in Figure 1 and the thresholded phase spectrum in time and frequency is shown in Figure 2.

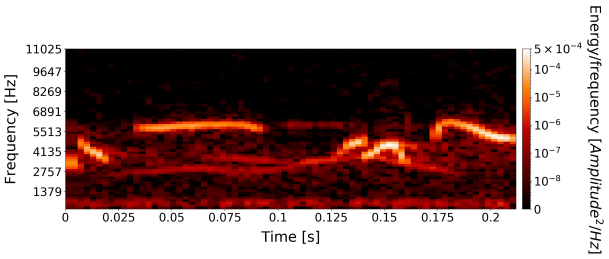


Figure 1: Spectrogram of the bird chirp.

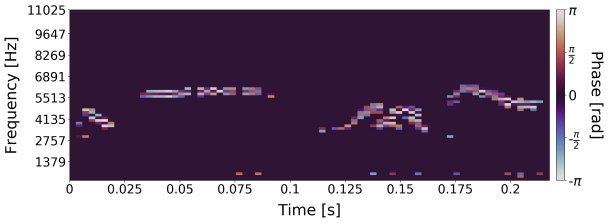


Figure 2: Thresholded phase spectrum in time and frequency of the bird chirp.

2.3 Smoothed Pseudo Wigner-Ville Distribution

Among many methods for time-frequency analysis, Wigner-Ville Distribution is considered the most advanced. It represents energy density, the same as a spectrogram, but with higher resolution, although artifacts called cross terms are present. Smoothed Pseudo Wigner-Ville Distribution (SPWVD) reduces unwanted cross terms with time-frequency filtering; however, time and frequency resolution are deteriorated. Smoothed Pseudo Wigner-Ville Distribution is defined with (4), where $g(s)$ and $h(\tau)$ are window functions for smoothing in time and frequency [8]. The result of Smoothed Pseudo Wigner-Ville Distribution on a recording of a bird chirp is shown in Figure 3.

$$SPWVD_x(t, f) = \int_{-\infty}^{\infty} h(\tau) \left[\int_{-\infty}^{\infty} g(s-t)\, x(s + \frac{\tau}{2}) \right.$$
$$\left. x^*(s - \frac{\tau}{2})\, ds \right] e^{-i2\pi f \tau} d\tau \quad (4)$$

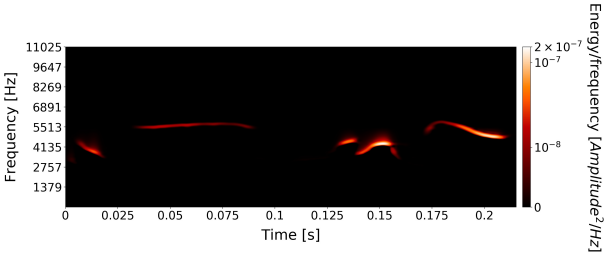


Figure 3: SPWVD of the bird chirp.

2.4 Continuous Wavelet Transform

Continuous Wavelet Transform (CWT) is an alternative to Short-Time Fourier Transform. An important property of Continuous Wavelet Transform is that it does not use Fourier transform. Instead, convolution of the signal is

performed using a function, called a wavelet. Time localization is carried out with translation τ of mother wavelet ψ. Its length is adjusted with scale s, thus effecting time and frequency resolution. Computing squared amplitude is represented by a scalogram, which shows energy density in time and scale. Continuous Wavelet Transform is defined by (5) [9]. Results of Continuous Wavelet Transform with a complex Morlet wavelet of a bird chirp are visible on Figures 4 and 5.

$$CWT_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t)\, \psi^*\left(\frac{t-\tau}{s}\right) dt \qquad (5)$$

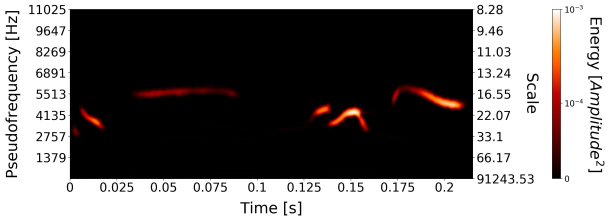


Figure 4: Scalogram of the bird chirp.

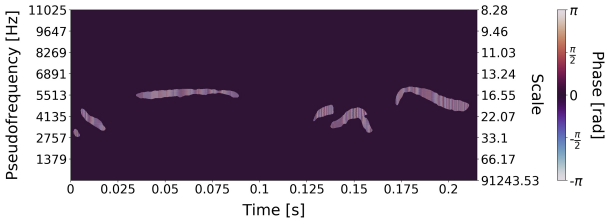


Figure 5: Thresholded phase spectrum in time and scale of the bird chirp.

3 Convolutional Neural Networks

The most established algorithms in the field of Deep Learning are CNNs, which are designed to learn spatial hierarchies of features automatically by using convolutional layers, pooling layers and fully connected layers. They are often used to solve computer vision problems in industrial processes and various fields of medicine. CNNs achieve exceptional results at segmentation and recognition of objects in images and at processing natural language [4].

3.1 Proposed CNN architectures

Figures 6 and 7 show the proposed CNN architectures for classification of time series of length N. The CNN-I architecture is designed for time-series classification, based on individual time-frequency representations. The CNN-C architecture performs time series classification with combined time-frequency representations, used as input. The input sizes $W \times H$ represent the width and height of the time-frequency representations, and the input sizes $N \times N \times 5$ are intended for combined time-frequency representations. The hyperparameter K represents the number of convolutional kernels, and F_p the size of pooling filters. Convolution layers and pooling layers use zero padding. The number of neurons in a fully connected layer is defined by the hyperparameter N_{fc}. The number of neurons in the output layer is equal to the number of classification classes N_C [4].

4 Results

Classification was performed on the datasets with pre-segmented time series described in Table 1 [10, 11]. All datasets (except Epileptic Seizure Recognition) had a predetermined train and test set. For individual time series dataset, five pipelines were defined to calculate the following normalized outputs for each time series of length N:

- Phase spectrum as a result of STFT,
- Spectrogram,
- Phase spectrum as a result of CWT,
- Scalogram,
- Smoothed Pseudo Wigner-Ville Distribution.

The output of each pipeline is a time-frequency representation of the time series, which is then used as input data to one of five CNNs having CNN-I architecture. Because the sizes of representations are not equal, they need to be scaled to size $N \times N$ with nearest neighbors interpolation before combining them and forwarding them to a CNN having CNN-C architecture. The hyperparameters of each CNN were adjusted based on 10-fold cross validation.

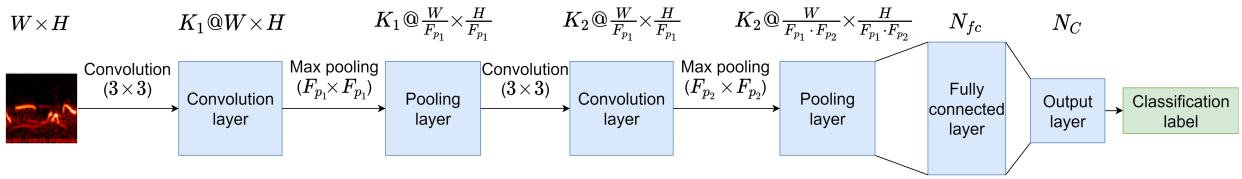


Figure 6: Architecture CNN-I for time series classification, using individual time-frequency representations as the input.

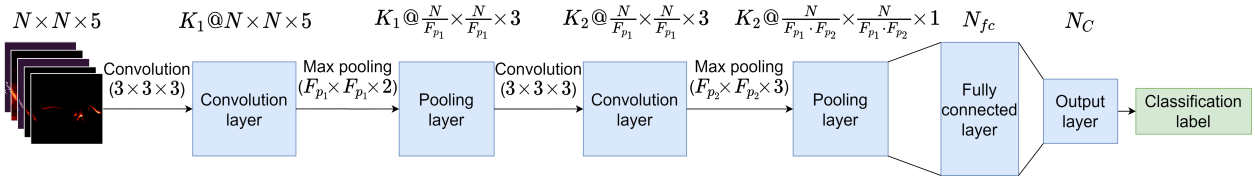


Figure 7: Architecture CNN-C for time series classification, using combined time-frequency representations as the input.

Table 1: Time series datasets.

Dataset name	Description	Time series length	No. of classes
ECG5000	ECG recordings of 15 patients with five categorizations of cardiovascular diseases	140	5
InsectEPGRegularTrain	EPG signals of insect interaction with plants	601	3
Epileptic Seizure Recognition	Sections of EEG recordings of 500 individuals with epilepsy attacks	178	5
MelbournePedestrian	Hourly number of pedestrians at ten locations in the city of Melbourne	24	10
ElectricDevices	Electricity consumption of seven groups of appliances in UK households	96	7
Wafer	Control measurements of sensors in the processing of silicon wafers	152	2
PowerCons	Electricity consumption of households in summer and winter	144	2

The results of the proposed classification method are shown in Table 2. Each column contains classification accuracies, obtained with selected time-frequency representations, and used as inputs to CNNs with a chosen architecture (CNN-I or CNN-C). The highest classification accuracy of each dataset is marked with bold text. Using individual time-frequency representations as inputs to CNNs with architecture CNN-I, the highest classification accuracies in all datasets were achieved with phase spectrums or scalograms, obtained with Continuous Wavelet Transform.

Convolutional Neural Networks with architecture CNN-C, that performed classification based on combined time-frequency representations, achieved higher classification accuracies of datasets ECG5000, InsectEPGRegularTrain, Epileptic Seizure Recognition and MelbournePedestrian.

Table 2: Time series classification accuracies.

Dataset name	STFT - phase spectrum (CNN-I)	Spectrogram (CNN-I)	CWT - phase spectrum (CNN-I)	Scalogram (CNN-I)	SPWVD (CNN-I)	Combined representations (CNN-C)
ECG5000	93,29%	92,98%	93,69%	93,51%	93,33%	**94,27%**
InsectEPGRegularTrain	95,58%	83,13%	99,60%	99,60%	83,13%	**100,00%**
Epileptic Seizure Recognition	91,04%	92,35%	94,87%	96,70%	95,78%	**97,39%**
MelbournePedestrian	61,96%	63,56%	83,22%	75,30%	68,36%	**88,39%**
ElectricDevices	57,23%	65,27%	61,47%	**69,14%**	63,91%	68,28%
Wafer	98,96%	99,04%	99,43%	**99,53%**	98,60%	99,33%
PowerCons	91,11%	**98,89%**	96,11%	**98,89%**	96,67%	98,33%

5 Conclusion

In this paper, a new method for time series classification was presented, where data preparation pipelines compute time-frequency representations, which are then used as input to Convolutional Neural Networks for classification. Convolutional Neural Networks were most successful at classifying phase spectrums and scalograms, which were calculated with Continuous Wavelet Transform, achieving an average accuracy of 90,07 %. Classification accuracies of some time series datasets have been improved by combining different time-frequency representations, achieving an average accuracy of 92,28 %. In the future, the proposed Convolutional Neural Network architectures will be adapted to perform multivariate time series classification.

Acknowledgment

I thank Assist. Prof. Dr. Niko Lukač for guidance and professional assistance in writing my Master's thesis, which was the basis for this paper.

References

[1] N. Elgendy and A. Elragal. Big Data Analytics: A Literature Review Paper. *Advances in Data Mining: Applications and Theoretical Aspects*. 14th Industrial Conference on Data Mining, St. Petersburg, 16.–20. July 2014. Leipzig: Springer, 214–227, 2014.

[2] G. A. Susto, A. Cenedese and M. Terzi. Time-Series Classification Methods: Review and Applications to Power Systems Data. *Big Data Application in Power Systems*. Amsterdam: Elsevier, 179–220, 2018.

[3] J. Zhang, J. Tian, Y. Cao, Y. Yang and X. Xu. Deep time–frequency representation and progressive decision fusion for ECG classification. *Knowledge-Based Systems*, 190, 2020.

[4] R. Yamashita, M. Nishio, R. K. G. Do and K. Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 2018.

[5] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, et al. InceptionTime: Finding AlexNet for Time Series Classification, 2019.

[6] M. Johansson. *The Hilbert Transform*. Växjö: Växjö University, 2005.

[7] F. Hlawatsch and F. Auger. *Time-Frequency Analysis: Concepts and Methods*. London: ISTE, 2008.

[8] A. Djebbari and F. Bereksi-Reguig. Detection of the valvular split within the second heart sound using the reassigned smoothed pseudo Wigner–Ville distribution. *BioMedical Engineering OnLine*, 12(1):1–21, 2013.

[9] L. Aguiar-Conraria and M. J. Soares. The Continuous Wavelet Transform: moving beyond uni and bivariate analysis. *Journal of Economic Surveys*, 28(2):344–375, 2014.

[10] A. Bagnall, E. Keogh, J. Lines, A. Bostrom and J. Large. UEA & UCR Time Series Classification Repository, 2020.

[11] D. Dua and C. Graff. UCI Machine Learning Repository. Irvine: University of California, School of Information and Computer Science, 2020.

Avtomatska segmentacija Golgijevih aparatov v volumetričnih podatkih z grobo označenim zlatim standardom

Eva Boneš

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
eb1690@student.uni-lj.si

Automatic Segmentation of Golgi Apparatus in Volumetric Data with Roughly Annotated Ground Truth

Golgi apparatus (GA) is a cellular organelle involved in the processing and sorting of proteins in all eukaryotic cells. Due to its numerous functions, structural complexity and organizational dynamics, the role of GA in normal and pathological processes is still under intensive research. Electron microscopy enables direct visualization of GA, yet only recent development of dual beam microscopes (i.e. focused ion beam combined with scanning electron microscope) enabled its organizational studies within relatively large volume of the cell. The bottle neck of the process is manual GA segmentation in the volumetric data. In this work, we present a tripartite approach to automatic segmentation of GA based on a pipeline consisting of i) neural network trained on roughly annotated data, ii) active contours for a more precise segmentation and iii) filtering of false positive results. Using the pipeline on the volume derived from urinary bladder epithelial cell, all of GAs were marked, with only a few false positives where stacks of fusiform vesicles were detected as GA. The proposed algorithm is able to annotate a wast majority of GAs with 89% sensitivity and 99% specificity and 20% of time needed for manually annotating the input data.

1 Uvod

Evkariontske celice so celice, ki gradijo vse mnogocelične organizme. Za razumevanje njihovega delovanja moramo natančno poznati strukturo ter funkcijo številnih celičnih organelov v njihovi notranjosti, vključno z njihovo prostorsko razporeditvijo in velikostjo.

V naši raziskavi se osredotočamo na urotelijske celice, ki sestavljajo urotelij – specializirano krovno tkivo, ki pokriva sečni mehur. V teh celicah se nahajajo fuziformni vezikli (FV), ki vsebujejo, shranjujejo in prenašajo urotelijske plake, ki so pomemben faktor pri vzpostavitvi krvno-urinske pregrade sečnega mehurja sesalcev. Še vedno ni povsem jasno kako nastajajo, dokazano pa je bilo, da pri tem sodeluje tudi Golgijev aparat (GA).

GA je organel biosintetske poti, kjer potekata funkcionaliziranje in razvrščanje membranskih proteinov, hkrati pa predstavlja križišče ekso- in endocitotske poti v celici. Poznavanje njegove prostorske razporeditve, ki jo pridobimo s segmentacijo v volumetričnih podatkih, bi pripomoglo k boljšemu razumevanju temeljnih procesov delovanja celic.

Razvoj nove tehnologije za zajemanje volumetričnih mikroskopskih slik z metodo usmerjenega ionskega snopa in vrstične elektronske mikroskopije (FIB-SEM, angl. *focused ion beam combined with scanning electron microscopy technique*) je omogočil neposredno opazovanje GA-jev in možnost njihovega raziskovanja v večjem volumnu celice.

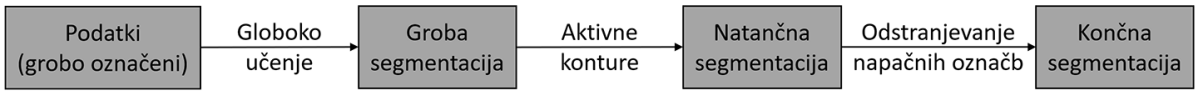
Ročna segmentacija tovrstnih tridimenzionalnih volumnov poteka z občrtavanjem organela na vsaki od rezin, kar zahteva izurjenega uporabnika in predstavlja veliko časovno obremenitev. V članku zato predstavljamo postopek za avtomatsko segmentacijo GA-jev v volumetričnih podatkih elektronske mikroskopije, ki uporablja grobo označen zlati standard za učenje globokega segmentacijskega modela, s čimer smo pohitrili pridobivanje učnih podatkov.

Problem segmentacije v izbrani vrsti celic je posebej kompleksen, saj so FV-ji v citoplazmi diferenciranih urotelijskih celic organizirani v skladovnice, ki so po teksturah zelo podobne skladovnicam GA-jev. Zato kljub temu, da smo pristop ovrednotili samo na tej vrsti celic, pričakujemo, da bi bil predlagan postopek dovolj splošen in uporaben tudi za segmentacijo GA-jev v drugih vrstah celic.

1.1 Segmentacija biomedicinskih slik

Z razvojem novih metod za zajem biomedicinskih slik se hitro razvija tudi analiza le-teh. Ker je ročna segmentacija volumetričnih podatkov časovno zelo zahtevna, raziskovalci razvijajo avtomatske metode, ki segmentacijo pohitrijo. Kot na vseh področjih analize slik večina trenutnih najsodobnejših pristopov za segmentacijo biomedicinskih slik temelji na globokih konvolucijskih nevronskih mrežah.

Prve sodobne metode, kot sta FCN [1] in U-Net [2], so bile ustvarjene za segmentacijo 2D podatkov, z novimi načini zajemanja slik pa se je kmalu pokazala potreba po izkoriščanju informacij, prisotnih v 3D volumetričnih podatkih. V ta namen so kmalu razvili več arhitektur, vsaka izmed njih je bila razvita za svojo vrsto podatkov, npr.: 3D U-Net kot nadgradnjo U-Net [3] so razvili za slike ledvic, V-Net [4] za segmentacijo prostate, DeepMedic



Slika 1: Predlagani pristop k avtomatski segmentaciji Golgijevih aparatov

[5] za segmentacijo lezij v možganih.

Vse omenjene arhitekture na svojih področjih dosegajo vrhunske rezultate, kar pa ne pomeni, da najboljše rezultate dosegajo na vseh vrstah podatkov. Zato je pri razvoju novih pristopov pomembno, da upoštevamo lastnosti ciljnega področja [6]. Nedavno je bila za segmentacijo mitohondrijev in endolizosomov v volumetričnih podatkih elektronske mikroskopije predstavljena arhitektura HighRes3DZMNet [7], ki kot nadgradnja HighRes3DNet [8] dodaja uporabo kontrastnih konovolucijskih filtrov na prvih nivojih konvolucijske nevronske mreže za izenačevanje svetlosti vhodnih podatkov, kar je pomembno predvsem pri mikroskopskih slikah, pri katerih hitro pride do svetlobnih artefaktov. Čeprav je na področju segmentacije celičnih organelov tovrstnih namenskih arhitektur veliko, pa pristopa za avtomatsko segmentacijo GA-jev v volumetričnih podatkih po nam znanih podatki še ni.

V tem članku zato predlagamo nov pristop k reševanju tega problema, ki poleg avtomatske segmentacije tudi pohitri označevanje vhodnih podatkov, kar je uporabno za segmentacijo vseh vrst vhodnih podatkov.

2 Segmentacija

Globoke arhitekture za učenje navadno potrebujejo veliko množico označenih podatkov. Zaradi velike raznolikosti GA-jev znotraj celice in zaradi njihovega razgibanega roba je njihovo natančno ročno označevanje v volumetričnih podatkih časovno zelo zahtevno, prav tako pa zelo hitro pride do napak.

Zato predlagamo pristop v treh korakih. V prvem koraku globoko arhitekturo nadzorovano učimo na grobo označenih podatkih, kar nam kot rezultat vrne model, ki je sposoben avtomatsko grobo označiti območja, kjer se nahajajo GA-ji. Na teh podatkih nato uporabimo metodo aktivnih kontur za pridobitev bolj natančnih robov GA-jev. Na koncu iz dobljene segmentacije še odstranimo napačne označbe. Pristop je v celoti avtomatski in je prikazan na Sliki 1.

Posamezni deli predlaganega pristopa so bolj podrobno opisani v naslednjih podpoglavjih.

2.1 Podatki

Mikroskopski podatki, ki smo jih uporabili, so del podatkovne baze UroCell [7]. Pridobljeni so bili z elektronskim mikroskopom FIB-SEM in prikazujejo krovne celice iz epitelija mišjega mehurja. Dimenzija enega voksla je približno $16 \times 16 \times 15\ nm$. Celoten volumen je bil razdeljen na manjše dele velikosti $256 \times 256 \times 256$ vokslov, tako da vsak izmed podvolumnov zavzema približno $4 \times 4 \times 4\ \mu m$. Slika 2a prikazuje enega izmed podvolumnov.

GA-je smo ročno označili z odprtokodnim orodjem Slicer3D [9]. Ročne oznake smo naredili za 9 podvolumnov. Na Sliki 3a je vidno, da imajo GA-ji razgiban rob, ki smo ga mi le grobo obkrožili. Vse označbe je pregledal strokovnjak s področja celične biologije.

2.2 Globoko učenje

Za segmentacijski model smo uporabili globoko arhitekturo HighRes3DZMNet [7], ki je bila razvita na podatkih, ki jih za segmentacijo GA-jev uporabljamo tudi mi. HighRes3DZMNet je konvolucijska nevronska mreža, sestavljena iz 20 slojev, ki na izhodu določijo pripadnost posameznega voksla ciljnim razredom.

Za učenje smo uporabili 9 podvolumnov, ki smo jih razdelili na učno (7) in testno (2) množico. Volumni so vsebovali od 0 do 3 GA-je. Ker HighRes3DZMNet uporablja obogatitev podatkov, se je, kljub sicer majhni učni množici, izkazalo, da so rezultati zadovoljivi.

2.3 Aktivne konture

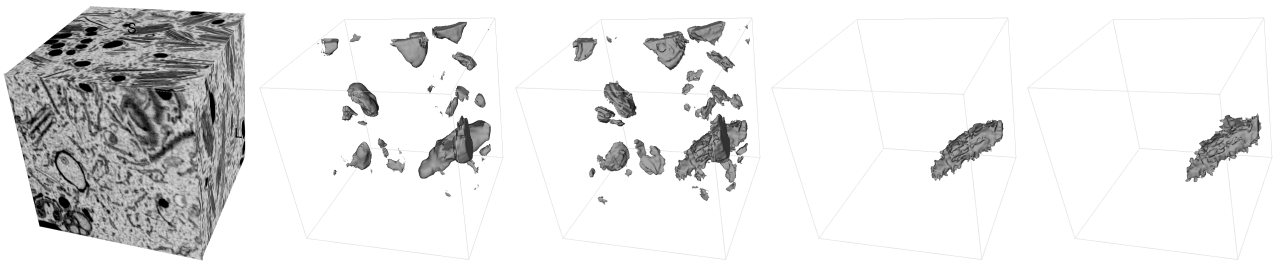
Aktivne konture [10] se pogosto uporabljajo pri obdelavi medicinskih slik, predvsem za segmentacijo in določanje meja objektov. Algoritem glede na prvotno oceno meje objekta to mejo iterativno prilagaja, pri čemer minimizira energijsko funkcijo in se s tem približuje dejanski meji.

Ker metoda deluje na 2D podatkih, jo izvedemo na vsaki izmed rezin volumetričnih podatkov (po vseh treh dimenzijah), kjer za prvotno oceno meje objekta uporabimo grobo mejo segmentiranih GA-jev, pridobljeno z globokim modelom. Na koncu združimo podatke o novih mejah tako, da voksle, ki so bili vsaj na eni izmed dimenzij označeni kot GA, označimo kot GA tudi na končni segmentaciji.

Pristop se je izkazal kot primeren za reševanje našega problema. Kot je vidno na Sliki 3, pristop dobro deluje tudi v primeru, ko je del začetne obrobe znotraj objekta, ki ga želimo obkrožiti. Z minimiziranjem energije se namreč na tistem mestu obroba razširi navzven.

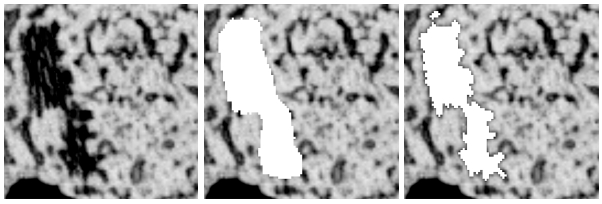
2.4 Odstranjevanje napačnih označb

V urotelijskih celicah se nahajajo tudi FV-ji, ki na mestih, kjer so združeni v sklade, teksturno izgledajo zelo podobno kot GA-ji, kar se vidi na Sliki 4a. Posledično predlagana metoda proizvede tudi nekaj napačno pozitivnih segmentacij, kjer sklade FV-jev označi kot GA-je. Z analizo napačnih segmentacij smo odkrili, da lahko tovrstne napake v veliki meri odpravimo že s filtriranjem po



(a) (b) (c) (d) (e)

Slika 2: Vhodni podatki in segmentacija GA-ja na enem podvolumnu. (**a**) Vhodni podatki. (**b**) Rezultat globokega učenja. (**c**) Rezultat metode aktivnih kontur. (**d**) Končna segmentacija – rezultat odstranjevanja napačnih označb. (**e**) Zlati standard.



(a) (b) (c)

Slika 3: (**a**) Del 2D rezine enega izmed podvolumnov z GA-jem. (**b**) Segmentacija GA-ja pred uporabo aktivnih kontur. (**c**) Segmentacija GA-ja po uporabi aktivnih kontur.

velikosti, saj so skladi navadno manjši. Tako odstranimo vse segmentirane predelke, ki so po volumnu manjši od 80 % velikosti najmanjšega GA-ja (pribl. 0,19 μm^3).

3 Rezultati

3.1 Vrednotenje

Za vrednotenje smo uporabili mere uspešnosti, ki se najpogosteje uporabljajo pri analizi biomedicinskih slik: Diceov koeficient (DSC, angl. *Dice Similarity Coefficient*), stopnja pravilno pozitivnih primerov – občutljivost (TPR, angl. *true positive rate*) in stopnja pravilno negativnih primerov – specifičnost (TNR, angl. *true negative rate*). Mere so definirane v članku Taha et. al. [11]. Izmerili smo tudi čas, ki ga porabimo za grobo in natančno ročno označevanje volumnov.

3.2 Rezultati

Stopnja metode	DSC	TPR	TNR
Po globokem učenju	0,515	0,861	0,983
Po uporabi aktivnih kontur	0,501	0,876	0,982
Končna segmentacija	0,926	0,885	0,999

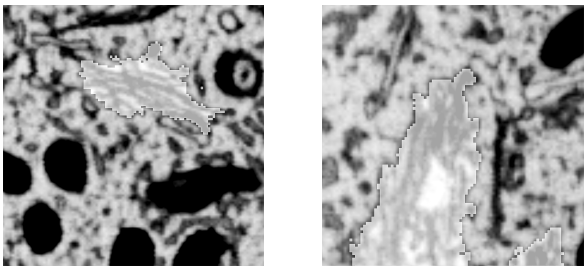
Tabela 1: Rezultat na različnih stopnjah naše metode. Zadnja vrstica predstavlja rezultat celotnega predlaganega pristopa.

Rezultati so prikazani v Tabeli 1. Izračunali smo jih na dveh podvolumnih, ki smo ju uporabili kot testno množico. Za potrebe evalvacije smo oba podvolumna ročno označili z natančnim robom. Rezultati v Tabeli 1 predstavljajo povprečni rezultat obeh testnih podvolumnov. Ker druge metode za avtomatsko segmentacijo GA-jev v volumetričnih podatkih še nismo zasledili, našega pristopa nismo mogli primerjati z drugimi.

Čas natančnega ročnega označevanja povprečno velikega GA-ja je bil 2,5 ure, medtem ko je čas grobega ročnega označevanja znašal 0,5 ure, kar je petkratna pohitritev.

3.3 Razprava

Rezultati po globokem učenju, uporabi aktivnih kontur in odstranjevanju napačnih označb dosegajo visoke mere uspešnosti. Diceov koeficient je 0,92, prav tako je zelo visok TNR, nekoliko nižji pa je TPR. To lahko pripišemo sledečim pomankljivostim.



(a) (b)

Slika 4: Napačne napovedi segmentacije. (**a**) FV označen kot GA. (**b**) Prazen prostor v sredini, označen kot GA.

Problem, ki se še vedno pojavlja, so napačne označbe, ki jih tudi filtriranje ni odstranilo. V vseh primerih so to označbe FV-jev, kot je prikazano na Sliki 4a.

Prav tako del napak, kot je vidno na Sliki 4b, doprinesejo tudi prazni prostori na sredini GA-jev, ki jih ročno označimo kot ozadje, ker pa je naš globoki model naučen na grobo označenih podatkih, sredinskih prostorov ne odstrani. Slednje sicer ni velika težava, saj nam zunanje meje GA-jev že dajejo veliko informacij – število GA-jev v celici, njihova prostorska razporeditev in odnosi med njimi. Notranja struktura GA-jev nas trenutno ne zanima toliko, je pa to eden izmed vidikov, ki ga želimo obravnavati v prihodnosti.

Nizke mere uspešnosti pred uporabo odstranjevanja napačnih označb so razumljive, saj je v celicah precej skladov FV-jev, ki so zelo podobni GA-jem, vendar jih s filtriranjem večino odstranimo. Zanimivo pa je, da vidimo zmanjšanje uspešnosti po uporabi metode aktivnih kontur. Do tega pride predvsem zaradi napačno označenih skladov FV-jev, kjer je samo del sklada označen kot GA, torej rob segmentacije na nekaterih delih poteka znotraj sklada, kar povzroči širjenje roba pri uporabi aktivnih kontur in s tem še večje število napačno segmentiranih vokslov.

Kljub temu da rezultatov nismo mogli primerjati z ostalimi pristopi, nam tudi kvalitativna analiza kaže, da pristop deluje dobro. Na Sliki 2 so prikazani zlati standard in rezultati algoritma na enem izmed testnih podvolumnov.

Izkazalo se je, da naša metoda na testni množici nikoli ni zgrešila celotnega GA-ja, razen v enem primeru, ko je bil znotraj podvolumna samo njegov manjši del in je bila posledično njegova velikost manjša kot meja filtriranja pri odstranjevanju napačnih označb. Pri segmentiranju celotnega volumna celice do takih napak ne bi prišlo.

Prednost predlaganega pristopa je tudi pohitren postopek ročnega označevanja. Pri povprečno velikem GA-ju smo za natančno ročno označevanje porabili petkrat več časa kot za grobo označevanje, ki ga uporabljamo pri globokem učenju. To pomeni, da bi za podatke, na katerih smo testirali naš pristop, z natančnim označevanjem porabili približno 50 ur več. Kljub temu da uporabljamo grobo označen zlati standard, je predlagan segmentacijski pristop popolnoma avtomatski.

4 Zaključek

Z razvojem tehnologije za zajem vse bolj natančnih mikroskopskih podatkov se kaže vedno večja potreba po pristopih za analizo le-teh, predvsem tistih, katerih raziskovanje je bilo prej zaradi nezadostne tehnologije oteženo. V članku smo zato predstavili pristop za avtomatsko segmentacijo Golgijevih aparatov v volumetričnih podatkih elektronske mikroskopije, ki je po nam znanih podatkih prvi na tem področju. Poleg uspešne segmentacije naš pristop z uporabo grobo segmentiranih učnih podatkov znatno zmanjša čas ročnega označevanja, ki je potrebno za nadzorovano globoko učenje. Predlagani pristop bo lahko uporabljen tudi na drugih področjih analize slik, kjer želimo segmentirati objekte z zapletenimi robovi, katerih ročno označevanje bi sicer vzelo preveč časa.

Pri nadaljnjem delu bomo izboljšali način odstranjevanja napačnih označb z vključevanjem dodatnih parametrov. Nadaljevali bomo tudi z izboljšanjem same segmentacije in pristop preizkusili na drugih podatkovnih zbirkah.

Celice sestavljajo vsa živa bitja, pa vendar njihovega delovanja še ne razumemo popolnoma. S predlaganim pristopom olajšujemo nadaljnje raziskovanje celičnih struktur in njihovih odnosov.

Zahvale

Za vso pomoč, podporo in spodbudo pri izdelavi dela se zahvaljujem asist. Manci Žerovnik Mekuč in mentorju izr. prof. dr. Matiju Maroltu iz Laboratorija za računalniško grafiko in multimedije na Fakulteti za računalništvo in informatiko ter doc. dr. Samu Hudoklinu iz Inštituta za biologijo celice na Medicinski fakulteti.

Literatura

[1] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1–1, 2016.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *LNCS*, vol. 9351, pp. 234–241, 2015.

[3] Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," pp. 424–432, 2016.

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," pp. 565–571, 2016.

[5] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61 – 78, 2017.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. [van der Laak], B. [van Ginneken], and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017.

[7] M. Žerovnik Mekuč, C. Bohak, S. Hudoklin, B. H. Kim, R. Romih, M. Y. Kim, and M. Marolt, "Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data," *Computers in Biology and Medicine*, vol. 119, p. 103693, 2020.

[8] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task," pp. 348–360, 2017.

[9] R. Kikinis, S. Pieper, and K. Vosburgh, *3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support*, 2014, vol. 3.

[10] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[11] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BioMed Central*, vol. 15, p. 29, 2015.

Predvidevanje vrednosti delnice in problem pretiranega prilagajanja modela

Miha Breznik[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, 1000 Ljubljana*
E-pošta: mihab2031@gmail.com

Predicting stock value and model overfitting

Paper presents how to minimize overfitting in predicting value of stocks using neural networks. Predicting value of stocks is very complex, so accuracy of predictions is very poor. Methods that do not require increasing volume of data are explained in detail. Data used in the paper was available free of charge which caused problems with overfitting and poor accuracy, due to limited range and density of data. Different methods to increase accuracy of the model, such as changing structure of a model, changing method how to approach a problem and adding new types of values, were used. Model is based on LSTM (long short term memory) recurrent neural network. Best results that show real improvement are achieved using binary output with a relatively large model and number of epochs.

1 Uvod

Problem napovedovanja vrednosti delnic je zelo kompleksen, saj na vrednost delnice vpliva veliko različnih faktorjev. Kljub temu je pregled literature pokazal večje število obstoječih rešitev izbranega problema. Literatura, ki predstavi smiselnost uporabe nevronskih mrež za napovedovanje vrednosti delnic, opisuje različne vrste mrež [1] in predstavi osnovno teorijo in smernice načrtovanja [2]. To tehnologijo že uporabljajo investicijske družbe, kot so UBS, Goldman Sachs ipd., ter tehnološki giganti, kot so Google in Yahoo.

Vplivi na delnice so lahko notranji, kot so na primer nov izdelek ali menjava vodstva. Vplivi so lahko tudi zunanji, kot je na primer trenutna epidemija ter s tem povezana gospodarska kriza. Vsi ti vplivi morajo biti zajeti v vhodnih podatkih. Odzivi morajo biti ponovljivi, da jih lahko model nevronske mreže (v nadaljevanju model) predvidi. Ti pojavi so lahko modelu predstavljeni kot internetna iskanja imena podjetja. Za zunanje nepovezane vplive na delnice so podatki dodani v obliki ključnih besed »bear market«, »bull market«, ki predstavljajo tendenco padanja (»bear«) ali rasti (»bull«) na finančnih trgih.

Vrednost delnice je potrebno napovedati tudi za nekaj dni naprej, da vidimo, za koliko bo vrednost delnice narastla ali padla. Model je napovedoval vrednost delnice na dveh časovnih intervalih. Prvi model je vsakodnevno napovedoval vrednost obravnavanih delnic naslednjega dne. Drugi, tedenski model, je napoved izvajal za sedem dni naprej. Iz dobljenih podatkov je razvidno, da takšno predvidevanje ni najbolj natančno in učinkovito. Problem napovedovanja vrednosti delnic si je mogoče zastaviti tudi drugače, tako da predvidimo binarno, ali delnica naraste ali pade v izbranem časovnem okviru v prihodnosti. Slednji model lažje predvidi spremembo vrednosti delnice, posledično se poveča natančnost modela in zmanjša pretirano prilagajanje (angl. overfitting).

Za izdelavo modela je bila uporabljena knjižnica tenserflow [3]. Vsi podatki, ki so bili uporabljeni v tem članku, so brezplačno dostopni. Podatki so dostopni na Yahoo finance [4] in Google trends [5]. Za večjo natančnost je mogoče uporabiti primernejše podatke, ki so za to plačljivi (običajno z gostejšim vzorčenjem in daljšimi časovnimi obsegi).

Model je mogoče uporabiti tudi za napovedovanje kripto valut in vrednost naravnih surovin, kot sta zlato in nafta. Za testiranje in prikaz podatkov so bile uporabljene delnice Tesle, Googla, Appla in Boieng.

2 Opis problema

Pretirano prilagajanje modela je problem, kadar ima model veliko natančnost pri treniranju in majhne izgube [6] (angl. loss), medtem ko v fazi validacije natančnost modela izrazito pade. Izguba je razlika med pravo vrednostjo in vrednostjo, ki jo model predvidi. Pretirano prilagajanje je mogoče zmanjšati s spremembo modela, predčasno zaustavitvijo treniranja modela, dodajanjem/odvzemanjem vrst podatkov, večjim številom podatkov in z uporabo tehnike regularizacije izpadanja [6] (angl. dropout). Tekom članka so podrobneje predstavljeni pristopi, ki so se izkazali za uspešne v obravnavanem primeru.

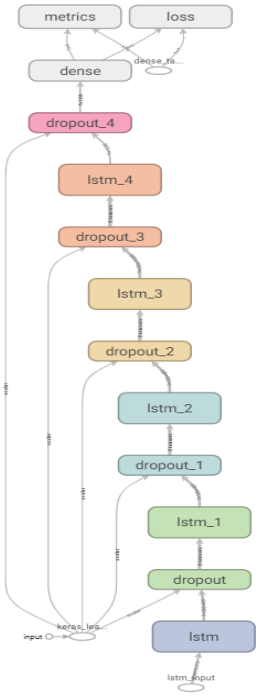
3 Zgradba modela

Osnovna struktura modela je izdelana iz nevronske mreže z dolgoročnim in kratkoročnim pomnilnikom LSTM [6], ki temelji na osnovi rekurentne nevronske mreže. Aktivatorska funkcija je usmerjena linearna enota kratice reLU [6]. Definirana je z enačbo (1).

$$f(x) = x^{+} = \max(0, x) \tag{1}$$

Za optimatizacijo modela je uporabljen algoritem adam [6], ki se uporablja za posodobitev uteži v mreži. Izgube so izračunane z efektivnim odstopanjem vrednosti (angl. Root mean square). Med plastmi je uporabljena tehnika izpadanja s koeficientom 0,20. Model je bil prilagojen za zmanjševanje pretiranega prilagajanja, tako da so se

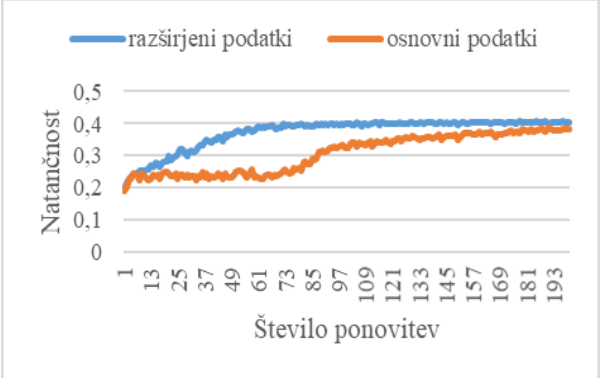
dodajale plasti mreže in se je spreminjalo število enot v plasti. Zaradi povečevanja modela je čas treniranja naraščal eksponentno. Iz obravnavanih podatkov in testiranja je nastal model, ki je prikazan na sliki 1. Ta model je bil uporabljen pri nadaljnjem raziskovanju. Število vhodnih podatkov v modelu je 30, kar pomeni, da model predvideva vrednost delnice na podlagi informacij pridobljenih v zadnjih 30 dneh.



Slika 1: Struktura modela mreže

4 Vhodni podatki

Vhodni podatki so bili sestavljeni iz vrednosti delnice ob odprtju (OPEN), vrednosti delnice ob zaprtju trga (CLOSE), najnižje vrednosti (LOW), najvišje vrednosti (HIGH) in prometa delnic (VOLUME). Za zmanjševanje pretiranega prilagajanja so bila dodana še spletna iskanja, »bear market«, »bull market« in ključna beseda recesija ekonomije. Uporaba slednjih je povečala učinkovitost modela za nekaj procentov, kar je predstavljeno na grafu (slika 2), v obliki natančnosti in izgube modelov. Vidno je, da se je s povečanjem števila vhodov povečala tudi natančnost.





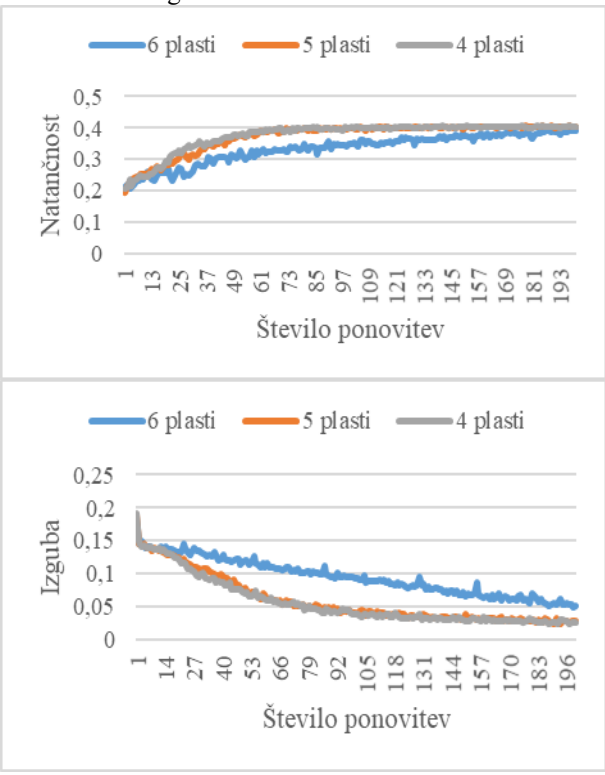
Slika 2. Primerjava števila vhodov

5 Eksperimenti

Izvedeni so bili naslednji eksperimenti: zmanjševanje pretiranega prilagajanja s spreminjanjem strukture, dodajanjem dni (dodajanje števila izhodov) za napoved izhoda in ponovna zastavitev problema.

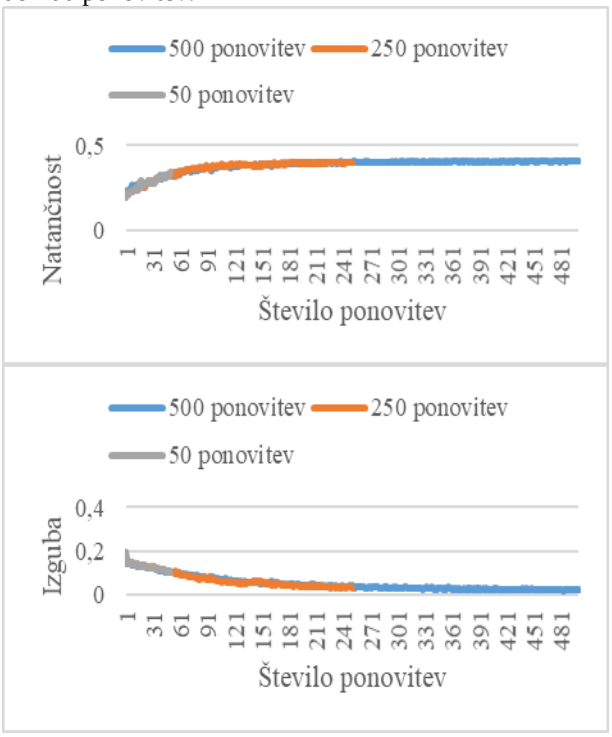
5.1 Spreminjanje modela

Zmanjšanje pretiranega prilagajanja je najlažje doseči s spreminjanjem modela in uporabo tehnike izpadanja. Tehnika temelji na izločanju naključnih nevronov iz modela. S povečevanjem faktorja izpadanja lahko naredimo model nestabilen, zaradi tega je nastavljen faktor na 0,20. Model lahko podaljšamo z dodajanjem plasti v mrežo in razširimo z dodajanjem celic v posamezno plast. Na sliki 3 so prikazane razlike med štirimi, petimi ter šestimi plastmi. Razvidno je, da spreminjanje strukture nima bistvenega vpliva na natančnost in izgubo modela.



Slika 3. Primerjava različnega števila plasti

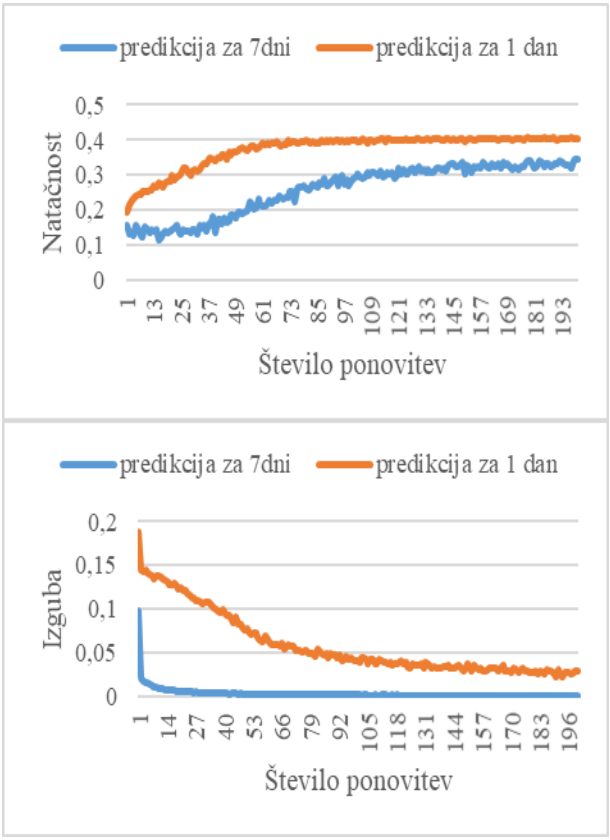
Optimalen je model s petimi plastmi, zaradi kasnejših eksperimentov, ki imajo boljšo natančnost s kompleksnejšo mrežo. V nekaterih primerih povečevanje kompleksnosti mreže pripomore k pretiranemu prilagajanju, kar se vidi pri šestih plasteh. Z dodajanjem plasti se zvišuje čas treniranja in prilagajanje modela na podatke. Spreminjanje števila celic ni pokazalo očitne spremembe pri predvidevanju, naraščal je samo čas učenja. Povečevanje števila ponovitev omogoči modelu, da se na različne načine nauči podatke, kar pomeni, da se poveča natančnost. Pretirano prilagajanje se pojavi tudi pri večjem številu ponovitev, zaradi tega je v članku uporabljeno 200 ponovitev. Iz slike 4 je vidna razlika natančnosti in izgube pri različnih ponovitvah. Vidno je, da se pri okrog 200 ponovitvah približujemo limiti natančnosti, zaradi tega se v nadaljevanju uporabi treniranje modela do 200 ponovitev.



Slika 4. Primerjava števila ponovitev

5.2 Povečanje število izhodov oz. predvidenih dni

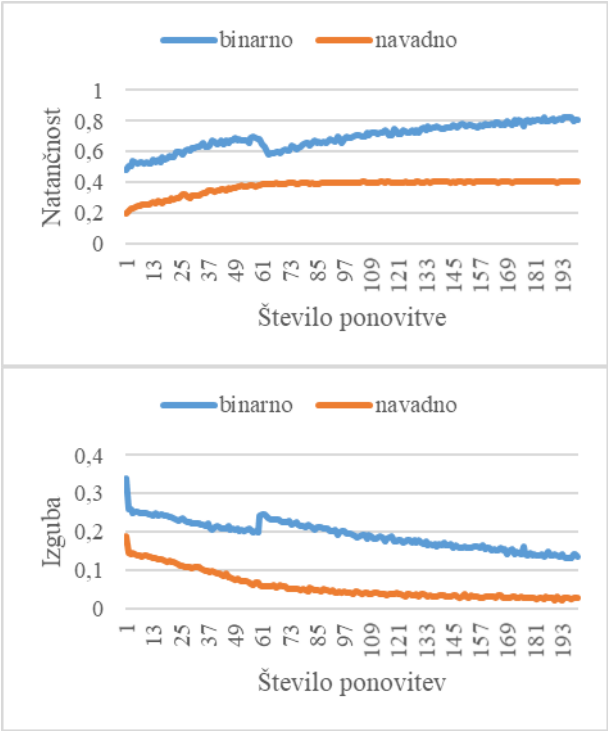
S povečevanjem števila izhodov zvišujemo kompleksnost problema, ki ga mora model ugotoviti. Tako povečamo možnost pretiranega prilagajanja. Slika 5 prikazuje predvidevanje za sedem dni v prihodnost. V primerjavi s predvidevanji vidimo, da model v večini primerov napačno ugotovi stanje delnice. Natančnost modela je v tem primeru zelo nizka.



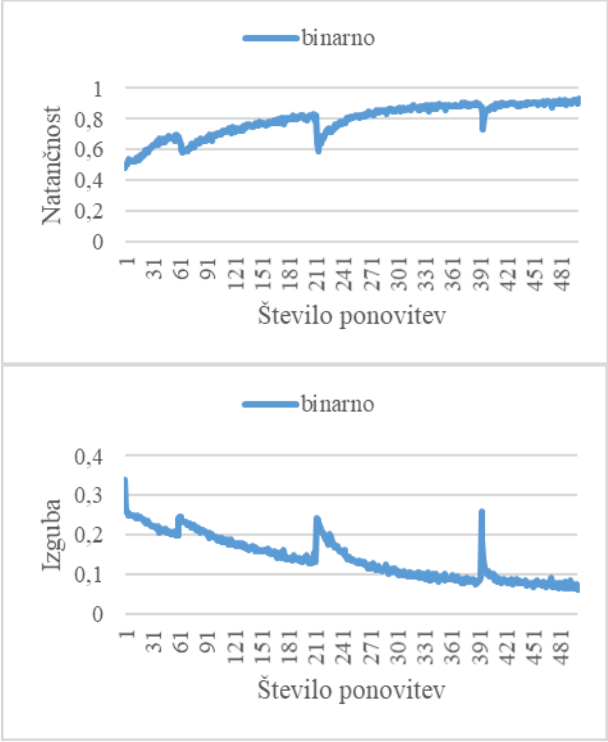
Slika 5. Primerjava napovedovanja sedem dni proti enemu dnevu

5.3 Sprememba zastavitve problema

Za izboljšanje natančnosti modela je mogoče zastaviti problem tako, da mora model zgolj napovedati, ali bo delnica narastla ali padla. Kompleksnost problema se tako zmanjša in posledično se zmanjša tudi možnost pretiranega prilagajanja. Iz pridobljenih podatkov je vidno, da je natančnost modela padla med učenjem, vendar natančnost predvidevanja naraste. Tu vidimo, kako se zmanjša pretirano prilagajanje z drugačno zastavitvijo problema. Na sliki 6 vidimo natančnost in izgubo v primerjavi s prejšnjimi modeli. Vidno je, da zaradi drugačne zastavitve modela natančnost izrazito naraste v primerjavi s prejšnjim. Za uporabo se je nov model naučil na 500 ponovitev, kar je vidno iz slike 7. Njegova natančnost je okoli 90 procentov, kar je že lahko primerno za praktično uporabo. Izguba konvergira proti napaki 0,1, kjer tudi ostane s povečevanjem ponovitev.



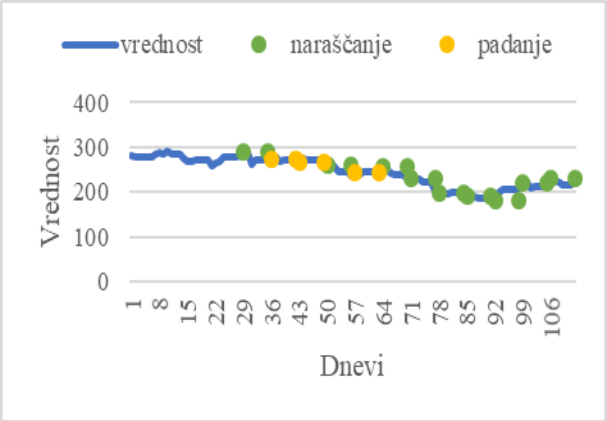
Slika 6. Primerjava binarnega z dnevnim napovedovanjem



Slika 7. Natančnost binarnega napovedovanja

6 Zaključek

V članku je opisano, kako optimizirati model LSTM mreže, z namenom zmanjšanja pretiranega prilagajanja na različne načine, ki niso povezani z uporabo večjega števila podatkov. Za najboljše rezultate si je potrebno problem zastaviti na binarni način, uporabiti pravilno mrežo, uporabiti dovolj velik faktor izpadanja, število želenih izhodov in vhodov. Z uporabo več vrst podatkov omogočimo modelu lažje zaznavanje kompleksnejših sprememb delnic. Na grafu je prikazano napovedovanje vrednosti oziroma, ali delnica narašča ali pada z uporabo binarnega modela. Zelena pika prikazuje, da bo delnica narastla, rumena pika pa, da bo delnica padla. Razvidno je, da v najboljšem primeru model še vedno naredi nekaj napak. Uporabljeni so bili podatki za 120 delovnih dni (cca. 4 meseci).



Slika 8. Napoved vrednosti z binarnim modelom

Literatura

[1] Mahdi Pakdaman Naeini, H. Taremian in Homa Baradaran Hashemi, "Stock market value prediction using neural networks," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krackow, 2010, pp. 132-136, doi: 10.1109/CISIM.2010.5643675

[2] Y. Li in W. Ma, "Applications of Artificial Neural Networks in Financial Economics: A Survey," 2010 International Symposium on Computational Intelligence and Design, Hangzhou, 2010, pp. 211-214, doi: 10.1109/ISCID.2010.70.

[3] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., "Tensorflow: a system for large-scale machine learning", OSDI, vol. 16, pp. 265-283, 2016.

[4] Yahoo! Finance, https://finance.yahoo.com/

[5] Google Trends, https://www.google.com/trends/

[6] I. Goodfellow, Y. Bengio, in A. Courville, Deep learning, MIT Press, 2016

Estimation of the derating PV power due to inverter clipping using Artificial Neural Network

Jakob Bevc

Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
E-mail: jakob.bevc@gmail.com

Abstract

A common PV plants design practice in climates with high irradiation is oversizing the DC capacity with regard to the AC capacity, which increases the inverter load ratio. The later reflects in a limited power output during summer time caused by inverter clipping. We propose the use of Artificial Neural Networks to estimate the amount of derating PV power due to aforementioned reasons. Our approach considers the learning phase where the model parameters are learnt on part of data not affected by clipping (operational data and environmental variables as input) and the evaluation phase where the model predicts the amount of derated PV power at times of clipping. We tested the approach in seven PV systems operating in different climate zones (Slovenia and the United States). The results show a clear agreement between predicted and true clipping losses ratio and improve the baseline model predictions in 5 out of 7 PV systems for up to 34 % lower relative nRMSE value in the best case. The high prediction accuracy can be beneficial to retrofit the design of PV systems, quantify the energy losses due to grid issues, and help to understand the behaviour of PV modules when operational data are limited.

1 Introduction

The worldwide energy transition towards renewables is forcing wide and fast deployment of Photovoltaic (PV) systems, showing increasing competitiveness with respect to conventional and other renewable energy sources [1]. The installed PV capacity grows exponentially in most of the climate regions, especially in the areas with high irradiation [2]. Consequently, the economy of scale forces down the costs of the PV module components, while trying to maintain or even increase its reliability and life time [3, 4].

In practice, the lower costs allow more flexible PV system designs, where a larger number of PV modules can be installed in the PV system of the same nominal output power. This consequently increases the inverter load ratio (ILR), defined as the ratio between the power capacity of PV module and AC inverter. Higher ILR values lead to higher share of maximum power generation and clipping mode time of the inverter. The latter reflects into flattened production curve that occurs mostly during the summer months. Previous research on this topic suggests an optimum ILR of 1.25 for minimal clipping loses [5]. Besides the inverter clipping, the limited power production can be caused by curtailment (e.g, forced power production by the grid operator) or when inverter failures.

For research purposes, the access to monitoring data at the DC side is usually unavailable, while the power injection to grid in some countries is shared by the grid operators. Even though having data at AC side is limited, still significant assessments can be carried out, such as performance modelling, quantification of the energy losses due to inverter clipping, curtailment or failures, as well as estimating of the operating behaviour of the PV modules for energy yield predictions and long-term degradation studies [6, 7, 8]. Our previous study shows that Machine Learning (ML) approaches can increase the accuracy of the estimation of the PV energy yield over time, using historical operational and weather data. [9].

In this paper we took a different approach regarding the training strategy. We estimate the amount PV power lost at times when inverter clipping occurs, based on the ML model trained on the data not suffering from clipping. Although the data are given in the form of time series, here the time component is disregarded.

We use the Artificial Neural Networks (ANN) to predict the PV energy yield in seven experimental PV systems operating in different Köppen-Geiger-Photovoltaic (KGPV) climate zones [10] in the Slovenia and United States. We begin by defining the general methodology for predicting the available PV power using ANN. Then, we define the datasets used in this research, the feature engineering and data filtering processes. Next, specifications of the ANN are given, together with the definition of generalized evaluation metrics. Finally, the results are presented in terms of modeling accuracy and estimation of energy losses due to inverter clipping.

2 Methodology

Here we evaluate selected field exposed PV systems installed in the Slovenia and United States. The available datasets do not suffer inverter clipping, thus, for experimental purposes, the data are altered with a power clipping value. Similarly as for the forecasting, environmental variables are used to train an ANN model to estimate the energy yield on the DC side of a PV system.

The training of ANNs is done on the data points not affected by clipping whereas the testing of the approach is carried out with the clipped data points, aiming to accurately predict the amount of derated PV power. Such training/testing data split simulates real life inverter clipping conditions and it enables us to evaluate the predictive performance of ML models.

2.1 Datasets

We perform experiments on measured operational data of several PV systems, situated at different locations: Ljubljana (Slovenia), Nevada (USA), Florida (USA) and New Mexico (USA) [11], (Table 1). All the datasets include the main climate variables, measured plane-of-array irradiance (G_{poa}) and air temperature (T_{air}), as well as the measured power output on the DC side (P_{DC}).

Location (file name)	$P_{nominal}$ [kW]	KGPV zone
Ljubljana (LPVO)	17.24	DM
New Mexico 1 (c10hov6)	3.24	CK
New Mexico 2 (t3pg1sv)	3.24	CK
Nevada 1 (wxzsjaf)	3.24	BK
Nevada 2 (kobdpi8)	3.24	BK
Florida 1 (wca0c5m)	3.24	DH
Florida 2 (z0aygry)	3.24	DH

Table 1: Description of the datasets: name, file name, nominal DC power and KGPV zone.

2.2 Artificial neural network

Artificial neural networks (ANN) are a model class that can be applied to wide variety of tasks. The core elements of ANNs are mathematical units called "neurons" which take multiple input values and multiply those with their respective weights, later refereed as parameters. The result of multiplications are followed by the sum up, which is then passed trough non-linear function called "activation function". A number of the neurons is then stacked together to form a "hidden layer". A typical neural network architecture is represented by the input layer, one or more hidden layers and the output layer, all off those densely connected.

The main strength of ANNs is their ability to fit complex nonlinear functions, with appropriately selecting parameters values. Those parameters are learned during the supervised learning process, where input data of form *(variables, label)* is used. During this process, neural network parameters are repeatedly adjusted, to minimize the selected loss function, which is mean squared error (MSE) in case of regression task. Additionally, a regularization stage is used to prevent the overfitting, where L2-norm weight penalty is added to the loss function [12].

2.3 Feature engineering and data filtering

In order to improve the performance of ML models, multiple new features where created. This involves logarithmic transformation of G_{poa} (named as G_{poa_log}) and also taking the square of the log transformed values (named as $G_{poa_log_sq}$). The motivation behind using log transformations is that it usually makes highly skewed distributions less skewed, which results in more interpretable data. In addition to that, one of the most accurate empirical models, the so-called PVGIS model, also includes the log transformation of irradiance following the physical behaviour of PV solar cells [13]. As reported in [9], the sun position, the zenith and azimuth angles, can also enhance the modelling, therefore, they are also considered in the datasets used.
To reduce the uncertainties due to faulty measurements, the data is pre-processed, aggregating the values on hourly basis, and filtering out samples with Performance Ratio (PR) below 0.4.

2.4 Generalized evaluation metrics

To evaluate the model accuracy using different feature sets, we consider the normalized RMSE (nRMSE) as presented in Equations 1 and 2.

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_{i,true} - y_{i,predicted})^2} \quad (1)$$

$$nRMSE = \frac{RMSE}{P_{nominal}} \quad (2)$$

Additionally, we estimate the ratio of energy losses due to the clipping. Ratio of energy losses is calculated as:

$$clipping_r = \frac{clipped_energy}{total_energy}, \quad (3)$$

where $total_energy$ is the total energy production without inverter clipping, and the $clipped_energy$ represents the total amount of energy above the clipping value, in this study set at 70 % of the nominal power.

3 Results

To evaluate the models, we first select a one year period of operational data and then, split the data on train and test set, for the seven PV systems. The limit clipping value is set to 70 % of nominal power. Thus, the respective ILR value is 1.43, where we consider power capacity of AC inverter as the limit clipping value, for each PV system. Experiments were made on subsets of original and generated features from which the best features set were selected. In addition to ANN, we have also established a common baseline, for all the datasets, which utilized the PVGIS empirical model [9, 13].

3.1 Accuracy of predictions

We present the base feature set (FS1) and three best performing feature sets (FS2, FS3, FS4) (Table 2) and the respective results for each data set (Table 3). Results show that ANN in combination with an appropriate feature set can outperform the baseline model for 5 out of 7 datasets. Proposed features also improve the results in most cases. We have also estimated the uncertainty of results by multiple iterations of learning process on the same dataset, while considering different initial ANN parameters.

Features set	Variables
FS1	G_{poa}, T_{air}
FS2	$G_{poa}, T_{air}, azimuth, zenith$
FS3	$G_{poa}, T_{air}, G_{poa_log}, G_{poa_log_sq}$
FS4	$G_{poa}, T_{air}, azimuth, zenith,$ $G_{poa_log}, G_{poa_log_sq}$

Table 2: Input variables used in different feature sets.

An example of the train/test split for clipped dataset is presented (Figure 1), where one year of operational data in Ljubljana (LPVO) is analysed with FS2. The results show that ANN outperforms the baseline PVGIS model by 0.31 % in terms of nRMSE (Table 3).

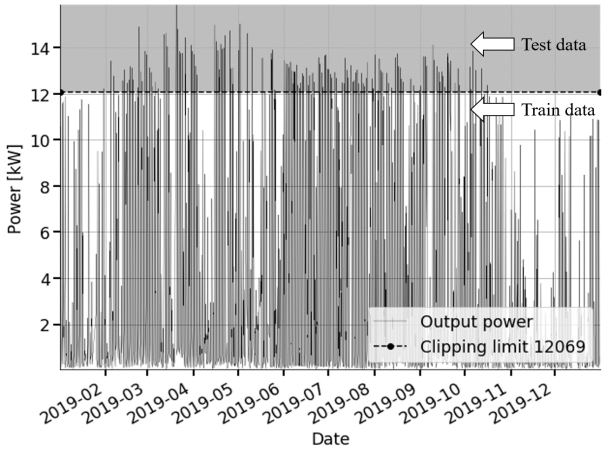


Figure 1: Power output through the year 2019 for Ljubljana.



Figure 2: CDF plot of power output values for Ljubljana.

Additionally, we analyze the results with the cumulative distribution function (CDF), where a certain value point on the curve represents the probability of the system to generate less or equal power than the observed. The comparison of true and predicted probabilities at certain power value can serve as an indicator if the prediction model is prone to the over or underestimation. Higher probability of predicted over the true values, near the clipping limit, (Figure 2), indicates that ANN model is underestimating in that part. On the other hand, minor overestimations occur in the middle range of power values.

3.2 Energy losses

The ratio of energy losses (3) and amount of days with power limitations, highly depends on the location, and consequently the KGPV climate zone (Table 4). The choice of a clipping value of 70 % of nominal power is reflected in a large number of underperforming days due to the derating power (from 101 days in Florida 1 to 316 days a year in New Mexico 1). This fictional scenario could be worse than expected in reality, however it helps us to evaluate the ANN under a harder scenario. In terms of percentage, the applied clipping value forces the systems to lose up to 5.1 % of the annual energy yield. For all the locations, excluding Florida, the ratio of predicted energy losses is overestimating the true one by a small margin.

	Days number		Clipping ratio (%)	
Location	Total	Clipping	True	Predicted
Ljubljana	360	148	1.69	1.76
New Mexico 1	364	316	5.10	5.38
New Mexico 2	364	303	3.92	4.09
Nevada 1	318	263	3.51	3.53
Nevada 2	360	299	3.97	4.52
Florida 1	362	101	0.63	0.58
Florida 2	362	111	0.79	0.74

Table 4: Number of clipping days and clipping ratio at clipping value of 70 % of nominal power and ILR of 1.43.

4 Discussions

The proposed approach and its validation stage are the first attempts in the prediction of possible PV energy yield when power limitations occur at the AC side. The use of high-quality climate data such as irradiance and ambient temperature helps to understand the electrical and thermal behaviour of the PV system. Improving the prediction accuracy of standard empirical models (e.g, logaritmic functions) with ML approaches, gives PV analysts an opportunity to even further optimize their project's inverter load ratio, and consequently improve the cost effectiveness of their designs. Evaluation of the proposed technique showed that it outperforms the baseline model for 5 out of 7 datasets in different climate zones for up to 34 % better relative nRMSE value in best case. We also observed that PVGIS baseline outperforms the ANN on the feature set FS1 in all cases. The latter is mainly due to the fact that PVGIS model implicitly already contains the logarithmic and other feature transformations, and is therefore able to learn accurate relations between those and the output power. Even though the current approach utilizes manually clipped datasets, which is not the case in a real environment, we aim to apply the same approach on real cases and utility scale PV systems. The latter can corroborate that ML approaches can help to better understand the behaviour, model and forecast the performance of PV modules when limited data is available.

In a desert climate as Nevada, the nRMSE scores of baseline model were not improved. Possible reasons for that are stronger thermal effects at high irradiances than in other climates. Additionally we also observe high uncertainties in some cases (Table 3), this could be reduced

		nRMSE (average ± standard deviation %)						
		Ljubljana	N. Mexico 1	N. Mexico 2	Nevada 1	Nevada 2	Florida 1	Florida 2
ANN	FS1	3.79 ± 0.67	3.14 ± 1.14	2.52 ± 1.09	2.70 ± 1.19	2.81 ± 1.11	2.35 ± 0.31	1.56 ± 0.32
ANN	FS2	$\mathbf{1.39 \pm 0.15}$	$\mathbf{1.79 \pm 0.13}$	$\mathbf{1.89 \pm 0.05}$	1.70 ± 0.06	2.04 ± 0.10	$\mathbf{1.44 \pm 0.05}$	$\mathbf{1.41 \pm 0.13}$
ANN	FS3	1.86 ± 0.17	$\mathbf{1.52 \pm 0.08}$	$\mathbf{1.67 \pm 0.16}$	1.78 ± 0.41	1.59 ± 0.07	$\mathbf{1.52 \pm 0.21}$	$\mathbf{1.42 \pm 0.13}$
ANN	FS4	$\mathbf{1.47 \pm 0.13}$	$\mathbf{1.58 \pm 0.06}$	$\mathbf{1.69 \pm 0.07}$	1.60 ± 0.10	1.62 ± 0.07	$\mathbf{1.39 \pm 0.04}$	$\mathbf{1.29 \pm 0.03}$
BASE	FS1	1.70 ± 0.01	2.30 ± 0.01	1.91 ± 0.01	1.45 ± 0.01	1.49 ± 0.01	1.62 ± 0.01	1.53 ± 0.01

Table 3: nRMSE average value and standard deviations both given in percentage. Bolded values indicate improved baseline values.

by normalizing the input variables as well as including new variables and mathematical operations in the input data. Another important note is that not all ML approaches can be used for such learning task. For example, most decision tree based algorithms are not suitable, since by their definition they cannot predict values larger than those in training data. For this reason, they completely fail on test set, even though they can perform very well on the classic prediction task [9].

Further work will address the problem of estimating PV energy yield not only for values above clipping limit, but also at the time-scale, which presents double challenges: prediction of yield in the future and prediction of yield under power limitations.

5 Conclusions

In this paper we estimated the possible PV energy yield from the limited data available on the AC side. The limited power on the AC side is mainly the consequence of inverter clipping, or even possible failures and curtailment. Predicting the possible energy yield is done by utilizing ML approaches, more specifically ANN, using environmental variables (G_{poa}, T_{air}) as input features. Evaluation of our methodology is done on the non clipped data, which is then manually clipped to simulate the real life power limitations on AC side. With only environmental variables as input to the ANN we could not outperform the PVGIS baseline. On the other hand, after constructing new features and combining them in different features sets, we were able to outperform the baseline for 5 out of 7 datasets.

Acknowledgements

I would like to thank my mentor Julián Ascencio-Vásquez for leading me through the research process and helping me improve the final manuscript. Additionally I want to express my great appreciation to Assoc. Prof. Dr. Marko Jankovec and Prof. Dr. Marko Topič for their help and advice in the process of writing scientific paper. The work was carried out mainly at the Laboratory of Photovoltaics and Optoelectronics (LPVO), Faculty of Electrical Engineering, University of Ljubljana.

References

[1] E. Vartiainen, G. Masson, C. Breyer, D. Moser, and E. Román Medina, "Impact of weighted average cost of capital, capital expenditure, and other parameters on future utility-scale pv levelised cost of electricity," *Progress in Photovoltaics: Research and Applications*, vol. 28, no. 6, pp. 439–453, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3189

[2] S. Europe, "Global market outlook for solar power 2020-2024," 2020.

[3] J. Ascencio-Vásquez, I. Kaaya, K. Brecl, K.-A. Weiss, and M. Topič, "Global climate data processing and mapping of degradation mechanisms and degradation rates of pv modules," *Energies*, vol. 12, no. 24, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/24/4749

[4] I. Kaaya, M. Koehl, A. P. Mehilli, S. de Cardona Mariano, and K. A. Weiss, "Modeling outdoor service lifetime prediction of pv modules: Effects of combined climatic stressors on pv module power degradation," *IEEE Journal of Photovoltaics*, vol. 9, no. 4, pp. 1105–1112, 2019.

[5] J. Good and J. X. Johnson, "Impact of inverter loading ratio on solar photovoltaic system performance," *Applied Energy*, vol. 177, pp. 475–486, 2016.

[6] A. Zurita, A. Castillejo-Cuberos, M. García, C. Mata-Torres, Y. Simsek, R. García, F. Antonanzas-Torres, and R. A. Escobar, "State of the art and future prospects for solar PV development in Chile," *Renewable and Sustainable Energy Reviews*, vol. 92, no. C, pp. 701–727, 2018. [Online]. Available: https://ideas.repec.org/a/eee/rensus/v92y2018icp701-727.html

[7] L. R. Camargo], J. Valdes, Y. M. Macia], and W. Dorner, "Assessment of on-site steady electricity generation from hybrid renewable energy systems in chile," *Applied Energy*, vol. 250, pp. 1548 – 1558, 2019.

[8] S. Lindig, I. Kaaya, K. Weiss, D. Moser, and M. Topic, "Review of statistical and analytical degradation models for photovoltaic modules and systems as well as related improvements," *IEEE Journal of Photovoltaics*, vol. 8, no. 6, pp. 1773–1786, 2018.

[9] J. Ascencio-Vásquez, J. Bevc, K. Reba, K. Brecl, M. Jankovec, and M. Topič, "Advanced pv performance modelling based on different levels of irradiance data accuracy," *Energies*, vol. 13, no. 9, p. 2166, 2020.

[10] J. Ascencio-Vásquez, K. Brecl, and M. Topič, "Methodology of köppen-geiger-photovoltaic climate classification and implications to worldwide mapping of pv system performance," *Solar Energy*, vol. 191, pp. 672 – 685, 2019.

[11] A. Curran, C. B. Jones, J. L. Braid, J. S. Stein, and R. H. French, "US DOE-RTC-BaselineSystems," 2019. [Online]. Available: https://osf.io/yvzhk/

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[13] T. Huld, R. Gottschalg, H. G. Beyer, and M. Topič, "Mapping the performance of pv modules, effects of module type and data averaging," *Solar Energy*, vol. 84, no. 2, pp. 324 – 338, 2010.

Vector Space Representation of Movies and Music Albums Using Node Embedding Methods

Vid Keršič[1]

[1] *University of Maribor, Faculty of Electrical Engineering and Computer Science*
Koroška cesta 46, 2000 Maribor
E-mail: vid.kersic@um.si

Abstract

The performance of machine learning methods improves every year. While commonly used machine learning methods that process data in vector form can efficiently perform many tasks, non-Euclidean data structures like graphs are often present in different real-world problems. Therefore, in recent years, machine learning methods for node embedding, which transform nodes of graphs into vector space, have gained a lot of traction. In this paper, we conduct an analysis of movies and music albums to explore their usage in recommender systems. We construct graphs from Wikipedia articles by following the contained hyperlinks. Different node embedding methods are used to embed nodes of the obtained graphs in a vector space, which allows us to analyze them more efficiently, focusing on visualization, similarity (recommendations), and classification of movies and albums into genres. We achieved a classification accuracy of 88.5% for movies and 89.3% for albums.

1 Introduction

Machine learning has made great strides in recent years, successfully performing various tasks like classification, regression, and clustering [1]. Although entirely different machine learning methods are used for the mentioned tasks, they have a similar input data structure, which is a one-dimensional, two-dimensional, or high-dimensional vector [2]. But nowadays, many scientific fields like social networks deal with non-Euclidean data structures like graphs [2]. For that reason, machine learning methods for node embedding, which transform nodes of graphs into vector space, are becoming very popular.

Node embedding methods have been successfully used to increase the effectiveness of recommender systems. Embedding nodes of graphs constructed from Wikipedia articles improved the ability to find similar historical figures, perform a recommendation of articles to editors of Wikipedia, and classify books into genres [3, 4, 5]. In this paper, we use node embedding methods to analyze movies and music albums and explore their usability in recommender systems, e.g. movie and music album recommendation. We construct a graph from Wikipedia articles by following the available hyperlinks and then using machine learning methods for node embedding on the obtained graphs analyze movies and albums, focusing on visualization, similarity (recommendations), and classification of movies and albums into genres.

In the next sections, we review different machine learning methods for node embedding, describe the data collection and graph construction process, and present the results of the analysis of movies and music albums.

2 Node Embedding

A graph is a data structure defined as $G = (V, E)$, where V is a set of nodes, and E is a set of edges. Node embedding is defined as:

$$f : V \rightarrow \mathbb{R}^d, \tag{1}$$

where d is the dimension of vector space. Node $v_i \in V$ is transformed into vector $\mathbf{z}_i \in \mathbb{R}^d$. The objective of node embedding is to preserve the relations between the nodes and the structure of the graph. Over the past few years, many machine learning methods for the vector embedding of graph nodes have been introduced. In this section, we review the methods that were used in this analysis.

The first widely used node embedding methods were shallow embedding methods, where a single matrix $\mathbf{Z} \in \mathbb{R}^{d \times |V|}$ represents the embedding of all nodes (i-th column is vector $\mathbf{z}_i$) [6]. One of the most popular shallow embedding methods is Deepwalk [7]. DeepWalk is a random-walk based method that uses the Skip-gram model to learn node embedding. Skip-gram is trained on the generated random walks in an unsupervised manner. Its objective is to embed a pair of nodes that are close to each other (inside the sliding window) in random walks to also be close in the vector space, where the similarity between vectors is measured using the inner product. The probability of node v_j appearing close to node v_i is defined by the softmax function:

$$p(v_j|v_i) \approx \frac{e^{\mathbf{z}_i{}^T \mathbf{z}_j}}{\sum_{v_k \in V} e^{\mathbf{z}_i{}^T \mathbf{z}_k}} \tag{2}$$

During training, Skip-gram attempts to minimize the following cost function:

$$\mathcal{L} = \sum_{(v_i, v_j) \in \mathcal{D}} -\log\left(p(v_j|v_i)\right), \tag{3}$$

where $\mathcal{D}$ is the training set of pairs of nodes from the generated random walks. Since the normalization factor in

softmax is expensive to calculate, DeepWalk uses hierarchical softmax [7]. Another popular random-walk based method is node2vec, which differs slightly from DeepWalk in the process of generating random walks and the calculation of the normalizing factor [8]. Random walks are generated in a biased way, using the hyperparameters p and q. While p determines the likelihood of the walk returning to the previous node, q determines the likelihood of going further away from the last node. Therefore, random walks can be biased towards a breadth-first search (BFS) and depth-first search (DFS) of a graph. Instead of hierarchical softmax, node2vec employs negative sampling [8].

Shallow embedding methods have some limitations, like the inability to capture non-linearity in a graph structure [6]. Methods based on deep autoencoders like SDNE (Structural Deep Network Embedding) were introduced to solve the non-linearity problem [9]. Deep autoencoders consist of two deep neural networks - an encoder and a decoder. The encoder transforms input data into the embedding, while the decoder tries to reconstruct the input data from the embedding. SDNE is a semi-supervised deep model that attempts to preserve the graph's local and global structure using first-order and second-order proximities. First-order proximity describes the local structure and aims to produce similar embeddings for directly connected nodes. Second-order proximity describes the global structure and seeks to embed nodes with similar neighborhood structures closer to one another in the vector space. The loss functions of proximities and regularization define the loss function of SDNE:

$$\begin{aligned} \mathcal{L}_{1st} &= \alpha \sum_{i,j=1}^{n} s_{i,j} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|_2^2 \\ \mathcal{L}_{2nd} &= \sum_{i=1}^{n} \left\| (\hat{\mathbf{s}}_i - \mathbf{s}_i) \odot \mathbf{b}_i \right\|_2^2 \\ \mathcal{L}_{mix} &= \mathcal{L}_{1st} + \mathcal{L}_{2nd} + \gamma \mathcal{L}_{reg}, \end{aligned} \tag{4}$$

where $\mathbf{s}_i$ is the neighborhood vector of node v_i ($s_{i,j} \in \mathbf{s}_i > 0$ if edge between nodes v_i and v_j exists) and $\hat{\mathbf{s}}_i$ is the reconstructed neighborhood vector. The values α, $\mathbf{b}_i$, and γ are hyperparameters.

3 Data Collection and Graph Construction

The analysis of movies and music albums is done on the graphs constructed from Wikipedia articles and hyperlinks. Because we did not find any publicly available dataset containing movies and albums with hyperlinks to their Wikipedia articles, we created the dataset ourselves using web scraping. We extracted the list of movies from the website *TheNumbers*, where we chose five genres and 100 movies per genre: superhero, kids fiction, horror, musical, and fantasy [10]. A similar procedure was repeated for albums. We extracted the list from the website *DigitalDreamDoor* and also chose five genres (100 albums per genre): country, rap, metal, rock, and electronic [11]. To obtain Wikipedia articles of movies and albums, we wrote a custom script. Because some movies and albums do not have Wikipedia articles, we had to check the lists manually. Table 1 shows the number of movies and albums per genre after data cleansing.

Table 1: The number of movies and albums per genre after data cleansing.

Genre	Movies	Albums
Genre 1	100	93
Genre 2	100	100
Genre 3	100	100
Genre 4	93	86
Genre 5	92	99
Total	485	478

From the lists of movies and albums, we constructed the graphs by following hyperlinks in the articles. We created two graphs: one for movies and one for albums. Nodes in graphs are Wikipedia articles, and edges are hyperlinks between articles. Both graphs are undirected and unweighted. We only included hyperlinks from sections of the articles, which in our opinion best characterize movies and albums. The hyperlinks selection is one of the most critical steps because wrongly selected hyperlinks can connect completely different movies and albums (e.g. hyperlinks to film studios). We constructed graphs by running DFS with depth limited to 2 from every movie and album from the list. All of the discovered Wikipedia articles, not only those describing movies and albums, were added to the graphs. The constructed graph for movies consists of 3,216 nodes and 7,520 edges, while the graph for albums consists of 3,601 nodes and 9,289 edges. Figure 1 shows the graph for movies.

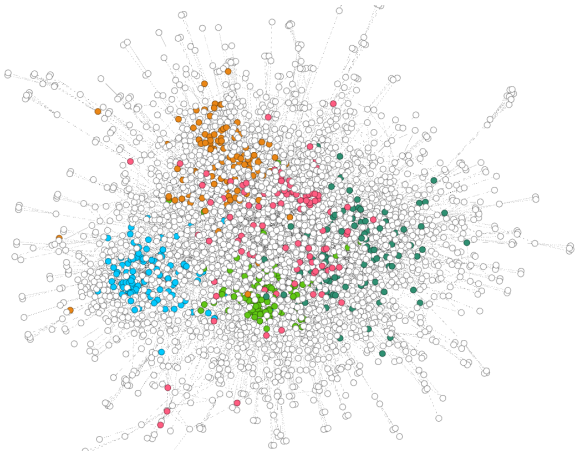


Figure 1: The constructed graph for movies. The colors of the nodes represent genres of movies: superhero (blue), kids fiction (green), horror (brown), musical (turquoise), and fantasy (pink). White nodes represent other Wikipedia articles.

4 Results

We compared the embeddings produced by four methods: DeepWalk, node2vec-BFS, node2vec-DFS, and SDNE.

The dimension of the node embedding was set to 128. The hyperparameters of the methods were tuned using a random and grid search. We generated 25 random walks from every node in random walk-based methods. For DeepWalk, we used *window size* 5 for movies and 15 for albums, and *walk length* 80 for movies and 60 for albums. For node2vec-BFS, we used p=0.5, q=2.0, *window size* 10 for movies and 15 for albums, and *walk length* 80 for both movies and albums. For node2vec-DFS, we used p=2.0, q=0.5, *window size* 15 for both movies and albums, and *walk length* 80 for both movies and albums. For SDNE, we used α=10^{-6} and β=150 for both movies and albums. The deep neural networks of SDNE consisted of two layers (3216-128 and 128-3216) for movies and three layers (3601-512-128 and 128-512-3601) for albums.

In this analysis, we only kept the embeddings of Wikipedia articles for movies and albums from the lists. We reduced the dimensionality of embeddings from 128 to 2 using principal component analysis (PCA) for visualization. Figure 2 shows the closest vector embeddings to the superhero movie The Dark Knight. Figure 3 shows the embeddings of all movies and albums colored by their genre.

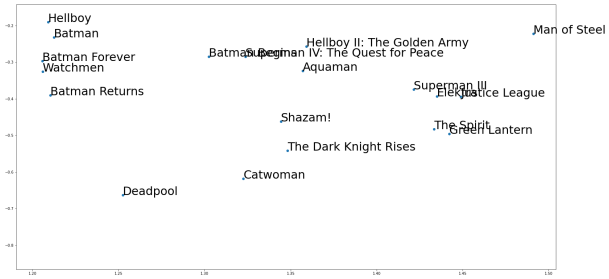


Figure 2: The closest vector embeddings to the superhero movie The Dark Knight produced by the method node2vec-DFS. All visualized movies are superhero movies, and many are from the Batman film series.

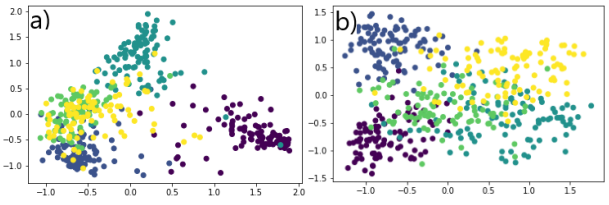


Figure 3: The embeddings of movies (a) and albums (b) produced by the method node2vec-DFS. Colors of the visualized vector embeddings represent genres (movies: superhero (violet), kids fiction (blue), horror (turquoise), musical (green), fantasy (yellow); albums: country (violet), rap (blue), metal (turquoise), rock (green), electronic (yellow)). Almost all movies and albums within the same genre have similar vector embeddings. Many musicals, fantasy, and kids fiction movies overlap. The reason is that many kids fiction movies are also fantasy movies, and they contain similar hyperlinks as fantasy movies. Some electronic, rock, and metal albums also overlap.

Cosine similarity was used to calculate the similarity between vector embeddings:

$$cos(\theta) = \frac{\mathbf{z}_i{}^T \mathbf{z}_j}{|\mathbf{z}_i||\mathbf{z}_j|} \quad (5)$$

The similarity between vectors is used for recommending movies and music albums. Table 2 shows the five most similar movies to the superhero movie X-Men: Dark Phoenix, and Table 3 shows the five most similar albums to the rap album Escape. SDNE embeds similar movies and albums much closer in a vector space than node2vec-DFS. However, because the graphs contain all articles from Wikipedia, the movie (album) articles can be far apart from each other. Because SDNE uses first and second-order proximities as similarity measurements, it fails to embed well articles that do not have a lot of similarity with others and are further away. On the other hand, random walk-based methods can explore a deeper part of the graphs.

Table 2: The top five most similar movies to the superhero movie X-Men: Dark Phoenix using cosine similarity as a similarity measurement between embeddings. For SDNE, all movies are from the same film series, while for node2vec-DFS the last one is not, but is still a superhero movie from the same film studio.

Method	Movie	Cos. sim.
node2vec-DFS	X-Men: Days of Future Past	0.852
	X-Men: The Last Stand	0.815
	X-Men 2	0.787
	X-Men: First Class	0.785
	Avengers: Endgame	0.759
SDNE	X-Men: Days of Future Past	0.929
	X-Men: First Class	0.922
	X-Men 2	0.904
	X-Men: The Last Stand	0.904
	X-Men: Apocalypse	0.878

Table 3: The top five most similar albums to the rap album Escape using cosine similarity as a similarity measurement between embeddings. In this case, SDNE generates similar embeddings for entirely different albums, while in the case of node2vec-DFS, all albums are from the same genre.

Method	Album	Cos. sim.
node2vec-DFS	Raising Hell	0.621
	How Ya Like Me Now	0.587
	King of Rock	0.583
	Run-D.M.C.	0.566
	The Low End Theory	0.501
SDNE	Ride the Lightning	0.728
	Mantronix: The Album	0.723
	Demon Days	0.722
	Black Sabbath	0.712
	Post	0.709

Multiclass logistic regression (one-vs-all) with 10-fold cross-validation was used for classification into genres. We evaluated classification performance using accuracy, precision, recall, and F1 score. Table 4 shows the classification accuracy of different methods. All methods

achieve high classification accuracy for movies, while SDNE fails for albums. Because node2vec-BFS and SDNE achieve lower classification accuracy for albums, we concluded that album articles are more distant from each other than movie articles. Tables 5 and 6 show detailed classification results for movies and albums using node2vec-DFS. We achieved F1 scores higher than 0.75 for all genres.

Table 4: Classification accuracy for movies and albums. The method node2vec-DFS achieved the highest classification accuracy - 88.5% for movies and 89.3% for albums.

Method	Movies	Albums
DeepWalk	0.885	0.887
node2vec-BFS	0.882	0.867
node2vec-DFS	0.885	0.893
SDNE	0.854	0.662

Table 5: Detailed classification results for movies using the method node2vec-DFS. The worst classified movie genres were kids fiction and fantasy, which can also be seen on the node embedding visualization.

Genre	Precision	Recall	F1
superhero	0.973	0.954	0.964
kids fiction	0.811	0.856	0.833
horror	0.942	0.948	0.945
musical	0.890	0.884	0.887
fantasy	0.806	0.774	0.789

Table 6: Detailed classification results for albums using the method node2vec-DFS. The worst classified album genres were rock and electronic, while the other three genres had an F1 score higher than 0.92.

Genre	Precision	Recall	F1
country	0.903	0.942	0.922
rap	0.937	0.950	0.943
metal	0.957	0.928	0.942
rock	0.796	0.781	0.788
electronic	0.861	0.853	0.857

5 Conclusion

In this paper, we used node embedding methods on the graphs constructed from Wikipedia articles to analyze movies and music albums, and explored their usability in recommender systems, focusing on visualization, similarity (recommendations), and classification into genres. We compared embeddings produced by four node embedding methods. The node2vec-DFS method achieved the best classification accuracy for both movies and albums: 88.5% for movies and 89.3% for albums. While all methods achieve high classification accuracy for movies, SDNE fails for albums. We have shown that calculating similarities between vectors produced by node embedding methods can be used for recommending movies and albums.

In the future, the analysis can be expanded in several ways. The nodes of the graphs can be extended with features, and different neighborhood aggregation methods (e.g. a graph convolutional network) can be used to produce node embeddings. An analysis of less extensive Wikipedia articles, which do not contain as many hyperlinks as movie and album articles, can also be performed.

6 Acknowledgements

This paper is closely related to my bachelor's thesis. I want to thank associate professor Damjan Strnad and assistant Štefan Kohek for their guidance and assistance.

References

[1] G. Rebala, A. Ravi, and S. Churiwala: An Introduction to Machine Learning. Cham: Springer International Publishing, 2019.

[2] M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst: Geometric Deep Learning: Going beyond Euclidean data, IEEE Signal Processing Magazine, vol. 34, no. 4. Institute of Electrical and Electronics Engineers Inc., pp. 18–42, 01-Jul-2017, doi: 10.1109/MSP.2017.2693418.

[3] Y. Chen, B. Perozzi, and S. Skiena: Vector-based similarity measurements for historical figures, Inf. Syst., vol. 64, pp. 163–174, Mar. 2017, doi: 10.1016/j.is.2016.07.001.

[4] O. Moskalenko: Convolutional Graph Embeddings for article recommendation in Wikipedia, 2019.

[5] Building a Recommendation System Using Neural Network Embeddings. [Online]. Available: https://towardsdatascience.com/building-a-recommendation-system-using-neural-network-embeddings-1ef92e5c80c9. [Accessed: 19-Jun-2020].

[6] W. L. Hamilton, R. Ying, and J. Leskovec: Representation Learning on Graphs: Methods and Applications, Sep. 2017, arXiv: 1709.05584.

[7] B. Perozzi, R. Al-Rfou, and S. Skiena: Deepwalk: Online learning of social representations, in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–710.

[8] A. Grover and J. Leskovec: node2vec: Scalable Feature Learning for Networks, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 855–864, Jul. 2016.

[9] D. Wang, P. Cui, and W. Zhu: Structural Deep Network Embedding, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, vol. 13-17-Augu, pp. 1225–1234, doi: 10.1145/2939672.2939753.

[10] The Numbers - Where Data and the Movie Business Meet. [Online]. Available: https://www.the-numbers.com/. [Accessed: 19-Jun-2020].

[11] DigitalDreamDoor.com - Greatest Music, Movie, and Book lists. [Online]. Available: https://digitaldreamdoor.com/. [Accessed: 19-Jun-2020].

Razvoj IoT aplikacije za pametni dom

Jaša Vid Meh Peer

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, Ljubljana
jasa.meh.peer@gmail.com

Abstract. Internet-of-Things (IoT) devices have been on the market for some time now, but they are not yet accessible to everyone, mainly due to price concerns. We wanted to create a proof of concept device and application that is more affordable. We designed and built basic appliances such as smart switch for turning light over the internet and smart blinds control that can control windows shades over the internet. Our simple IoT devices consist of simple electronics and run on microcontrollers that are based on the ESP8266 integrated circuit that can be connected over wireless network. The appliances connect to the network and they communicate using Message Queuing Telemetry Transport (MQTT) protocol over the public internet to the cloud service called Adafruit IO, which runs control application for connected appliances, save their data or inter-connect these appliances to other services like If That Then That (IFTTT) and WebHooks. We also tested our appliances and application to verify their security and reliability making it useful for development of IoT devices in the future.

1 Uvod

Želja po razvoju in uporabi aplikacij Interneta stvari (*ang. Internet-of-Things, IoT*) je zagotovo človeška lenoba in težnja k izboljšanju življenjskega sloga. Prvi koncept o povezanemu sistemu internetnih naprav je bil predstavljen leta 1982 na primeru prodajnega aparata s povezljivostjo na internet, ki je svojemu upravniku lahko sporočal ali je katerega izdelka že zmanjkalo in če so se novi izdelki že ohladili [6]. Od takrat naprej se je uporaba in implementacija različnih IoT aplikacij in naprav razširila na vsa področja, saj z njimi olajšamo tako tehnološke postopke kot vsakodnevna opravila.

Posebno področje IoT aplikacij so pametni domovi, vdelani pametni sistemi kot so KNX, Wireless in Loxone so za gospodinjstva v današnjem časupostali nepogrešljivi. Ker imajo vsi profesionalni sistemi sorazmerno visoko ceno za uporabo v manjšem ali starejšem stanovanju, smo želeli ta problem rešiti z uporabo vdelanih naprav, ki komunicirajo z internetom in nam omogočajo upravljanje izhodnih enot, ki so pravilno prilagojene za različne porabnike in opravila.

Za demonstracijske namene smo pripravili napravo in vezje za avtomatizacijo luči in zaves, ki jih lahko izdelamo sorazmerno enostavno, s pomočjo preprostih enostavno dobavljivih komponent. V drugem poglavju bomo predstavili uporabljene tehnologije, v tretjem strojno in programsko opremo ter komunikacijski protokol. V četrtem poglavju bomo predstavili rezultate testiranja naše aplikacije. Na koncu bomo zaključili z rezultati in razpravo o možnihnadgradnjah sistema.

2 Uporabljene tehnologije

2.1 Telemetrijski transportni protokol za čakalne vrste

Message Queuing Telemetry Transport (MQTT) je odprtokodni International Organization for Standardization (ISO) standard, ki je nizko-zahteven glede strojne opreme, in deluje po vzorcu publish/subscribe.

Protokol se zelo pogosto implementira v nizko nivojnih sistemih in v mobilnih aplikacijah, saj je uporabniku ključnega pomena zanesljivost in nizka poraba energije [7].

2.2 Platforma Adafruit IO

Oblačna storitev Adafruit IO se uporablja za povezovanje naprav in aplikacij preko interneta ter za shranjevanje in pošiljanje podatkov. Ponuja tudi mnogo drugih storitev, kot so prikazovanje stanj povezanih naprav v realnem času, upravljanje naprav in aplikacij preko interneta ter pošiljanje podatkov, kot je vremenska napoved, napravam in aplikacijam. Adafruit IO je pogosto uporabljena storitev, razlog za to je predvsem, ker je večji del storitve brezplačen.

2.3 Mikrokrmilnik ESP8266

ESP8266 je mikrokrmilnik starejše generacije, ki je cenovno ugoden, omogoča pa Wi-Fi povezljivost, in ima implementiran protokolni sklad Transmission-Control-Protocol/Internet-Protocol (TCP/IP). Gre za 32 bitni procesor, ki deluje z logičnimi napetostnimi nivoji na 3,3 V in s frekvenco programske ure 80 MHz ali 160 MHz.

Med izdelovalci je postal mikrokrmilnik zelo priljubljen v letu 2014, ko je na trg stopila prva razvojna platforma, ki ga je uporabljala. Od tistega leta dalje je na osnovi ESP8266 nastalo veliko različnih razvojnih platform, ki so delovanje in uporabo mikrokrmilnika izboljšale in poenostavile. Med najbolj priljubljene spadajo ESP 1.0, NodeMCu, Wemos, Adafruit Huzzah in še mnogo drugih.

3 Razvoj IoT sistema in naprav

S predstavljenim sistemom želimo pokazati, kako iz osnovnih elementov sestaviti napredno IoT aplikacijo. Pripravili smo dve prototipni IoT napravi, ki sta povezani na oblačno platformo preko lokalne brezžične povezave

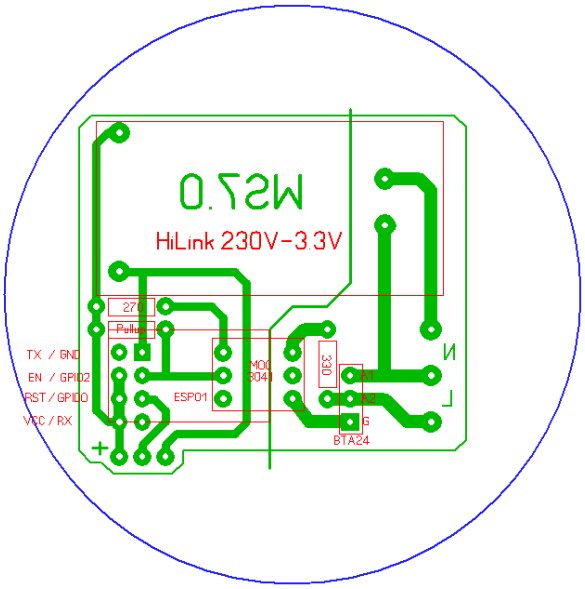
s protokolom MQTT. Glavna komponenta IoT naprav je mikrokrmilnik ESP8266, ki krmili ostale naprave in je povezan s spletom. Za spletno kontrolo smo uporabili brezplačno platformo Adafruit IO, na katero se povežejo vse naprave v našem sistemu.

Uporabo komponent in povezavo vezij je narekovalo več zahtev. Najosnovnejši zahtevi sta bili sposobnost brezžične povezave in avtomatizacija vodenih komponent, kot so stikala in motorji. Kasneje pa smo naprave pripravili v velikostnih okvirih potrebnih za montažo in prilagodili delovanje glede na specifično aplikacijo naprave.

3.1 IoT pametno stikalo

Zahteve pri izdelavi pametnega stikala so bile vezane predvsem na velikost samega vezja, saj je bilo potrebno celotno napravo vgraditi v inštalacijsko dozo kot nadomestilo klasičnemu stikalu. Ohranili smo tudi klasično delovanje stikala torej na pritisk z uporabo kapacitivnega ali navadnega preklopnega stikala.

3.1.1 Strojna oprema



Slika 1. Vezje za pametno stikalo na katerem lahko vidimo vse povezave med mikrokrmilnikom ESP01 in ostalimi uporabljenimi komponentami. Moder krog je za primerjavo vezja z velikostjo inštalacijske doze.

Pametno stikalo sestavljajo 3 W usmernik HiLink, ki iz 230 V izmenične napetosti pretvori na 3,3 V enosmerne napetost, ki napaja mikrokrmilnik. Izbrali smo mikrokrmilnik ESP 1.0, saj je najmanjša razvojna platforma, ki uporablja ESP8266 in ustreza vsem zahtevam.

Ima 2 Večnamenska Vhodno/Izhodna priključka (*ang. General-Purpose-Input/Output (GPIO)*) od katerih je priključek 2 uporabljen za vhodne enote, ki so lahko kapacitivno stikalo na dotik, navadno preklopno stikalo ali gumb, v našem primeru smo izbrali preklopno stikalo. Izhodne enote so povezane na drugi GPIO priključek, priključek 0, na katerega je bil sprva povezan rele, vendar smo zaradi zmanjšanja vezja naprave prešli na vezavo optotriak-triak [1], ki zavzame dosti manj prostora.

Izhodna enota preko glavnega triaka preklaplja izmenično breme v našem primeru je to žarnica, bi pa na takšen način lahko krmilili katerokoli napravo.
Vezje je veliko 41x37mm in ga je, glede na Sliko 1, mogoče vstaviti v inštalacijsko dozo.

3.1.2 Programsko delovanje

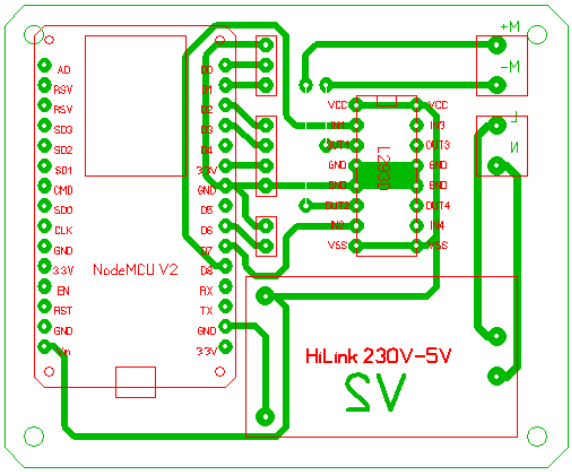
Program, ki se izvaja na mikrokrmilniku, vključuje štiri knjižnice in dve dodatni funkciji [2, 3]. Knjižnice omogočajo vzpostavitev WiFi povezave ter povezovanje preko MQTT protokola do storitve Adafruit IO v oblaku [4, 5, 10].

Ob priklopu na napajanje, mikrokrmilnik začne z izvajanjem programa. Sprva se izvede funkcija za inicializacijo sistema. Vzpostavi se povezava na brezžično omrežje in povezava na Adafruit MQTT posrednik. Določijo se tudi namembnosti priključkov mikrokrmilnika. Nato mikrokrmilnik preide v stanje preverjanja lastnega stanja in odzivanja na zahteve iz strežnika. V tem stanju kličemo dve funkciji za spreminjanje stanja na izhodu. Eno za ročen vklop razsvetljave, ki spremlja stanje na vhodnem priključku, ter drugo, ki spremlja podatke, ki prihajajo iz strežnika, in na podlagi teh ukazov spremeni stanje izhoda.

3.2 IoT pametne zavese

Zahteve pri tej napravi so bile drugačne kot pri IoT stikalu. Velikost naprave ni bila odločilnega pomena, zato smo lahko uporabili različico mikrokrmilnika ESP8266 z več priključki. Za krmiljenje motorja smo potrebovali vsaj 2, več priključkov pa smo potrebovali za generiranje prikaza na štiri znakovnem sedem segmentnem zaslonu.

3.2.1 Strojna oprema



Slika 2. Vezje za pametne zavese na katerem lahko vidimo vse povezave med mikrokrmilnikom NodeMCU 12E in ostalimi uporabljenimi komponentami.

Ker so bile zahteve za to napravo bolj obsežne, smo potrebovali več vhodno-izhodnih enot in posledično tudi

mikrokrmilnik z več GPIO priključki, kar je bil največji razlog, da smo se odločili za NodeMCU 12E.

Vhodni napravi sta ročica za krmiljenje in gumb za izbiro menijev. Ročica potrebuje dva GPIO vhoda, gumb pa je povezan na en GPIO vhod. Na vseh vhodnih enotah smo uporabili že integrirane *pull-up* upore in zato posledično nismo potrebovali dodatnih zunanjih.

Izhodne enote so modul s štirimi sedem segmentnimi zasloni (vsak za prikaz ene števke) za prikazovanje menija, za kar potrebujemo 4 GPIO priključke ter dva GPIO priključka, ki sta uporabljena za krmiljenje vhodov integriranega vezja L293D [9], ki smo ga uporabili za krmiljenje motorja, ki je v našem primeru 6V DC z reduktorjem.

3.2.2 Programsko delovanje

Program pametnih zaves je po veliki večini enak programu za pametno stikalo, sestavljajo ga iste knjižnice in glavne funkcije [2, 3].

Programa se razlikujeta v dveh knjižnicah in treh funkcijah. Program pametnih zaves vsebuje dve dodatni knjižnici in sicer eno za štiri znakovni sedem segmentni zaslon ter drugo, ki omogoča povezavo preko Omrežno časovnega protokola (*ang. Network Time Protocol (NTP)*) na server za podatke o času. Funkcije, *Manual*, *Automation* in *wifiControl*, ki spremljajo vhode in podatke iz interneta ter na podlagi le-teh spreminjajo stanje izhoda preko funkcije Output. Ter funkcija timeSet v kateri je meni, za nastavljanje podatkov za avtomatsko delovanje.

3.3 Komunikacijski protokol

MQTT protokol uporablja model publish/subscribe, ki je alternativa tradicionalnemu Hipertekstovnemu prenosnemu protokolu (*ang. HyperText Transfer Protocol (HTTP)*) modelu klient/server [7].

Najpomembnejša točka v MQTT sistemu je MQTT posrednik [8], katerega naloga je prenašanje podatkov ob spremembah od pošiljateljev do prejemnikov, v našem primeru je posrednik in prejemnik/strežnik ista stvar, Adafruit IO. Sporočila morajo imeti ustrezen naslov, da lahko posrednik pošlje podatek pravemu prejemniku.

3.4 Oddaljeno upravljanje

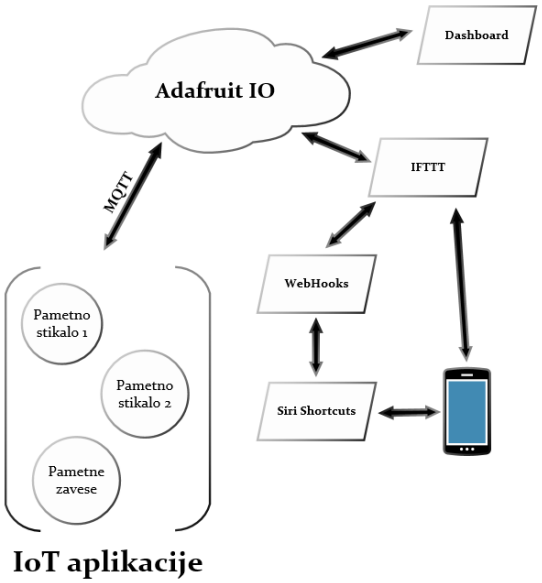
Za oddaljeno upravljanje IoT aplikacije uporabljamo storitev Adafruit IO. Na javno dostopni strežnik se povežemo preko MQTT protokola s pomočjo knjižnice *AdafruitIO*. Ker za uporabo storitve potrebujemo uporabniški račun in edinstven identifikator, smo za potrebe naše aplikacije to ustvarili na spletni platformi.

V platformi Adafruit IO spremljamo stanje izhodov in jih na kontrolni plošči tudi smiselno spreminjamo. Prav tako lahko s pomočjo platforme, naše naprave povežemo z drugimi spletnimi storitvami, kot so If That Then That (IFTTT).

IFTTT nam omogoča povezavo z vrsto ostalih oblačnih storitev, kot so npr.: Amazon Alexa, Instagram, lokacijski podatki, Facebook, Gmail in WebHooks. Mi smo implementirali uporabo IFTTT aplikacije na mobilnem telefonu tako, da smo povezali Adafruit IO in IFTTT ter s pomočjo *button widget*, navideznega gumba na namizju telefona, in WebHooks povezav spreminjali stanja na izhodnih enotah naših aplikacij.

WebHooks so zelo pogosto uporabljeni v aplikacijah, ko želimo, da se določena procedura izvede, ko odpremo ustrezno spletno povezavo. S pomočjo WebHooks in aplikacije za iPhone Siri Shortcuts smo povezali Siri in naše naprave, kar nam je omogočilo upravljanje IoT storitve z glasom.

4 Testiranje sistema



Slika 3. Shema sistema, na kateri lahko vidimo vsak posamezen gradnik sistema in kje v verigi delovanja stoji.

4.1 Splošno delovanje sistema

Delovanje sistema je glede na uporabljene komponente in odprtokodno okolje, skoraj brezhibno. Čas potovanja informacije od uporabnika do IoT naprave je v razponu 2-5 sekund. Do razlik pride zaradi moči brezžičnega signala in načina pošiljanja podatkov, od telefona do Adafruit IO platforme in do IFTTT ali Siri s pomočjo WebHooks (Slika 3).

Normalno uporabo sistema lahko razdelimo na ročno, na dotik in prostoročno. Kot smo že prej omenili lahko stikalo in zavese še vedno upravljamo s fizičnimi stikali torej ročno. S pomočjo IFTTT lahko s pritiskom na zaslon telefona spremenimo stanje izhodne enote, upravljanje na dotik ali pa sistem upravljamo s pomočjo glasovnih komand preko Siri, prostoročno upravljanje, v primeru, ko imamo roke polne ali pa smo zaposleni z čim drugim. Vsi ti različni načini upravljanja sistema nam dajo veliko izbire in s tem nam oziroma uporabniku izboljšajo izkušnjo.

4.2 Varnost sistema

Ko obravnavamo IoT sisteme, je zelo pomembno vprašanje varnosti. MQTT protokol ima zelo visok nivo varnosti, če uporabljamo šifriranje. Kljub temu da uporablja TCP protokol, je možna popolna ali samo delna enkripcija podatkov [7].

V naši storitvi se pojavi težava z varnostjo, ko povezavi med napravami in Adafruit IO dodamo druge

storitve preko IFTTT. Večina storitev, ki jih lahko povežemo z IFTTT deluje brez motenj, vendar v določenih primerih, kot so WebHooks, lahko pride do nezaželenega dogajanja. Ko naredimo spletno povezavo kot sprožilec za določeno dejanje naše IoT aplikacije, vsebuje spletna povezava ime dogodka in naš varnostni ključ. Če nekdo pozna ta dva podatka, nam lahko preko splošnega razporeda podatkov v spletni povezavi sproži naš dogodek. Seveda se lahko zgodi, da določena oseba vzpostavlja svojo spletno povezavo in ima isto ime dogodka ter se pri varnostnem ključu zmoti in vtipka našega nenamerno in nam tako sproži nezaželen vklop porabnika.

4.3 Zanesljivost sistema

Težave pri uporabi se še pojavijo pri zanesljivosti našega sistema, ki je predvsem odvisna od delovanja storitve v oblaku, načina delovanja MQTT protokola in močjo signala brezžičnega omrežja, na katerega so naše IoT aplikacije povezane. Če ena od naštetih ne deluje, pomeni, da podatki ne bodo prišli od IoT naprave na strežnik oz. iz telefona do IoT naprave. Odvisno od napake je ali se bodo podatki izgubili ali pa shranili v čakalno vrsto MQTT posrednika in se kasneje, ko bo povezava ponovno vzpostavljena, poslali k čakajoči napravi. V takem primeru se zgodi nepričakovana reakcija sistema na že pretečeni dogodek.

5 Zaključek

S tem člankom smo pokazali, da je mogoče s pomočjo preprostih komponent in odprtokodnih orodij izdelati preprosto, a vseeno široko uporabno IoT storitev. Med izdelavo aplikacije se je pokazalo, da je načrtovanje in izdelava enoslojnih vezij preprost postopek. Z nekaj različnimi prototipi smo izdelali končni produkt, ki deluje dobro glede na različne pogoje v okolju in z različnimi komponentami, ki jih upravljamo.

Našemu sistemu je težavo predstavljalo predvsem delovanje programa. Da je bilo možno izhode spreminjati na tradicionalen način s stikalom in preko interneta, smo morali preizkusiti kar nekaj različnih variacij kode. Seveda smo na koncu tudi programsko izvajanje optimizirali za najboljše možno delovanje. To sicer ni odpravilo vseh težav, saj zaradi komunikacijskega protokola in slabe brezžične povezave še kdaj pride do motenj v delovanju.

Strojno delovanje sistema in IoT aplikacije je že sedaj brez napak, vendar je rezkanje zgolj prototipni postopek, tako bi lahko v prihodnje naročili profesionalno izdelana večslojna vezja pri kakšnemu priznanemu proizvajalcu ali uporabili tehniko jedkanja.

Izboljšali bomo tudi kontroliranje zaves, saj enosmerni krtačni motorji niso zelo zanesljivi z vidika regulacije. Tako bomo za dviganje in spuščanje zaves rajši uporabili koračni motor in primerno krmilno enoto. S tem povečamo zanesljivost in natančnost ter tako dodamo več stopenj dviganja.

Nasprotno kot pri strojni opremi je programsko delovanje zagotovo potrebno še izboljšati, da bo delovanje aplikacije potekalo popolnoma brez motenj. Izboljšanja delovanja se lahko lotimo tako, da vztrajamo pri načinu, ki smo ga do sedaj uporabljali in popravimo napake, katere ima. Ali poskusimo drugo metodo pristopa in uporabimo mikroračunalnik (npr. Raspberry Pi) za vodenje storitev v oblaku in tako lokalno namestimo MQTT posrednik ter razvijemo lastno storitev podobno Adafruit IO za kontrolo in shranjevanje podatkov. Tako bi našo storitev popolnoma ločili od svetovnega spleta ter povečali varnost ter zanesljivost.

Zahvala

Za pomoč in nasvete se zahvaljujem as. Matevžu Hriberniku.

Literatura

[1] Horowitz Hill, The Art of Electronics. Cambridge University Press, 1980.

[2] Iztok Fajfar, Programiranje mikrokrmilnikov in programiranje 2, zapiski predavanj, UL FE, 2020.

[3] Tomaž Dobravec, abC, Založba FE in FRI, Ljubljana, 2010.

[4] Internet of Things Class, https://www.instructables.com/class/Internet-of-Things-Class/ (30.8.2020).

[5] Esp8266 control, https://www.instructables.com/id/Control-ESP8266-Over-the-Internet-from-Anywhere/ (30.8. 2020).

[6] Internet of things, https://en.wikipedia.org/wiki/Internet_of_things (30.8.2020)

[7] MQTT – vodilni komunikacijski protokol za Industrijo 4.0, https://www.tipteh.si/mqtt/ (20.8.2020).

[8] Matevž Hribernik, Razvoj varnega komunikacijskega posrednika in učinkovite arhitekture za povezovanje interneta stvari in spletnih aplikacij, magistrsko delo, UL FE, 2019.

[9] L293 dokumentacija, https://www.ti.com/lit/ds/symlink/l293.pdf (2.9.2020)

[10] ESP8266 as a WiFi Switch, https://www.instructables.com/id/ESP8266-Wifi-Switch/ (30.8.2020).

Index

Kranjc Matej: 384
Rimac Drlje Snježana: 393
Vidic Amadej: 78
Ćirić Dejan: 409
Amadasi Giovanni : 296
Andonovski Goran: 136, 140
Andrejc Žiga: 33
Antić Miloš: 189
Arnuga Sašo: 376
Atanasković Aleksandar: 409
B. Marković Goran: 425
Babić Zdenka: 405
Babnik Gregor: 340
Batagelj Boštjan: 25, 29, 33, 53, 74, 82, 86
Begeš Gaber: 256
Beguš Samo: 252, 288
Beković Miloš: 181
Bertalanič Blaž: 29
Bevc Jakob: 456
Bevk Nejc: 292
Blatnik Aljaž: 29, 103
Blažič Aljaž: 132
Blažič Boštjan: 205
Blažič Sašo: 136, 144
Blažun Vošner Helena: 309
Bogovič Tajda: 309
Bojkovski Jovan: 252
Bojovic Zivko: 421
Boneš Eva: 448
Borovina Nihad: 417
Bošković Borko: 301
Bošković Dušanka: 397, 405, 417
Bošnak Matevž: 136, 196
Bovcon Narvika: 123
Božić Marija : 425
Brest Janez: 301
Breznar Jan: 331
Breznik Miha: 452
Brugola Marcello : 280
Burnik Urban: 78, 115, 292, 393, 397
Cafnik Marko: 107
Cestnik Breda: 205
Chowdhury Amor: 177, 317, 322
Čebašek Eva: 164
Čegovnik Tomaž: 127
Čehovin Zajc Luka: 359
Čermelj Vid: 70, 260
D. Bedford Michel: 66
D. Ćertić Jelena: 425
Delić Vlado: 393, 405
Demirović Damir: 363
Dimc Franc: 45
Dobrišek Simon: 351
Dolinšek Marjan: 272
Drobnič Klemen: 231, 235
Đulabić Martina: 363
Emeršič Žiga: 119
Fabijan Žiga: 231
Fišer Rastko: 231, 235
Flisar Karel: 380
Fošnarič Miha: 326, 376
Gabrijelčič Tomc Helena: 107
Gams Andrej: 156
Gardašević Gordana: 413
Gartner Nina: 78
Gavric Nikola: 421
Gazič Jakob: 78
Genovese Jessica: 384
Geršak Gregor: 288
Gleich Dušan: 21, 37
Golob Ožbej: 344
Gošek Domen: 317
Gotlih Karl: 317
Gruden Timotej: 95
Hace Aleš: 148
Hamler Anton: 181, 211, 215
Hiti Miha: 256
Hrastič Aleksander: 70
Hribernik Matevž: 58
Hrovat Andrej: 66
Hudomalj Uroš: 8
Huljenić Darko: 397
Humar Iztok: 272
Igrec Dalibor: 177, 322
Jakus Grega: 95, 111
Jamšek Janez: 429, 438
Jan Matevž: 372
Jarm Tomaž: 372
Javornik Tomaž: 66
Jelenc David: 335
Jeon Myounghoon: 284
Jereb Tjaša: 49
Jesenik Marko: 215
Karer Gorazd: 132
Kastelic Gašper: 91
Kavran Domen: 444
Keršič Vid: 460
Keš Erik: 58
Kitak Peter: 41
Klančar Gregor: 136, 144, 196
Klanjšček Nejc: 376
Klenovšek Mario: 168
Kljun Maša: 367
Kmecl Peter: 172
Ko Sangjin: 284
Kocijan Juš: 185
Kokol Peter: 309
Kolmanič Simon: 99, 331
Koprivica Mladen: 393, 405, 425
Kos Andrej: 49
Kos Anton: 58
Kosec Špela: 429

Košir Andrej: 78
Košir Andrej: 348
Kovačič David: 235
Koželj Romi: 119
Kraljić David: 200
Krč Janez: 74
Krivec Tadej: 185
Kropivšek Leskovar Rebeka: 160
Kukar Matjaž: 248
Lavrič Andrej: 74
Lavrič Henrik: 231
Leitgeb Erich: 21
Lekić Milica: 413
Leskovec Jure: 2
Lešnjak Gorazd: 41
Likar Gal: 86
Lipuš Bogdan: 331
Loknar Martina: 144
Lombergar Peter: 380
Lončarević Zvezdan: 156
Luin Blaž: 45
Lukač Niko: 99
Mahnič-Kalamiza Samo: 380, 384, 388
Malačič Veronika: 107
Malajner Marko: 21
Malenšek Miha: 123
Maleš-Ilić Nataša: 405, 409
Maleš-Ilić Nataša P.: 397
Mandeljc Ana: 164, 256
Mandla Christopher: 8
Marcuzzi Francesco: 397
Markovič Jure Janez: 78
Martinčič Aljaž: 49
Mavsar Matija: 152
Meh Peer Jaša Vid: 464
Meža Marko: 115, 292, 348
Mihelj Matjaž: 164, 168, 172
Miklavčič Damijan: 372, 380, 384, 388
Miklavčič Peter: 82
Milavec Matej: 140
Milić Dejan: 409
Mlačnik Vid: 268
Mlinar Tomi: 62
Muhovič Jon Natanael: 355
Mujčić Aljo: 393, 405
Munih Marko: 164, 168, 172
Mušič Jernej: 53
Nabergoj David: 248
Nadri Chihab: 284
Nemec Mitja: 12, 219, 223
Nikolič Gregor: 4
Nimac Peter: 78
Ornik Timi: 301
Pahič Rok: 156
Peer Peter: 344
Penič Samo: 326, 376
Perkovič Marko: 45
Pernuš Martin: 351
Perš Janez: 433
Petkovšek Marko: 239
Petrič Tadej: 160
Planer Nejc: 301
Planinšič Peter: 21, 41
Plattner Markus: 8
Podobnik Janez: 164, 172
Policardi-Antoncich Franc: 264, 280, 296
Pongrac Blaž: 37
Popič Jan: 301
Pregelj Boštjan: 313
Pridgar Blaž: 123
Prislan Rok: 277
Pušnik Igor: 244, 268
Rabzelj Matej: 49
Radovanović Mladen: 25
Radović Marko: 41
Regoršek Žan: 326
Rihar Andraž: 223
Rimac-Drlje Snježana: 401
Rizvić Selma: 99
Rocculi Pietro: 384
Rozman Tanja: 429
Sadek Uroš: 322
Sanghavi Harsh: 284
Sešek Aleksander: 16
Simončič Aleš: 78
Skočaj Danijel: 355
Skuber Tadej: 16
Sobot Srdjan: 421
Sodnik Jaka: 127, 284
Solar Mitja: 4
Stojmenova Duh Emilija: 70, 260
Stojmenova Kristina: 127, 284
Stradovnik Saša: 148
Strle Gregor: 348
Stržinar Žiga: 313
Subhan Hameed Muhammad: 8
Suljanović Nermin : 393, 405
Sušin Denis: 227
Svetina Matija: 78
Šafarič Riko: 317
Šajn Luka: 119, 340
Šerifović-Trbalić Amira: 363
Škrjanc Igor: 136, 189, 313
Šlajpah Sebastjan: 168
Šmerc Rok: 388
Štruc Vitomir: 351
Štrumbelj Erik: 367
Štublar Jernej: 372
Šuran David: 309
Tabernik Domen: 355
Tasić Dejan: 177
Tavčar Rok: 252
Teršek Matija: 367
Thebault Marion: 264
Tomažič Sašo: 127
Tomažič Simon: 136
Tonello Andrea: 393, 405
Trbušić Mislav: 211, 215

Trilar Jure: 260
Trlep Mladen: 215
Troha Miha: 200
Trontelj Janez: 16
Trost Andrej: 433
Umek Anton: 58
Urankar Bernarda: 438
Urška Levac: 78
Valič David: 115
Vidmar Rok: 78
Vitek Matej: 344
Vizjak Jakob: 181
Vuk Balič Kristjan: 272
Vukobratovic Dejan: 421
Vöröš Matjaž: 305
Zajc Matej: 393, 397, 405
Zajec Peter: 239
Zaletelj Janez: 272
Zamuda Aleš: 305
Završnik Jernej: 309
Završnik Tadej: 309
Zdešar Andrej: 136, 189, 196
Zejnilović Amar: 205
Zhezhov Anton: 211
Zoretič Uroš: 111
Zukic Merima: 417
Žagar Drago: 397, 401, 405
Žalik Borut: 99
Žibert Janez: 326
Žižek David: 372