ERK 2018

Portorož, Slovenija, 17. - 18. september 2018

Zbornik sedemindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2018

Proceedings of the Twenty-seventh International Electrotechnical and Computer Science Conference ERK 2018

Vabljena predavanja / Invited Lectures
Elektronika / Electronics
Telekomunikacije / Telecommunications
Multimedija / Multimedia
Avtomatika in robotika / Automatic Control and Robotics
Modeliranje in simulacija / Modelling and Simulation
Akustika in elektroakustika / Acoustics and Electroacoustics
Močnostna elektrotehnika / Power Engineering
Merilna tehnika / Measurement - ISEMEC 2018
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 sedemindvajsete Elektotehniške in računalniške konference ERK 2018,
17. - 18. september 2018, Portorož, Slovenija

Proceedings of the Twenty-seventh Electrotechnical and Computer Science Conference ERK 2018,
17 - 18 September 2018, 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.

IEEE Region 8
Slovenska sekcija IEEE
Fakulteta za elektrotehniko v Ljubljani

Sporočilo predsednika konference

v imenu Organizacijskega odbora imam prijetno dolžnost, da Vas ponovno pozdravim tokrat že na sedemindvajseti Elektrotehniški in računalniški konferenci - ERK 2018. Konferenca bo potekala od 17. do 18. septembra 2018 v hotelu Bernardin, v Portorožu v Sloveniji.

Naša konferenca se je utrdila kot vsakoletno srečanje strokovnjakov na elektrotehniškem in računalniškem področju. Pod okriljem Slovenske sekcije IEEE združuje enkrat letno interese velikega števila strokovnih društev. Strokovna društva skrbijo za program konference na svojih delovnih področjih, hkrati pa organizirajo svoje sestanke, okrogle mize in vabljena predavanja, kjer se je mogoče pogovoriti o strokovnih problemih, o tehnološkem in znanstvenem razvoju, mednarodnem povezovanju, izobraževalnih programih, financiranju, zaposlovanju in podobno. Zato bo tudi letošnje srečanje ERK 2018 mozaik, sestavljen iz mnogih drobnih, ozko usmerjenih strokovnih srečanj, ki bodo kvalitetna; skupno dogajanje pa bo omogočalo mnoga interdisciplinarna srečanja in obisk vabljenih predavanj, zanimivih za širši krog poslušalcev.

Zanimanje za srečanje je že 27 let. Tako vidimo, kako potrebujemo nek forum, kjer lahko naši raziskovalci predstavijo strokovnemu krogu svoje delo, ga primerjajo z drugimi prispevki in dobijo kritične pripombe in ustvarjalne nasvete za nadaljnje delo. Na ta način poročevalec dograjuje svoje delo in spozna skupine in posameznike za sodelovanje. Mladi avtorji vadijo nastop v sredini, kjer živijo, kjer se želijo uveljaviti.

Pri tem pa vsi skupaj razvijamo tudi slovensko tehniško besedo, zato se slovenskemu jeziku na konferenci nismo odpovedali. Skupno si prizadevamo, da ostaja ta jezik tudi naš dopoldanski jezik v službi, na delovnem mestu.

Vabljena predavanja smo postavili kot plenarna predavanja. Ker takrat posameznih sekcij ni, se lahko teh predavanj udeležimo vsi. Pričakujemo velik interes udeležencev in plodno razpravo z uglednimi predavatelji iz Slovenije in tujine.

Razpored referatov in predavanj je natisnjen v tem programu in na spletni strani **http://www.ieee.si/erk**.

Vabim Vas, da se udeležite konference v celoti, saj si prizadevamo, da bi naredili vsebino strokovnega dogajanja čim bolj bogato. Zato lahko vsak udeleženec izbira in se pri tem strokovno izpopolnjuje.

Prav bi bilo, da ostanete na konferenci oba dneva. Predvsem pa Vas vse vabim na otvoritveno svečanost. Nenazadnje naj nas naši gostje vidijo zbrane ob začetku našega dogodka.

Ob samem konferenčnem dogajanju bo dovolj priložnosti, da tudi neformalno navežemo medsebojne stike. To je tudi pomemben namen te konference poleg izobraževanja ob vabljenih predavanjih in osebne uveljavitve pri prezentaciji lastnega raziskovalnega dela.

Veseli smo, da se boste konference ERK 2018 udeležili in prizadevali si bomo, da boste z delom na konferenci zadovoljni in v našem krogu te dni tudi uživali.

Andrej Žemva
Predsednik konference

Baldomir Zajc
Častni predsednik konference

Organizatorji / Organizers

- IEEE Region 8
- 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 2018),
- 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

Častni predsednik konference / Honorary Conference Chairman
Baldomir Zajc

Podpredsednik konference / Conference Vicechairman
Matej Zajc

Predsednik program. odbora / Program Committee Chairman
Janko Drnovšek

Programski odbor / Programe Committee

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

Predsednik odbora za tisk / Publications Chairman
Andrej Trost

Recenzenti / Reviewers

Vanja Ambrožič
Boštjan Batagelj
Aleš Belič
Aljaž Blatnik
Sašo Blažič
Jovan Bojkovski
Božidar Bratina
Patricio Bulić
Gregor Burger
Urban Burnik
Luka Čehovin
Matjaž Divjak
Žiga Emeršič
Ratko Fišer
Helena Gabrijelčič Tomc
Gregor Geršak
Giovanni Godena
Klemen Grm
Jože Guna
Aleš Holobar
Iztok Humar
Grega Jakus
Marko Jankovec
Rihard Karba
Gorazd Karer
Anton Kos
Matej Kristan
Janez Križaj
Matevž Kunaver
Vito Logar
Uroš Lotrič
Alenka Maček Lebar
Marko Meža
Peter Miklavčič
Nenad Muškinja
David Nedeljković
Klemen Pečnik
Tomaž Perme
Janez Perš
Ratko Pilipović
Matevž Pogačnik
Franc Policardi-Antoncich
Matevž Pustišek
Danijel Skočaj
Boštjan Slivnik
Sara Stančin
Stanko Strmčnik
Luka Šajn
Božidar Šarler
Vitomir Štruc
Domen Tabernik
Andrej Trost
Anton Umek
Danjel Vončina
Matej Zajc
Janez Zaletelj
Andrej Zdešar
Borut Zupančič
Andrej Žemva

Vabljena predavanja

Invited Lectures

Challenges in Electric Motor Design for Automotive Applications

Damir Žarko

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
e-mail: damir.zarko@fer.hr

Abstract. Development of efficient and affordable power trains for electric and hybrid vehicles is in the focus of many researchers in industry and academia and is motivated by government regulations and consumer demands. The key components of a power train are electric motor, power converter, batteries and associated control systems. Electric traction motors are required to have the lowest possible size and weight, provide sufficient continuous and short-term torque and power to satisfy the requirements of vehicle kinematics in various conditions, and endure high rotational speeds without exceeding thermal constraints of winding insulation or permanent magnets, mechanical constraints of the rotor, and voltage and current limits of the battery and inverter. These conflicting requirements present a considerable challenge for machine designers. On an example of interior permanent motor most commonly used in automotive applications this lecture will demonstrate a multiphysical approach to traction motor design process combined with evolutionary optimization as a design tool capable of addressing the most important design challenges.

Game design for situated development of students' creativity

Maja Pivec

FH JOANNEUM University of Applied Sciences
Institute of Design and Communication
Alte Poststraße 152, 8020 Graz, Austria
e-mail: maja.pivec@fh-joanneum.at

Abstract. Industry is placing new challenges in front of educational institutions, emphasizing the necessity for young people entering the workforce to have the ability to work in teams, communicate and collaborate and be proactive i.e. soft skill set that educational institutions did not provide in needed extent to their graduates. To achieve this, educational institutions are increasingly looking into how to harness elements and game mechanics of leisure games and adapt them into appealing and motivating learning that transfer into their future work settings.

As an example of how this creativity and team work are needed and could possibly transfer into the real world, we can use the advancements and integration of AI, speech recognition, and computational power that have evolved from experimental research environments into products for everyday use. Furthermore, we witness the use of virtual assistants in everyday life, but how do we establish an emotional state with a virtual tutor as we often do with real people, and could we design virtual assistants' personas to enable users to make choices based on technology as well as the VA personality they prefer.
All this will take team work from UX designers as well as talented developers, both being creative and both working towards the same goals.

This talk will present four game-based learning designs with focus on situated development of students' creativity and active participation in the learning process. Presented cases from formal and non-formal learning setting, are based on role-play, interdisciplinary team work and co-creation, using games and gamification as additional motivation to engage students in project and problem based learning.

Total Audience Measurement

Matjaž Finc

Nielsen Lab d.o.o, Obrtniška ulica 15, Koper
e-mail: matjaz.finc@gmail.com

Abstract. Modern world is flooded with visual media. Consumer interest is rapidly shifting from broadband television broadcast towards more individually tailored content like video on demand and social media video streaming. Providers of video and commercial adds must be able to place the content of interest to specific user groups and individuals at the most convenient time. To get feedback about the penetration of content delivery, it requires precise measurements of audience coverage.

This lecture will outline the challenges and current state-of-the-art of technology for measuring video content delivery to audience in a household, with focus on television from broadband broadcasts to over the top video content and covering other fixed and mobile streaming video consumer devices (personal computers, mobile phones, tablets).

Single and Multiobjective Clustering

Ujjwal Maulik

Department of Computer Science and Engineering, Jadavpur University
Department of Computer Science and Engineering
Kolkata, West Bengal, India
e-mail: umaulik@cse.jdvu.ac.in

Abstract. Data clustering is a popular unsupervised data mining tool that is used for partitioning a given data set into homogeneous groups based on some similarity/dissimilarity metric. There are several single objective clustering algorithms available and popular in the literature including KMean and FCM. Although both KMean and FCM are very popular algorithms, they may stuck to local optima depending on the choice of initial cluster centers. In this lecture first we will demonstrate how Metaheuristic techniques can be used to solve the problem of KMean/FCM. Result will be demonstrated for pixel classification of satellite images.

In the second part of the problem we will discuss Multiobjective clustering, in which multiple objective functions are simultaneously optimized. Selecting one solution from the set of Pareto Optimal solutions is always a critical issue. We will also discuss how machine learning techniques like Support Vector Machine (SVM) can be used to combine the Pareto Optimal solutions to evolve a better solution. The result will be demonstrated for classification of Gene Micro Array Data.

Elektronika

Electronics

Digitalni signalni generator na napravi STEMlab

Matija Mavsar, Andrej Trost

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, 1000 Ljubljana
E-pošta: matija.mavsar@gmail.com, andrej.trost@fe.uni-lj.si

Digital signal generator on the STEMlab device

STEMlab is a low-cost programmable logic platform intended to replace many expensive laboratory measurement and control instruments. The STEMlab produced by Red Pitaya includes 2-channel oscilloscope and signal generator application. The open source software enables design of custom laboratory instruments. Our contribution is extension of the digital signal generator to an optional number of channels with arithmetic operations between the channels and the input signals. We implemented the changes at the levels of hardware description language, API interface and SCPI server. The extended signal generator can be used for digital signal processing experiments with 125MS/s RF outputs.

1 Uvod

Signalni generator je merilna oprema, ki jo skupaj z osciloskopom potrebujemo pri eksperimentiranju in razvoju elektronike. Eksperimentiranje z elektroniko je v izobraževalnih ustanovah del uvajanja v področja znanosti in tehnologije (ang. STEM - *Science, Technology, Engineering and Mathematics*). Namesto dragih merilnih instrumentov je podjetje Red Pitaya razvilo napravo STEMlab, ki deluje kot dvokanalni osciloskop in signalni generator [1]. Naprava ima odprto programsko kodo, ki omogoča razvoj različnih merilnih aplikacij [2][3].

Naprava STEMlab-14 ima dva vhodna in dva izhodna kanala s frekvenco zajemanja oziroma generiranja signalov 125 MHz, ločljivost signalov pa je 14-bitna. Na plošči je tudi razširitveni priključek z digitalnimi in analognimi vhodi ter izhodi. Jedro naprave je sistem na čipu (SoC) iz družine Xilinx Zynq-7000 [4] s programirljivim poljem logičnih vrat (ang. FPGA - *Field-Programmable Gate Array*), dvojedrnim procesorjem in bralno-pisalni pomnilnikom (RAM). Na procesorju teče operacijski sistem Linux s spletnimi aplikacijami za upravljanje naprave preko omrežja. Priključitev na omrežje je mogoča z Ethernet kablom ali z Wi-Fi USB priključkom.

V članku predstavljamo nadgradnjo signalnega generatorja ASG (ang. *Arbitrary Signal Generator*), pri katerem smo želeli povečati število kanalov, ki jih lahko uporabljamo za generacijo signalov, in omogočiti dodatne operacije med njimi. Naš cilj je bil razvoj strojnega opisa modula ASG s parametrično nastavljivim številom kanalov in ločljivostjo izhodov. Poleg logike generatorja signalov so potrebni tudi programski vmesniki za nastavljanje parametrov med delovanjem generatorja in izvajanje operacij med izbranimi kanali.

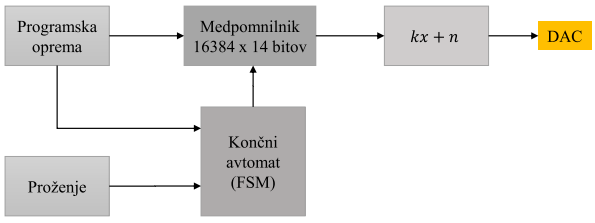
2 Signalni generator

Tabela 1 opisuje lastnosti obstoječega signalnega generatorja na napravi STEMlab. V osnovni izvedbi sta na voljo dva kanala z možnostjo nastavljanja oblike, amplitude, frekvence in enosmerne komponente signala.

Tabela 1: Lastnosti signalnega generatorja

Lastnost	Vrednost
Vzorčna frekvenca	125 MHz
Ločljivost DAC	14 bit
Pomnilnik vzorcev	16384
Napetostno območje	$\pm$ 1 V
Najvišja strmina napetosti	200 V / µs
Izhodna impedanca	50 Ω

Napravo STEMlab poganja sistem na čipu z zmogljivo programirljivo logiko in procesnim sistemom. Digitalna sinteza signala se izvaja v logiki programirljivega vezja, programska oprema pa nudi vmesnik za nastavitve parametrov izhoda in določanje oblike periodičnega signala. Slika 1 prikazuje glavne gradnike za izvedbo enega kanala.



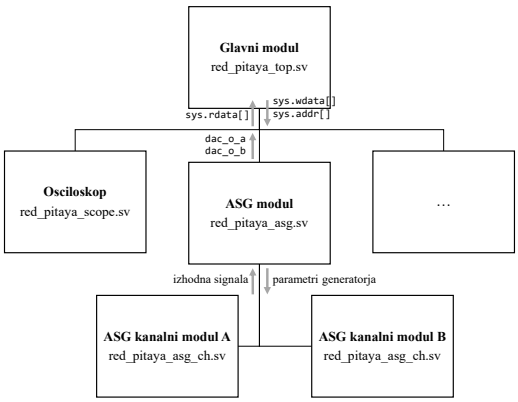
Slika 1: Gradniki enega kanala signalnega generatorja

Vsak kanal signalnega generatorja ima svoj končni avtomat (FSM), ki skrbi za periodično branje vzorcev iz medpomnilnika. Programska oprema ob inicializaciji zapiše obliko signala v medpomnilnik, med delovanjem pa nastavlja frekvenco, način proženja in vrednosti amplitude ter enosmerne komponente ($kx + n$). Izhod ob-

stoječega generatorja je neposredno povezan na digitalno-analogni pretvornik (DAC).

2.1 Modul signalnega generatorja

Signalni generator je na vezju FPGA realiziran s strojno opisnim jezikom SystemVerilog kot eden izmed 8 modulov oziroma IP (ang. *Intellectual Property*) komponent, vsaka pa ima na voljo 4 MB naslovnega prostora [1]. Uporaba že vnaprej načrtovanih in sintetiziranih IP komponent omogoča lažjo integracijo v projekte in tako zmanjša njihovo kompleksnost [5]. Hierarhijo modulov prikazuje slika 2.



Slika 2: Hierarhija modulov v jeziku Verilog

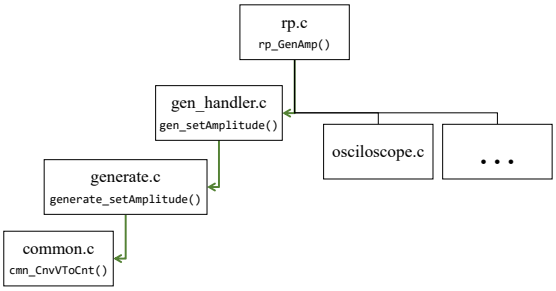
Spreminjanje parametrov generatorja poteka s pošiljanjem vrednosti na ustrezne naslove s pomočjo jezikov C (vmesnik API) ali Python (SCPI strežnik). Pri tem FPGA logika prebere naslove za pisanje in podatke zapiše v ustrezne registre. S procesorjem je povezana preko vmesnika AXI (ang. *Advances eXtensible Interface*), ki sestoji iz naslovnih in podatkovnih vodil za pisanje ter branje [6]. Na napravi je v FPGA logiki narejen pretvornik iz kompleksnega AXI vodila v enostavnejše vodilo *sys*. Nanj so v glavnem modulu priključene komponente programirljive logike (slika 2) in tako predstavlja sistemsko vodilo, ki omogoča zapisovanje v registre.

V ASG modulu je koda, potrebna za zapisovanje parametrov generatorja v registre in branje iz njih. Pri pisanju se glede na naslov, prejet preko naslovnega vodila, v ustrezen register zapiše trenutna vrednost na podatkovnem vodilu za pisanje. Pri branju registrov pa se vrednosti glede na naslov pošljejo na podatkovno vodilo za branje. V modulu je za vsak kanal generatorja definiran nov ASG kanalni podmodul, v katerega so posredovani parametri generatorja, ki skrbi za zajemanje vrednosti iz medpomnilnika in pošiljanje teh vrednosti na digitalno-analogni pretvornik. Vrednosti v medpomnilnik zapišemo s pomočjo programske opreme (prek signalov v ASG modulu), ki tudi nastavlja krmilni avtomat za branje vrednosti, način proženja in linearno transformacijo izhodov.

2.2 Nastavljanje parametrov signalnega generatorja

API vmesnik je napisan v jeziku C, sestavljajo pa ga različne datoteke, namenjene delovanju komponent programirljive logike. V glavni datoteki *rp.c* so združene funkcije, ki posredno upravljajo s parametri vseh komponent.

Nastavljanje parametrov signalnega generatorja je opisano v datoteki *generate.c*, kamor se preko glavnih funkcij iz *rp.c* posredujejo parametri. Ti se nato združijo v skupno strukturo in se s pomočjo kazalcev usmerijo na ustrezne naslove, ki jih FPGA logika dekodira in vrednosti zapiše v ustrezne registre. Na podoben način se v medpomnilnik zapišejo tudi vrednosti želenih izhodnih signalov, vendar s pomočjo drugih kazalcev. Potek klicanja funkcij za nastavitev amplitude je prikazan na sliki 3.



Slika 3: Potek klicanja funkcij za nastavitev amplitude

Spreminjanje parametrov ASG lahko poteka tudi z jezikom Python preko SCPI strežnika. Pri tem na ciljno napravo pošiljamo ukaze v obliki znakovnih nizov, po prejetju pa se kliče pripadajoča funkcija z želenimi parametri. Izvorna koda je napisana v jeziku C in se nanaša na kodo za API. Za delovanje generatorja je tako pomembna datoteka *generate.c*, v kateri so funkcije, ki ob prejetju določenega znakovnega niza kličejo ustrezno funkcijo iz vmesnika API. V datoteki *scpi-commands.c* pa je koda, ki določa, katera funkcija se bo klicala ob posameznem prejetem znakovnem nizu.

3 Večkanalni signalni generator

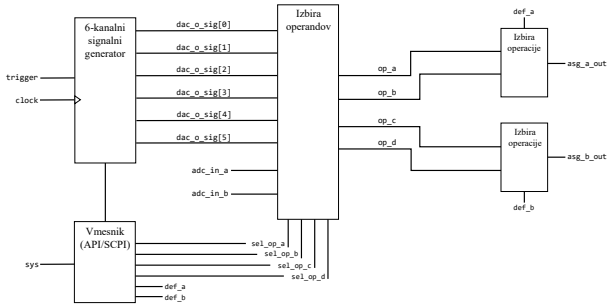
Blokovna shema večkanalnega signalnega generatorja je prikazana na sliki 4. Na sredini je prikazan blok z osmimi vhodnimi signali (šest iz signalnega generatorja in dva iz analognih vhodov naprave STEMlab), spodaj pa so štirje parametri, ki določajo, katere signale želimo izbrati kot operande. Vsak par operandov gre na blok procesnega modula, kjer vhodni parameter (*def_a* ali *def_b*) določa operacijo, ki se bo na paru operandov izvedla. Rezultata se nato pošljeta v glavni modul in na analogna izhoda naprave.

3.1 Logično vezje

Opis večkanalnega generatorja temelji na nadgradnji posameznih modulov referenčne kode v jeziku Verilog, ki je dostopna na GitHub repozitoriju Red Pitaye.

3.1.1 Glavni modul

V glavnem modulu smo dodali le kodo, ki s pomočjo pogojnega operatorja (ang. *ternary operator*) določa, ali naj se na vhod modula za osciloskop pošljeta izhodna



Slika 4: Blokovna shema večkanalnega generatorja

signala generatorja ali vrednosti iz analognih vhodov naprave.

3.1.2 Modul ASG

Pri opisu modula ASG smo v začetku deklaracije definirali parametra za nastavljanje števila kanalov in bitov ter pomožni parameter za maksimalno število bitov, dodali pa smo tudi dva 14-bitna vhoda za signala z analognih vhodov naprave, posredovana iz glavnega modula, in signal *send_scope* za pošiljanje izhodnih signalov na vhod osciloskopa. Nato smo deklarirali matriko vseh kanalov generatorja, signale, ki predstavljajo operande, in nove parametre iz tabele 2. Dodali smo stavke za izbiro operandov izmed vseh kanalov generatorja in analognih vhodov, ter spremenili način proženja kanalov, tako da se vsi lahko ponastavijo naenkrat.

Tabela 2: Registri novih parametrov

Parameter	Št. bitov	Opis
sel_op_a	3	izberi operand A
sel_op_b	3	izberi operand B
sel_op_c	3	izberi operand C
sel_op_d	3	izberi operand D
def_a	3	operacija med A in B
def_b	3	operacija med C in D
send_scope	1	pošiljanje na vhod
calib_amp_a	14	kalibracija prvega izhoda
calib_dc_a	14	
calib_amp_b	14	kalibracija drugega izhoda
calib_dc_b	14	
r_trig	$4 \cdot CN$	proženje kanalov
r_rst	CN	ponastavitev števcev

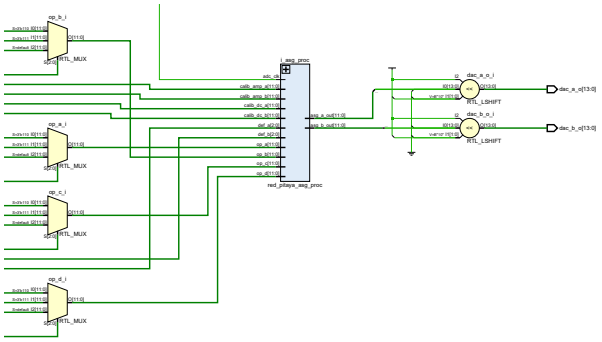
Kodi za zapisovanje in branje registrov smo dodali še nove parametre generatorja ter celotno kodo prilagodili za poljubno število kanalov in bitov.

3.1.3 Procesni modul

Za izvajanje aritmetičnih operacij med izbranimi operandi smo ustvarili nov procesni modul z dvema izhodoma (rezultata operacij), v katerega posredujemo parametre iz ASG modula (parametre za število kanalov, bitov, kalibracijo in izbiro operacij) in signale operandov. Deklarirali smo tudi vmesne signale za izvajanje operacij med operandi.

V *always* bloku je opisano računanje z operandi glede na vrednost parametra za izbiro operacije. Na rezultatu se nato upoštevajo kalibracijski parametri in decimalna mesta (pomik v desno). Končna signala sta zatem posredovana v glavni modul in na DA pretvornik.

Slika 5 prikazuje blokovno shemo procesnega modula, ki smo jo sintetizirali s programom Xilinx Vivado. Vezje sestavljajo multiplekserji za izbiro operandov, ki se skupaj z ostalimi parametri posredujejo v procesni modul, iz katerega izhajata dva izhoda, ki se pomakneta za ustrezno število bitov.



Slika 5: Shema sintetiziranega vezja s procesnim modulom

3.2 Aplikacijski vmesnik

Za uporabo nadgrajenega logičnega vezja signalnega generatorja smo naredili spremembe in napisali nove funkcije v vmesniku API.

- V datoteki *rp.h* smo dodali konstanto za število kanalov generatorja, podatkovni tip za izbiro kanala ter deklaracije novih funkcij.
- V *rp.c* smo definirali nove funkcije, ki se sklicujejo na funkcije iz datoteke *generate.c* z želenimi parametri.
- V *generate.h* smo dodali konstante (npr. začetni naslov za zapis novih parametrov v registre), definirali strukturo za spremenljivke novih parametrov (tabela 2) in deklarirali nove funkcije.
- V *generate.c* smo kodo posplošili za poljubno število kanalov, spremenili način proženja in kalibracije izhodnih signalov ter dodali še pošiljanje novih parametrov v registre FPGA logike.
- V datotekah *gen_handler.c* in *gen_handler.h* se je koda nekoliko poenostavila, saj smo namesto ločenega opisa delovanja kanal uporabili le indeks želenega kanala. Dodali smo funkcijo za ponastavitev kanalov in definirali nov parameter, ki ima v binarnem zapisu enice na mestu kanalov, ki jih želimo ponastaviti.

3.3 Strežnik SCPI

Delovanje signalnega generatorja se lahko omogoči tudi preko SCPI strežnika s pomočjo pošiljanja ukaznih nizov z jezikom Python. V nadaljevanju so opisani koraki za nadgradnjo obstoječe kode strežnika.

- V datoteki *generate.c* SCPI strežnika smo definirali funkcije za klicanje novih funkcij istoimenske datoteke API vmesnika. Vhodni parameter je znakovni niz iz katerega razberejo vrednosti in kličejo ustrezne funkcije API.
- V datoteki *scpi-commands.h* smo na seznam vseh obstoječih znakovnih ukaznih nizov dodali nize, ki pripadajo novim funkcijam. Tako se ob pošiljanju nekega niza z jezikom Python izvede ustrezna C funkcija.

4 Rezultati

Tabela 3 prikazuje zasedenost vezja FPGA po sintezi modula večkanalnega signalnega generatorja v programu Vivado. Sintetizirali smo eno- do osemkanalne generatorje s 14 bitno ločljivostjo.

Tabela 3: Zasedenost originalnega in nadgrajenega FPGA vezja v odstotkih glede na število kanalov

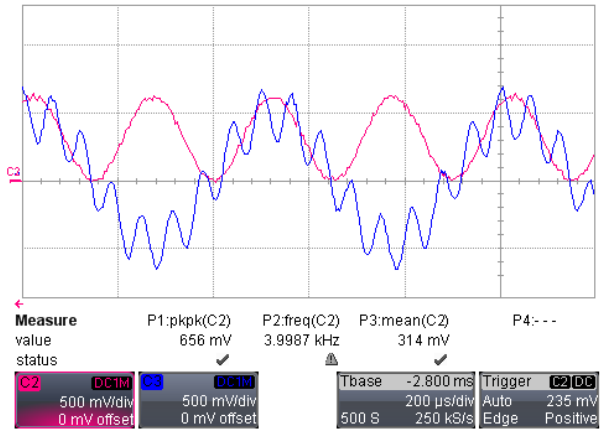
Tip enote	Orig.	Nadgrajen generator za različna št. kanalov							
		1	2	3	4	5	6	7	8
LUT – logika	5.18	4.28	6.91	10.10	12.72	15.19	18.07	20.46	23.30
Registri – flip-flop	2.61	1.76	3.00	4.27	5.50	6.78	8.01	9.15	10.37
F7 MUX	0.14	0.00	0.07	0.38	0.40	0.48	1.06	1.07	1.06
BRAM	23.33	11.67	23.33	35.00	46.67	58.33	70.00	81.67	93.33
DSP	2.50	6.25	7.50	8.75	10.00	11.25	12.50	12.50	12.50

Razvidno je naraščanje uporabljenih enot s številom kanalov generatorja. Nadgrajen generator za dva kanala porabi v primerjavi z originalnim nekoliko več LUT in registrov, trikratno število DSP-jev, enako število BRAM enot in pol manj F7 MUX.

Pri sintezi celotnega sistema se pri 6 kanalih in več uporabijo tudi LUT v obliki pomnilnika. Pri 6 kanalih je uporabljenih 0.02 %, pri 7 kanalih pa kar 85.35 % enot, saj pri 6 kanalih začne zmanjkovati BRAM enot.

4.1 Preizkus na napravi STEMlab

Slika 6 prikazuje oscilogram izhodnih signalov izmerjen na izhodu naprave STEMlab pri 4 kanalnem signalnem generatorju z nastavljenimi vhodi amplitude 0.8 V, 0.5 V in 0.2 V in frekvence 2 kHz ter 16 kHz.



Slika 6: Meritev izhodnega signala. Moder signal prikazuje vsoto dveh izhodov, pri čemer ima eden nižjo amplitudo in višjo frekvenco, rdeč signal pa je kvadrat sinusnega signala.

5 Zaključek

V pripevku smo opisali postopek nadgradnje strojne in programske opreme merilne in krmilne naprave STEMlab na primeru razvoja večkanalnega signalnega generatorja. Predstavili smo postopek nadgradnje kode na strojnem (Verilog) in programskem nivoju (C in Python), ki je podrobneje opisan v [7].

Z nadgradnjo signalnega generatorja smo uspešno omogočili izbiro poljubnega števila kanalov generatorja, dodali možnosti izbire vhodnih kanalov za operande in omogočili spreminjanje števila bitov, s katerim so signali opisani. Signalni generator smo preizkusili na napravi STEMlab z različnimi nastavitvami ter ugotovili, da je njegovo delovanje v skladu z želenimi lastnostmi.

Nadgradnja generatorja je lahko uporabna npr. pri modulaciji in obdelavi signalov ter nasploh pri raziskovanju oziroma eksperimentiranju, kjer potrebujemo nadzor nad več signali in sprotno izvajanje aritmetičnih operacij v realnem času.

Razvoj strojne in programske opreme za digitalno obdelavo signalov v sistemu SoC, ki je osrednji del naprave STEMlab zahteva specifična znanja iz obeh področij in poznavanje tehnologije. Možna rešitev za tiste uporabnike naprave, ki nimajo specifičnih znanj, je sinteza vezja iz programske kode in avtomatsko sestavljanje sistema iz ključnih komponent IP [8]. Večkanalni digitalni generator signalov je ena izmed takšnih komponent in ga bomo v nadalnjem delu uporabili pri razvoju računalniških orodij za eksperimentiranje z napravo STEMlab.

Literatura

[1] Red Pitaya. "Red Pitaya documentation", 2016. [Online]. Dosegljivo: http://rpdocs.readthedocs.io/en/latest/. [Dostopano: 7. 1. 2018].

[2] M. Ossmann: Red Pitaya – not just a USB scope module, Elektor, December 2014, str. 38–42.

[3] M. Cimerman, Z. Topčagić et al.: Laboratorijski merilni sistem na osnovi naprave Red Pitaya, zbornik ERK 2014, 22. - 24. september 2014, Portorož, Slovenija, zv. B, str. 215.

[4] L. H. Crockett, R. A. Elliot, M. A. Enderwitz, R. W. Stewart: The Zynq Book: Embedded Processing with the Arm Cortex-A9 on the Xilinx Zynq-7000 All Programmable Soc, Strathclyde Academic Media, UK, 2014.

[5] T. Thomas, "Technology for IP reuse and portability", *IEEE Design & Test of Computers*, vol. 16, št. 4, str. 7-13, 1999.

[6] Xilinx, "AXI reference guide", podatkovni list, marec 2011 [revidirano: januar 2012].

[7] M. Mavsar: Večkanalni signalni generator na merilni napravi STEMlab, Magistrsko delo, FE, Ljubljana 2018.

[8] A. Trost, A. Žemva: Pipeline circuit synthesis from Python code, 6th Mediterranean Conference on Embedded Computing (MECO), 11 - 15 junij, 2017, Bar, Črna Gora, str. 320-323.

Programirljiva prstna vrtavka za prikaz besedila ali grafike po principu vztrajnosti vida

Matej Nogić, Ernest Gungl, Zmago Brezočnik

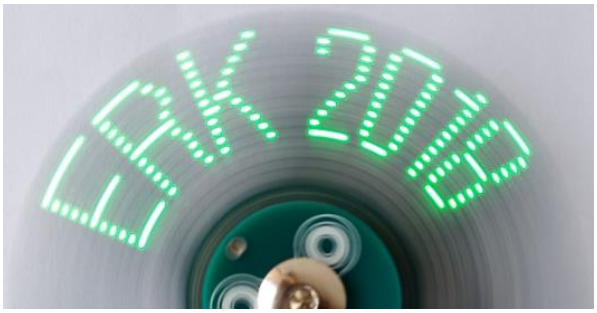
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Koroška cesta 46, 2000 Maribor
E-pošta: matej.nogic@student.um.si, ernest.gungl@um.si, zmago.brezocnik@um.si

Programmable Persistence of Vision Fidget Spinner for Display of Text or Graphics

Abstract. *This paper presents our four-wing fidget spinner, a finger-spun propeller that displays an arbitrary image, either text or graphics, as an optical illusion according to the persistence of vision. It has several advantages over many of the available persistence of vision fidget spinners on the market including the well-known device of Microchip.*

1 Uvod

Prstna vrtavka za prikaz besedila ali grafike je rotirajoča naprava, ki deluje po principu vztrajnosti vida (slika 1).



Slika 1: Prstna vrtavka za prikaz besedila ali grafike

Prikaz dosežemo s hitrim premikanjem stolpca svetlečih diod po krožni poti, kjer v vsakem trenutku z njim prikažemo le delček celotne slike ali besedila. Opazovalec vidi celotno sliko, če čas, v katerem so prikazani vsi posamezni deli slike, ne presega časa vztrajnosti vida človeškega očesa [1, 2].

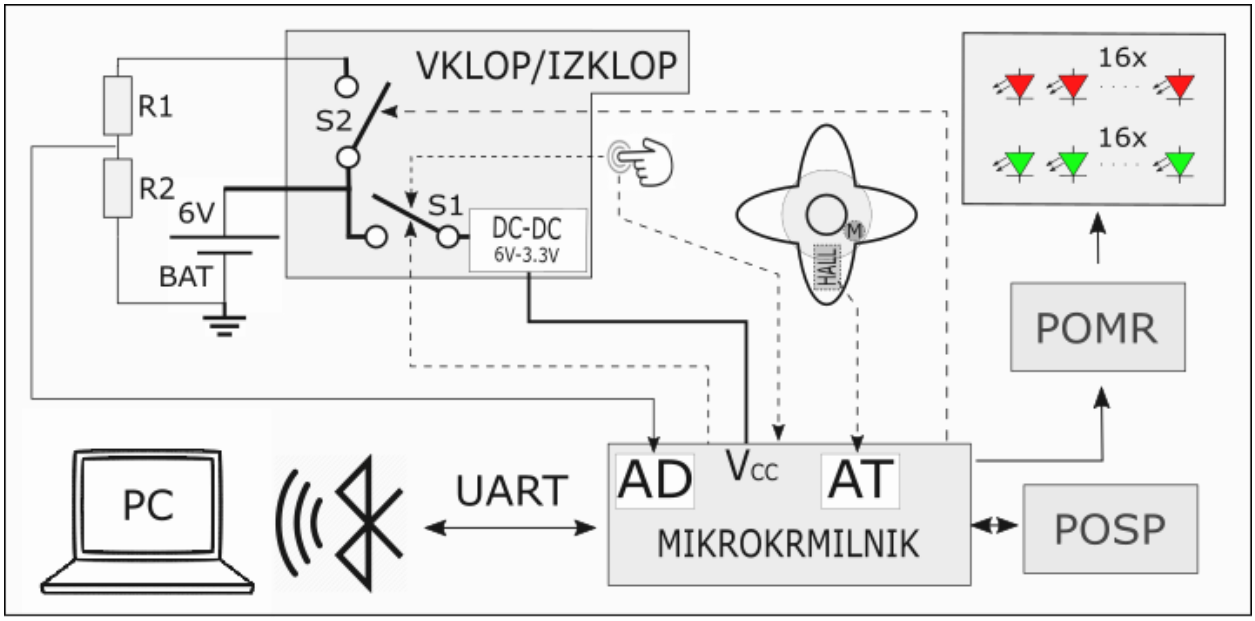
Naša izvedba naprave je nastala v okviru študentskega projekta po vzoru prstne vrtavke podjetja Microchip [3]. Obe izvedbi za natančno pozicioniranje besedila ali grafike med vrtenjem vrtavke, ne glede na hitrost vrtenja ali pospešek, namesto programske rešitve uporabljata namenski strojni modul, imenovan kotni časovnik (Angular Timer – AT) [4]. Vsebino prikaza je pri obeh izvedbah mogoče spreminjati tudi med vrtenjem vrtavke prek povezave Bluetooth Low Energy (BLE).

V primerjavi z vrtavko podjetja Microchip ima naša več prednosti. Z namenom doseganja različnih barvnih učinkov in večje ločljivosti besedila ali grafike ima namesto enega stolpca z osmimi svetlečimi diodami dva stolpca s po šestnajstimi. V enem stolpcu so diode rdeče, v drugem pa zelene. Namesto bipolarnega Hallovega senzorja in dveh magnetov, ki morata biti obratno orientirana, smo uporabili omnipolarni senzor [5] in samo en magnet, katerega orientacija ni pomembna, kar poceni čas in stroške montaže. Vrtavka podjetja Microchip se napaja s tremi zaporedno vezanimi gumbnimi baterijami CR2032, naša pa z zaporedno-vzporedno vezavo štirih takšnih baterij. Na ta način je dosežena večja učinkovitost pretvornika na 3,3 V napajalno napetost in s tem daljša življenjska doba baterij. Ta se poveča tudi zaradi samodejnega izklopa napajalne napetosti, če naprava ni aktivna. Iz neaktivnega stanja jo zbudimo z uporabo namenskega vezja, katerega osrednji del je kapacitivno stikalo. Merjenje kapacitete baterije je izvedeno s pomočjo integriranega stikala v pretvorniku za 3,3 V, ki omogoča dodatno zmanjšanje porabe energije. Vse komponente so skrbno izbrane iz skupine komponent z nizko porabo.

Blokovna shema naprave in opis uporabljenih komponent sta podana v poglavju 2. Zasnova programske opreme, ki se izvaja v sami napravi in tiste na osebnem računalniku, je opisana v poglavju 3. V poglavju 4 prikažemo realizacijo naprave in preizkus njenega delovanja. V sklepu ovrednotimo dosežene rezultate in nakažemo morebitne izboljšave in nadgradnje.

2 Strojna oprema

Blokovna shema strojne opreme naprave je prikazana na sliki 2. Osrednji del predstavlja 8-bitni mikrokrmilnik PIC16F1619 proizvajalca Microchip [6]. Njegova glavna naloga je v ustreznem časovnem sosledju polniti 16-bitna pomikalna registra (POMR), na katera je priključenih 16 zelenih in 16 rdečih svetlečih diod. Pomikalna registra sta namenska tokovna gonilnika svetlečih diod LED [7]. Vsaka polnitev pomikalnih registrov predstavlja en stolpec prikazane vsebine (npr. prvi stolpec črke E na sliki 1). Za natančno pozicioniranje besedila uporabljamo kotni časovnik [4], ki se nahaja v mikrokrmilniku. Kotni časovnik določa pozicijo besedila na osnovi informacije, ki jo prejme od Hallovega senzorja [5]. Hallov senzor se nahaja na rotorju vrtavke (del vrtavke, ki se vrti) in sporoča periodo vrtenja kotnemu časovniku ob preletu mimo magnetka (M) na statorju (negibni del, namenjen držanju vrtavke). Spreminjanje prikazane vsebine na vrtavki je omogočeno prek modula Bluetooth Low Energy RN4871 [8] prizvajalca Microchip in kapacitivne tipke. Ker uporabljene elektronske komponente za svoje delovanje potrebujejo napajalno napetost 3,3 V, v vezju pa uporabljamo baterije CR2032, ki v uporabljeni vezavi generirajo napetost 6V, uporabimo za znižanje napetosti DC/DC pretvornik TPS62745 [9] proizvajalca Texas Instruments. Da bi zmanjšali porabo energije, smo razvili namensko vezje, ki baterijsko napetost po potrebi priklopi oziroma odklopi (VKLOP/IZKLOP). Vezje vključimo z dotikom kapacitivne tipke, priključene na

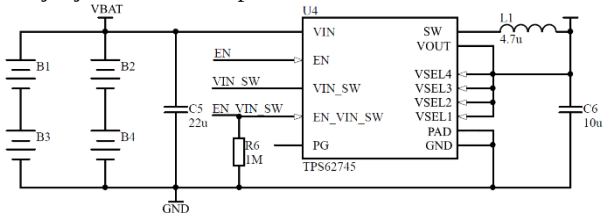


Slika 2: Blokovna shema strojne zasnove naprave

kapacitivno stikalo PCF8883 [10], ki je del vklopno-izklopnega vezja. Po preteku petih minut brez aktivnosti mikrokrmilnik napajalno napetost izklopi (S1). S pomočjo analogno-digitalnega pretvornika (AD) meri kapaciteto baterij na osnovi padca napetosti na napetostnem delilniku (pred tem vključi stikalo S2). Sledi podobnejši opis delovanja posameznih delov vezja.

2.1 Napajalni del naprave

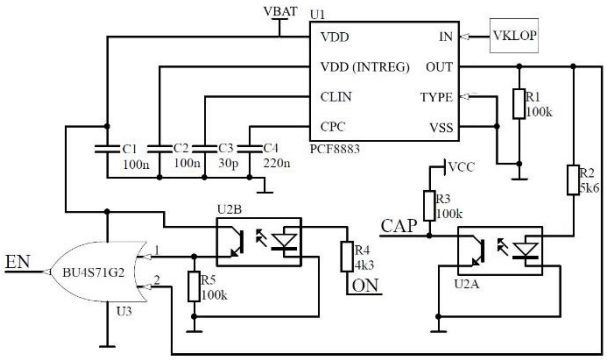
Slika 3 prikazuje napajalni del naprave. V vezju so nameščene štiri 3-voltne baterije CR2032. Po dve bateriji sta vezani zaporedno, s čimer dosežemo napetost 6V, ki jo peljemo na vhod pretvornika TPS62745 za znižanje enosmerne napetosti. Želeno izhodno napetost pretvornika 3,3 V smo izbrali z ustrezno nastavitvijo vhodov VSEL1 do VSEL 4. Pretvornik ima zelo majhno porabo (400 nA). V načinu izklopa (vhod EN je na nizkem nivoju) se poraba zniža na 130 nA. Pretvornik uporablja integrirano notranje napetostno stikalo, ki preklopi vhodno napetost VIN na izhodni priključek VIN_SW, ko je vhodni priključek EN_VIN_SW na visokem logičnem nivoju [9]. To lastnost smo s pridom izkoristili za merjenje napetosti baterije, ki jo s pretvornikom AD merimo kot padec napetosti na napetostnem delilniku, določenim z uporoma R1 in R2 (slika 2). Tako teče tok čez napetostni delilnik le v času merjenja električne napetosti.



Slika 3: Napajalni del naprave

2.2 Namensko vezje za znižanje porabe

Že sama izbira komponent zagotavlja nizko porabo energije, ki smo jo še dodatno znižali z izdelavo namenskega vezja za vklop in izklop naprave (slika 4). Ob dotiku kapacitivne tipke, priključene na kapacitivno stikalo PCF8883 [10], aktiviramo izhod vezja EN, ki omogoči pretvornik za znižanje enosmerne napetosti TPS62745 (slika 3). Ko pritisnemo na kapacitivno tipko (vhod CAP na sliki 4), se prek vhoda 2 vrat U3 aktivira izhod EN in naprava »oživi«. Zdaj se aktivira signal ON, ki prek vhoda 1 vrat U3 zagotavlja napajanje tudi po spustitvi tipke. Mikrokrmilnik po preteku petih minut neaktivnosti onemogoči pretvornik (poraba 130 nA) in s tem prekine dobavo električne napetosti 3,3 V.



Slika 4: Vezje za vključitev in izključitev naprave

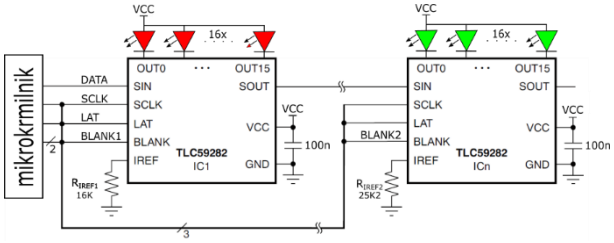
2.3 Gonilnik svetlečih diod

Za poganjanje svetlečih diod sta uporabljena 16-bitna pomikalna registra TLC59282 (slika 5), ki sta povezana v kaskado (podatkovni izhod SOUT prvega vodi na podatkovni vhod SIN drugega). Izhodni del pomikalnega registra (OUT0 do OUT15) je 16-kanalni gonilnik svetlečih diod z zmožnostjo ponikanja konstantne vrednosti toka, nastavljene z uporom na priključku IREF.

Vrednost upora R_{IREF} v kΩ izračunamo iz izraza (1), kjer je I_{LED} želeni tok skozi svetlečo diodo v mA in V_{REF} referenčna napetost 1,205 V [7].

$$I_{LED} = \frac{V_{REF}}{R_{IREF}} \cdot 41{,}9 \qquad (1)$$

Uporabljene svetleče diode so serije APTD in KPTD proizvajalca Kingbright. Ker je pri nominalnem toku 2 mA svetilnost uporabljenih rdečih svetlečih diod 150 mcd, svetilnosti zelenih pa 440 mcd, smo s primerno izbranima uporoma (R_{IREF1} = 16 kΩ in R_{IREF2} = 25,2 kΩ) dosegli tok 3,15 mA skozi rdeče in tok 2 mA skozi zelene diode in s tem približno enako svetilnost enih in drugih.



Slika 5: Priključitev svetlečih diod

2.4 Kotni časovnik

Kotni časovnik [4] je pomemben del naše naprave, ker zagotavlja natančno lociranje besedila ali grafike in razbremeni procesor. Ob vsakem polnem obratu vrtavke Hallov senzor generira impulz, ki gre na vhod kotnega časovnika. Čas med dvema zaporednima impulzoma tako ustreza kotu zasuka 360º. Strojna oprema kotnega časovnika razdeli periodični vhodni signal na 180 enakih intervalov in tako sproži prekinitev na mikrokrmilniku na vsaki 2º zasuka vrtavke. V prekinitvenem programu se za vsak položaj vrtavke, ki zahteva spremembo vsebine prikaza glede na prejšnji položaj, pošlje ustrezna bitna kombinacija na stolpca LED.

2.5 Modul Bluetooth Low Energy

Pri odločitvi za izbiro komunikacijskega modula Bluetooth Low Energy proizvajalca Microchip so nas prepričale njegove majhne dimenzije in predvsem nizka poraba energije (2,9 µA v mirovanju). Modul ima vgrajeno keramično anteno in komunicira z mikrokrmilnikom prek vrat UART [8].

Komunikacija med modulom in PC-jem poteka prek Bluetooth Low Energy protokola GATT (Generic Attribute Profile), ki določa organizacijo in izmenjavo podatkov med različnimi aplikacijami [11]. Osnovan je na protokolu ATT (Attribute Protocol), ki je njegov transportni protokol. GATT je namenjen odkrivanju storitev (Services) ter branju in spreminjanju karakteristik (Characteristics), povezanih s storitvijo. Karakteristika predstavlja vrednost podatka, ki ga prenašamo med odjemalcem (PC) in strežnikom (vrtavka). Storitve se razlikujejo po 16 ali 128-bitnem univerzalnem unikatnem identifikatorju UUID in so s pripadajočimi karakteristikami zbrane v navidezni zbirki, imenovani profili (Profiles). Vsaka karakteristika ima tudi svoj UUID oziroma oprimek (handle), ki omogoča njeno identifikacijo.

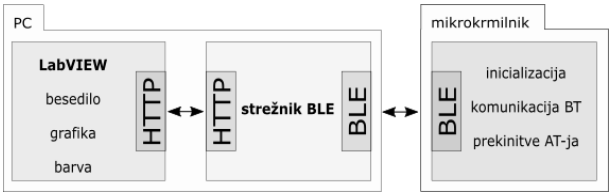
V tem projektu je uporabljen en servis s tremi karakteristikami (tabela 1). Prva karakteristika v tabeli je namenjena upravljanju 32-ih svetlečih diod v času mirovanja naprave, druga prikazu besedila med rotacijo vrtavke in zadnja izbiri barve besedila.

Tabela 1. Tri karakterstike znotraj storitve

Storitve UUID			
DABEB8B482AF42A8B35716F80B374EE2			
Oprimek	Namen	Velikost	Dovoljenja
0092	LED v mirovanju	0x14	W/R
0095	Rotirajoči LED	0x14	W/R
0098	Barva – rotirajoči LED	0x02	W/R

3 Programska oprema

Slika 6 prikazuje strukturo programske kode, ki se izvaja v mikrokrmilniku in na osebnem računalniku PC. Programska koda, ki se izvaja na mikrokrmilniku, je napisana v programskem jeziku C in izvede najprej inicializacijo uporabljene periferije (PLL sistemska ura 32 MHz, vhodno-izhodni priključki, kotni časovnik, UART, zunanje prekinitve). Po uspešni inicializaciji mikrokrmilnik prek povezave Bluetooth LE prejema od PC-ja zahteve o vsebini na vrtavki in čaka na prekinitve kotnega časovnika za prikaz grafike.



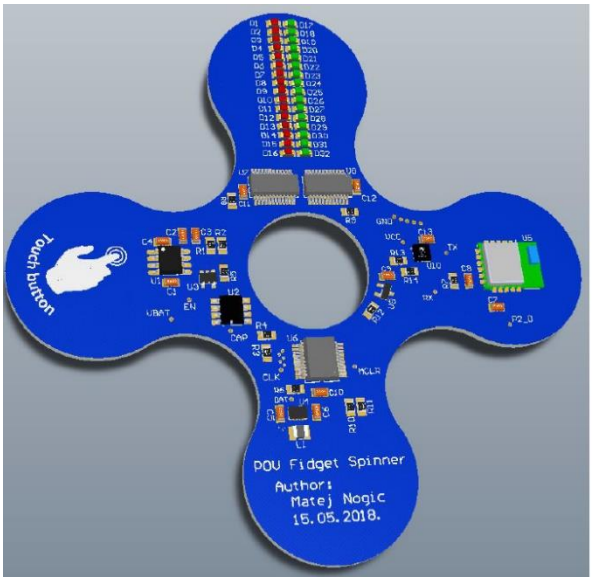
Slika 6: Struktura programske opreme

Aplikacija, ki se izvaja na PC-ju, je napisana v okolju NI LabVIEW podjetja National Intruments. Prek nje pošiljamo zahteve glede barve in vsebine prikazane grafike do mikrokrmilnika. Ker aplikacije, zapisane v okolju NI LabVIEW, trenutno ne znajo neposredno komunicirati prek protokola Bluetooth Low Energy, smo uporabili programski strežnik HTTP NI BLE podjetja National Instruments, ki ga namestimo na PC-ju [12]. Strežnik ustvari most med aplikacijo NI LabVIEW in vmesnikom Bluetooth Low Energy na PC-ju, tako da aplikacija LabVIEW komunicira prek protokola HTTP na vratih 5000 s strežnikom, ta pa z mikrokrmilnikom prek protokola Bluetooth Low Energy.

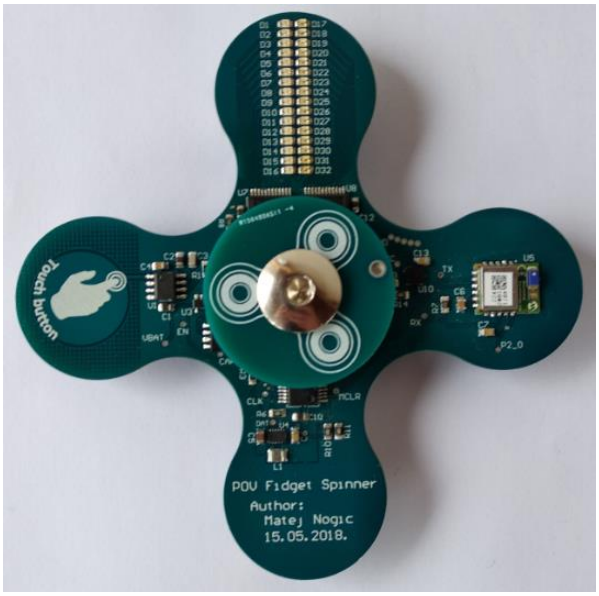
Za generiranje pisave smo uporabili brezplačno aplikacijo »The Dot Factory« in zapisali pomožno aplikacijo v LabVIEW za modificiranje generiranih pik za znake. Pomožna aplikacija dve 8-bitni vrednosti, ki predstavljata pike posameznega stolpca (diode LED), združi v 16-bitno vrednost in izvede ustrezne pomike bitov, s čimer kompenzira neskladnost fizičnega zaporedja svetlečih diod, do katerega je prišlo zaradi omejitev pri povezovanju linij na dvostranski tiskanini z zaporedjem bitov izvorne 16-bitne generirane vrednosti stolpca.

4 Izdelava in preizkus naprave

Za napravo smo načrtali dvoslojno tiskanino v programskem okolju Altium Designer. Tridimenzionalni izgled naprave v omenjenem programu je prikazan na sliki 7, fotografija izdelane vrtavke z vstavljenim statorjem s krogličnim ležajem za držanje na prstu pa na sliki 8.



Slika 7: 3D prikaz vrtavke v programu Altium Designer



Slika 8: Izdelana vrtavka

5 Sklep

Cilj projekta je bil izdelati prstno vrtavko za prikazovanje besedila ali druge grafike po principu vztrajnosti vida. Testiranje delovanja naprave je potrdilo uspešnost realizacije zastavljenega cilja. Grafika je prikazana jasno in na natančno definirani lokaciji. S tem upravičimo uporabo kotnega časovnika, ki procesor razbremeni sprotnega računanja kota zasuka vrtavke in mu daje več časa za osveževanje pik med rotacijo vrtavke. K temu pripomore tudi izbira svetlečih diod z visoko svetilnostjo pri majhnem toku.

Izbira komponent z nizko porabo in premišljeno načrtano vezje za znižanje porabe zagotavljata dolgo življenjsko dobo baterij.

Z uporabo pospeškometra (POSP - slika 1), ki je že del vrtavke, bi lahko dosegli, da bi bilo prikazano besedilo vedno pravilno obrnjeno, ne glede na to, kako držimo vrtavko.

Napravo lahko uporabljamo za pedagoške ali promocijske namene, kot igračo ali celo kot terapevtski pripomoček [13, 14].

Literatura

[1] G. J. Burak in D. P. Fiden, „Gaming Machine Having a Persistence-of-Vision Display“. Patent US 7,708,640 B2, 4 May 2010.

[2] D. Z. Ratnayake, C. K. Kulasekera, Y. V. Fernando, T. W. Meewella, J. P. Praveena in S. P. Yasakethu, „Persistence Vision Display,“ v *International Research Symposium on Engineering Advancements*, Malabe, Sri Lanka, 2015.

[3] Microchip Technology Inc., „Bluetooth Fidget Spinner,“ 21 February 2018. [Elektronski]. Dosegljivo na: http://www.microchip.com/promo/bluetooth-fidget-spinner.

[4] D. Hou, „Angular Timer Implementation, Microchip Technology Inc.,“ 2016. [Elektronski]. Dosegljivo na: http://ww1.microchip.com/downloads/en/AppNotes/90003143A.pdf.

[5] Texas Instruments, „DRV5033 Digital-Omnipolar-Switch Hall Effect Sensor,“ September 2016. [Elektronski]. Dosegljivo na: http://www.ti.com/lit/ds/symlink/drv5033.pdf.

[6] Microchip, „14/20-Pin, 8-Bit Flash Microcontroller,“ 2014-2017. [Elektronski]. Dosegljivo na: http://ww1.microchip.com/downloads/en/DeviceDoc/40001770D.pdf.

[7] Texas Instruments, „16-Channel, Constant-Current LED Driver with 4-Channel Grouped Delay,“ October October. [Elektronski]. Dosegljivo na: http://www.ti.com/lit/ds/symlink/tlc59282.pdf.

[8] Microchip Technology Inc., „RN4870/71 Bluetooth Low Energy Module User's Guide,“ 2017. [Elektronski]. Dosegljivo na: http://ww1.microchip.com/downloads/en/DeviceDoc/50002466B.pdf.

[9] Texas Instruments, „TPS62745 Dual-cell Ultra Low IQ Step Down Converter for Low Power Wireless Applications,“ June 2015. [Elektronski]. Dosegljivo na: http://www.ti.com/lit/ds/symlink/tps62745.pdf.

[10] NXP Semiconductors, „Capacitive touch/proximity switch with auto-calibration, large voltage operating range, and very low power consumption,“ 14 September 2016. [Elektronski]. https://www.nxp.com/docs/en/data-sheet/PCF8883.pdf.

[11] „GATT Overview,“ Bluetooth SIG, Inc., 2018. [Elektronski]. Dosegljivo na: https://www.bluetooth.com/specifications/gatt/generic-attributes-overview.

[12] National Instruments, „LabVIEW Bluetooth Low Energy (BLE) Instrument Drivers for Windows & cRIO,“ 2017. [Elektronski]. Dosegljivo na: https://forums.ni.com/t5/NI-Labs-Toolkits/LabVIEW-Bluetooth-Low-Energy-BLE-Instrument-Drivers-for-Windows/ta-p/3538595.

[13] E. J. Cohen, R. Bravi in D. Minciacchi, „The effect of fidget spinners on fine motor control,“ *Scientific Reports*, Izv. 8:3144, February 2018.

[14] L. Mesquita, G. Brockington, P. A. de Almeida, M. E. Truyol, L. A. Testoni in P. F. F. Sousa, „Using a fidget spinner to teach physics,“ *Physics Education*, Izv. 53, št. 4, 2018.

High-level synthesis and optimization of video signal processing IP

Tadej Skuber, Andrej Trost, Andrej Žemva

Faculty of Electrical Engineering, University of Ljubljana, Slovenia
E-mail: ts7305@student.uni-lj.si

Abstract. *Experiment described in the paper studies design methodology by implementing image rotation with bilinear interpolation. Algorithm is firstly coded and tested in high level language (C). Code is than converted to RTL design using special commands (directives). To reach desired result latency and small circuit size, code style and directives are experimented with. Effects of different fixed-point decimal places is also analyzed.*

1 Introduction

Advanced designs used in today's increasingly complex electronic products are stretching the boundaries of density and performance. They create a challenge for design teams to hit a target release window within their allocated budget. A productive methodology for addressing these design challenges is one where more time is spent at higher levels of abstraction, where verification time is faster and productivity gain is the greatest [1].

The main goal of this paper is to use High Level Synthesis (HLS) to find algorithm with optimal performance, from the choice of multiple versions of known algorithm for image rotation, evaluating performance and resource use. The usage of HLS enables us a fast way of checking both criteria without much coding effort.

Image rotation and translation are widely used operations in image processing with possible parallelism potential. The only problem is a need for additional resampling algorithm.

In computer vision and image processing, bilinear interpolation is one of the basic resampling techniques. It can be used where perfect image transformation with pixel matching is impossible, so that one can calculate and assign appropriate intensity values to pixels. Unlike other interpolation techniques, such as nearest-neighbor interpolation and bicubic interpolation, bilinear interpolation uses values of only the 4 nearest pixels, located in diagonal directions from a given pixel, in order to find the appropriate color intensity values of chosen pixel [2].

2 Rotation and bilinear interpolation

Used algorithm iterates over resulting image in x' and y' direction. For every pixel (x', y') it looks where does it mirror in original image (x, y) (Fig. 1). Algorithm uses the following formula:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \left(\begin{bmatrix} x' \\ y' \end{bmatrix} - T \right) A + \begin{bmatrix} II_H/2 \\ II_W/2 \end{bmatrix} \tag{1}$$

where

$$A = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{2}$$

$$T = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} OI_H/2 - t_{1d} \\ OI_W/2 - t_{2d} \end{bmatrix} \tag{3}$$

Matrix T translates original image for t_{1d} pixels in x direction and for t_{2d} pixels in y direction. Rotation is handled by matrix A, which rotates image for for angle θ in anti-clock wise direction. II_H and II_W are constants that represent original image (Input Image or II) width and height. Constants OI_H and OI_W represent width and height of Output Image (OI).

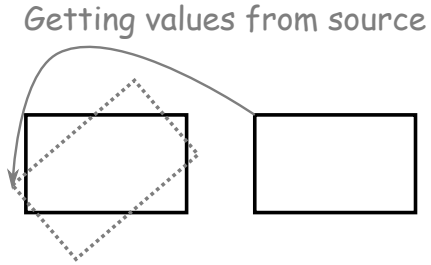


Figure 1: Representation of algorithm operation [3].

Image transformation does not result in integer coordinates values of the pixels. Therefore bilinear interpolation is used to find appropriate pixel values at given coordinates. Pixel color value $f(x, y)$ is calculated using the four neighbor pixels color values (Fig. 2) by the following formula:

$$f(x,y) = \begin{bmatrix} 1-x & x \end{bmatrix} \begin{bmatrix} f(0,0) & f(0,1) \\ f(1,0) & f(1,1) \end{bmatrix} \begin{bmatrix} 1-y \\ y \end{bmatrix} \tag{4}$$

Values $f(0,0), f(1,0), f(0,1), f(1,1)$ are neighbor pixel color values. Variables x and y are distances of calculated (x, y) from the nearest integer coordinate value.

Formula is applied to each of three RGB (Red, Green and Blue) pixel colors, giving three $f(x, y)$ result values that are combined in one piece of data.

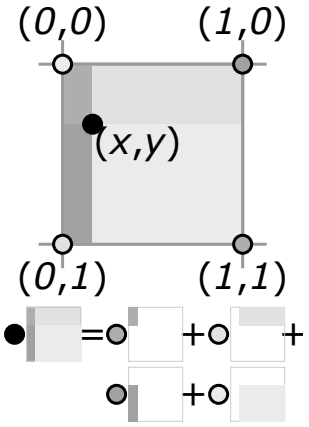


Figure 2: Geometric visualization of bilinear interpolation [2].

2.1 C code

The following code (Lst. 1) was firstly used to test the algorithm and asses performance of HLS (High Level Synthesis).

```
1  typedef int                cnt_t;
2  typedef ap_fixed<32, 16> fix_t;
3  typedef ap_uint<24>        col_t;
4  ...
5  void bilinear(
6  col_t input_image[II_H*II_W],
7  col_t output_image[OI_H*OI_W],
8  fix_t A[4],
9  fix_t T[2]){
10 ...
11 loopi: for (cnt_t i = 0; i < OI_H; i++){
12 loopj: for (cnt_t j = 0; j < OI_W; j++){
13
14 x = ((fix_t)(i) - OI_H_2 - T[1])*(A[0]) +
15 ((fix_t)(j) - OI_W_2 - T[0])*(A[1]) + II_H_2;
16 y = ((fix_t)(i) - OI_H_2 - T[1])*(A[2]) +
17 ((fix_t)(j) - OI_W_2 - T[0])*(A[3]) + II_W_2;
18 ...
19 if ((x0 >= 0 && y0 >= 0) &&
20 (x1 < II_H && y1 < II_W)){
21 dx = x - (fix_t)(x0);
22 dy = y - (fix_t)(y0);
23 dx_dy = dx*dy;
24
25 loopk: for (cnt_t k = 0; k < 3; k++){
26 shift = k << 3;
27
28 output_image[i*OI_W + j] |= ((col_t)(
29 (fix_t)(((input_image[x0 * II_W + y0]) &
30 (col_t)(0xFF << (shift))) >> (shift))*
31 (1+dx_dy-dx-dy)+
32 (fix_t)(((input_image[x1 * II_W + y0]) &
33 (col_t)(0xFF << (shift))) >> (shift))*(dx-dx_dy) +
34 (fix_t)(((input_image[x0 * II_W + y1]) &
35 (col_t)(0xFF << (shift))) >> (shift))*(dy-dx_dy) +
36 (fix_t)(((input_image[x1 * II_W + y1]) &
37 (col_t)(0xFF << (shift))) >> (shift))*dx_dy))<<
38 (shift);
39 ...
```

Listing 1: Basic unmodified C code.

Main algorithm input is 1D array input_image of size II_H*II_W, with each 24 bit element. This array contains 2D image that is transformed into single row. Other inputs are matrix A and T, containing information of image rotation and translation. Output is rotated and translated image output_image with size OI_H*OI_W, each element also being 24 bits in size.

The RGB color values are coded in 24 bit ap_uint type variables with identifier type col_t, giving each color a standard 8 bit value. Variables with identifier fix_t are of type ap_fixed and are used for calculations where decimal places are needed.

Basic algorithm implementation loops through resulting image with two for loops, each pixel assigned (with coordinates (i, j)) a proper pixel value form original image (with coordinates (x, y)). Pixel color calculation is done with bilinear interpolation according to equation (4).

3 Directives

Using Vivado HLS compiler directives, we can assist it in mapping C/C++ code to hardware [6]. In design four directives were used, unroll, pipeline, interface and array reshape. Unroll was used in for loop "loopk", for calculation of the individual RGB color value. Loops "loopi" and "loopj" are too large for unrolling, so directive was not used here. Directive pipeline was used on all loops, reducing the initiation interval for loops by allowing the concurrent execution of operations [5]. Directive interface with argument ap_memory was used on function array parameters, for implementation of arguments as standard RAM interface. Images were then conveniently stored in RAM. Array reshape was used for parallel access to data in RAM.

Algorithm synthesis settings were set for ZedBoard using Vivado HLS 2017.4. Synthesis clock period was set to default 10 ns, and implementation was done in C and supports various input (II_H being height and II_W width) and output (OI_H and OI_W) image sizes. For implementation, standard test image size of $512*512$ was used.

Results of using different directives are visible in the following table:

DIRECTIVE	LATENCY/INTERVAL
/	8127489
only pipeline and unroll	524299
all directives	524299

Table 1: Latency when using different directives.

It can be seen(Tab. 1) that it is possible to achieve latency of about two clock cycles per pixel ($512*512*2 = 524288$), but it can also be noticed that additional directives (interface memory and array reshape) do not help to decrease latency. Problem is that only one RAM array element can be accessed in one clock cycle. Solution would be in loading multiple pixel data in one read cycle, and since input data repeats when iterating true image pixels. This can be used to lower latency to one clock cycle per pixel (Lst. 2).

4 Code modification

Code modification was done so that it does consider repetition of data input [11]. That was done with 48 bit data input instead of 24. Image data was so divided in one array of size II_W/2*II_H. Each element of array therefore includes 2 pixel data. This small change in code gives us desired clock cycles per pixel. For image size $512 * 512$ giving us latency of 262155 ($262155/(512 * 512) \approx 1$).

But this method is only applicable for images of size $2^n * 2^m$. For all other sizes latency is the same. Algorithm with different image size inputs has latency of at least $II_H * II_W * 2$. The test with image size $650 * 975$ has latency of 2535038 ($2535038/(650*975) \approx 4$), or image of size $640 * 540$ has 691235 ($2535038/(640 * 540) \approx 2$) clock cycles of latency. This modification works only when directive array reshape is applied.

	basic	**modified**
DSP48E	64/29%	64/29%
FF	2620/2%	3135/2%
LUT	7690/14%	7295/13%

Table 2: Resources used by different algorithms. Both using 32 bit fixed point numbers with 16 bits for calculations.

```
...
if((y0 & 0x1) == (cnt_t)(0)){
temp_image[0] =
(color_t)(input_image[x0 * II_W/2 + (y0 >> 1)]);
temp_image[1] =
(color_t)(input_image[x0 * II_W/2 + (y0 >> 1)] >>
    24);
temp_image[2] =
(color_t)(input_image[x1 * II_W/2 + (y0 >> 1)]);
temp_image[3] =
(color_t)(input_image[x1 * II_W/2 + (y0 >> 1)] >>
    24);
}else{
temp_image[0] =
(color_t)(input_image[x0 * II_W/2 + (y0 >> 1)] >>
    24);
temp_image[1] =
(color_t)(input_image[x0 * II_W/2 + (y1 >> 1)]);
temp_image[2] =
(color_t)(input_image[x1 * II_W/2 + (y0 >> 1)] >>
    24);
temp_image[3] =
(color_t)(input_image[x1 * II_W/2 + (y1 >> 1)]);
}
...
loopk: for (cnt_t k = 0; k < 3; k++) {
shift = k << 3;
temp_out[k] = ((color_t)(
(fixed_t)(((temp_image[0]) & (color_t)(0xFF << (
    shift)))
>> (shift))*(1+dx_dy-dx-dy) +
(fixed_t)(((temp_image[2]) & (color_t)(0xFF << (
    shift)))
>> (shift))*(dx-dx_dy) +
(fixed_t)(((temp_image[1]) & (color_t)(0xFF << (
    shift)))
>> (shift))*(dy-dx_dy) +
(fixed_t)(((temp_image[3]) & (color_t)(0xFF << (
    shift)))
>> (shift))*dx_dy))<<(shift);
...
```

Listing 2: Modified C code.

Use of resources is visible in table 2. It can be seen that there are no major differences in use of resources between basic and modified algorithm for image of size $512 * 512$.

5 Effects of fixed point

Effects of using fixed point, instead of floating point, were also studied. Different number of fixed point decimal bits were used to see the effects on image quality. For assessment of image quality fixed point images were compared to floating point image. Code was only slightly changed (Lst. 3).



Figure 3: Fixed point images with 1, 2, 3 and 16 decimal bits of accuracy.

$$MSE = \frac{1}{m * n} \sum_{i=0}^{n} \sum_{j=0}^{m} [I(i,j) - K(i,j)]^2 \quad (5)$$

Results of comparison are given in Table 4. Image comparison was done with two algorithms, Mean Squared Error (MSE) and Structural Similarity Measure (SSIM) [6]. MSE was simply implemented according equation 5. More complicated image assessment with SSIM [7] was done using existing python library [9]. For simplifying image quality estimation, gray-scale image was used.

```
typedef int                     cnt_t;
//typedef ap_fixed<32, 16>      fix_t;
typedef float                   fix_t;
typedef ap_uint<24>             col_t;
...
```

Listing 3: Changing fixed point to floating point to compare images.

NDB	**DSP48E**	**FF**	**LUT**
8	17	2423	2423
9	17	2467	5982
10	17	2696	6452
11	17	2724	6826
12	17	2778	7014
13	17	2836	7188
14	23	2894	7324
15	43	2820	7351
16	51	3092	7334
17	51	2916	7332
18	51	2964	7311

Table 3: Resources used when changing number of fixed point bits. Using 10 bits for integer part and NDB for decimal part.

As can be seen in Fig. 3 and Tab. 4, image quality quickly becomes sufficient with the use of more than 8

decimal bits. Image distortion is visible only if a few decimal bits are being used. Assessment methods used enable insight of image quality, revealing that use of more bits does not significant add to image quality. But increasing the number of fixed point bits has an effect on resource use (Tab. 3), mainly on DSP48E blocks (FFs and LUTs are not greatly effected).

NDB	MSE	SSIM
1	4782	0.32853
2	4556	0.32055
3	2202	0.47513
4	951	0.60306
5	590	0.68887
6	230	0.81405
7	98	0.90796
8	43	0.96283
9	9.77	0.98872
10	1.56	0.99701
11	0.41	0.99887
12	0.13	0.99948
13	0.06	0.99967
14	0.029	0.99988
15	0.014	0.99993
16	0.011	0.99994
17	0.005	0.99997
18	0.002	0.99999

Table 4: Results with different Number of Decimal Bits(NDB). Using 10 bits for integer part.

6 CONCLUSIONS

As shown in this paper, HLS is useful at development of digital hardware design. As demonstrated in this experiment, algorithm that would take bigger effort to code in HDL (Hardware Description Language) language such as VHDL, can easily be implemented. Doing so with use of only a few lines of simple C code and some directives.

Important thing to consider when using HLS tool is that C code synthesis output does not always meet our idea of hardware. Change of hardware design can be made by exploiting directives. Use of directives like unroll or pipeline can for instance increase algorithm parallelism and shorten latency. More drastic and optimized change in hardware design usually requires slightly different coding style. For instance, Vivado HLS only managed to reduce latency after image array elements were grouped together.

One disadvantage of HLS is definitely unreadable HDL language code at the end of synthesis. Regardless of representation, the best algorithms for hardware realization are not necessarily the same as those used in software. While HLS tools will produce hardware for realizing the algorithm, there may be better or more appropriate algorithms [10].

References

[1] UltraFast High-Level Productivity Design Methodology Guide, Xilinx, UG1197 (v2018.2) June 6, 2018.

[2] Bilinear interpolation, Wikipedia, https://en.wikipedia.org/wiki/Bilinear_interpolation

[3] Image Rotation With Bilinear Interpolation, http://polymathprogrammer.com/2008/10/06/image-rotation-with-bilinear-interpolation/

[4] SDAccel Development Environment Help, HLS Pragmas, https://www.xilinx.com/html_docs/xilinx2018_1/sdaccel_doc/ijd1517252127848.html/

[5] Vivado Design Suite User Guide, High-Level Synthesis, Xilinx, UG902 (v2014.1) May 30, 2014.

[6] Xilinx Vivado HLS Beginners Tutorial, https://medium.com/@chathura.abeyrathne.lk/xilinx-vivado-hls-beginners-tutorial-custom-ip-core-design-for-fpga-59876d5a4119

[7] How-To: Python Compare Two Images, https://www.pyimagesearch.com/2014/09/15/python-compare-two-images/

[8] Zhou Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Transactions on Image Processing, Volume: 13, Issue: 4, April 2004.

[9] Image processing in Python, http://scikit-image.org/

[10] The Advantages and Limitations of High Level Synthesis for FPGA Based Image Processing.

[11] Ryan Kastner, Janarbek Matai, and Stephen Neuendorffer, Parallel Programming for FPGAs, May 11, 2018.

Standardizacija in umestitev ekološko osveščenega načrtovanja v elektrotehniki

Roman Strahija[1], **Tatjana Strahíja**

[1]*STMicroelectronics d. o. o.*
E-pošta: roman.benkovic@st.com

Standardization and Inclusion of Electrotechnics Eco-Design

Consumer electronics (CE) market expansion over last decades directs changes in assessment of our design, production and recycling. Since unchanged approach in production and handling of waist electronic industry will not sustain environmental resistance, an increasing demand to change our practices is required.

There are three main "life" phases of electronic equipment – the production stage, in which the product is developed, life time of the equipment and handling of it when it is no longer useful in its original form.

In order to reduce the impact of the electronic industry on the environment, it is necessary to strive to improvement at all stages of the equipment's life cycle; eg. in the production cycle, it is necessary to reduce the use of natural resources, increase recycling of by-products, alter production process by replacing harmful processing steps with less harmful ones; but further more, it is necessary to design the product in such way that it has the longest possible life and the highest possible value for consumer; and for the final step, it is necessary to plan for its, if possible, re-use or ability to recycle with the least possible impact on the environment.

The guidelines for environmental-friendly design are defined in several directives and standards that most companies are trying to apply; some due to awareness and responsibility towards the environment, other simply of legal basis requirements.

The article will give an overview of the eco-design field, its directives, standardization and guidelines for eco-friendly design.

Keywords: Eco Design, Environment, Sustainability, Life Cycle Assessment (LCA), Eco Design Directives (EDD), Standardization

1 Uvod

Ekološko osveščeno načrtovanje – *Eco-Design* (pogosto poimenovano tudi z drugimi izrazi: oblikovanje za okolje / *Design for Environment*, trajnostno oblikovanje – *Sustainable Design*, okolju prijazna zasnova – *Environmental Conscious Design*, čisti design – *Clean Design*, inženiring življenjskega cikla – *Life Cycle Engineering*) je definirano kot sistematično vključevanje okoljskih, zdravstvenih in varnostnih usmeritev v celoten življenjski cikel proizvodnega procesa in izdelka, v smislu ekologije, in ekonomije. Ob tem stremimo k trem glavnim ciljem [1]:

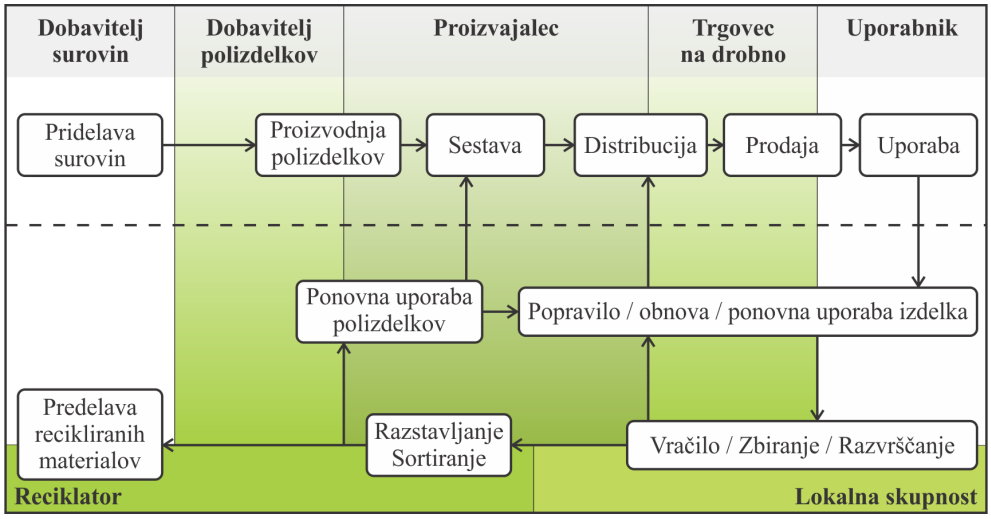
- minimizaciji uporabe obnovljivih in neobnovljivih virov,
- učinkovitem upravljanju obnovljivih virov,
- zmanjšanju in nadzoru nad toksičnimi vplivi na okolje

z namenom, da bi procese, izdelke ali storitve oblikovali tako, da se izognemo negativnim vplivom na ljudi, gospodarstvo ali okolje.

Poglavitne iniciative ekološkega načrtovanja (po EN ISO 14006; 2011 [2]) so:

- sprejemanje novih poslovnih modelov, povečanje znanja in konkurenčnosti ter zmanjšanje stroškov v podjetjih;
- z upoštevanjem regulativ – višja okoljska uspešnost in dostopnost do širšega tržišča;
- spodbujanje inovativnosti zaposlenih pri oblikovanju izdelkov ter v procesih izdelave in povečanje ugleda podjetja ter motivacije zaposlenih;
- izboljšanje, optimizacija ter sprememba zasnove in funkcionalnosti procesov ter izdelkov;
- zmanjšanje negativnih vplivov na okolje;
- celostna obravnava izdelkov čez celotno življenjsko obdobje;
- olajšanje recikliranja in ponovne uporabe izdelka.

Skozi pregled razlogov za vključevanje okoljskih vidikov v oblikovanje in razvoj izdelkov bomo predstavili še standardizacijo in umestitev področja ter načine vključevanja ekološko osveščenega načrtovanja na elektrotehničnem področju.



Slika 1: Soodvisnost korakov ekološko osveščenih procesov v podjetjih [3]

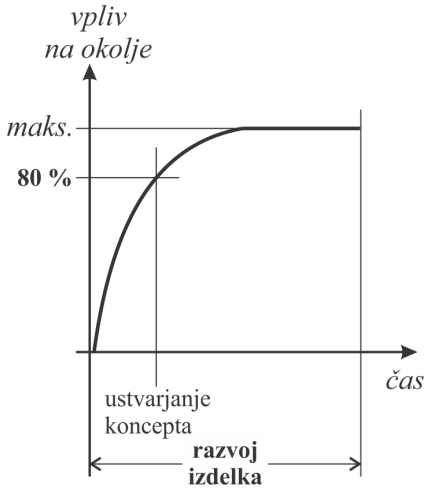
2 Pregled in razlogi za vključevanje okoljskih vidikov v oblikovanje in razvoj izdelkov

V zadnji tretjini stoletja se je, ob ekonomskem gonilu in tržnem razcvetu, povečala količina in dosegljivost raznovrstnih izdelkov na tržišču, kar je spremenilo miselnost potrošnikov in močno skrajšalo pričakovano uporabno dobo izdelkov. Tovrstno ravnanje ima posledično negativen vpliv na okolje, zaradi česar nastaja zahteva po vpeljavi sprememb v miselnost, načrtovanje in procese izdelave, po drugi strani pa po ustreznejšem ravnanju z odsluženimi, uporabniku nezanimivimi in nepotrebnimi izdelki.

Vzporedno z vse večjo potrošnjo so se, ob naraščajoči ekološki osveščenosti, razvile tudi različne direktive in standardi, ki naj bi, na eni strani, usmerjale proizvajalce v skupen boj proti odvečni porabi naravnih virov in zmanjšanem onesnaževanju, na drugi pa v razvoj in dopolnitve zakonodaje, ki zahteva trajnostni razvoj in boljše ravnanje v celotnem obstoju polizdelkov, izdelkov ali njihovih delov.

Začetne spremembe v podjetjih so se začele uvajati predvsem tam, kjer je bilo na voljo dovolj razvojno-raziskovalno pokritih finančnih virov (npr. Philips, IBM, Apple). Manjša podjetja so uvedbo smernic ekološkega načrtovanja vpeljala predvsem v namen zmanjšanja stroškov prispevkov za obremenitev okolja in porabe virov, ne da bi vpeljavo upoštevala že v samem procesu izdelave.

Načrtovanje določa večino (približno 80 % [4]) končnih vplivov na okolje in celoten potek življenjskega cikla izdelka. Še tako napredna kasnejša proizvodja ali končno recikliranje omogočata le manjše posege na vpeljavo sprememb, povečanje učinkovitosti proizvodnih procesov ter zmanjšanje okoljskih stroškov in vplivov. Zato je bistvenega pomena, da se ocena okoljske učinkovitosti vključi v zgodnjih fazah načrtovanja (slika 1) – v integralni del zasnove samega izdelka – saj tako zmanjšamo stroške kasnejših faz in preoblikovanj ter zajamemo tudi



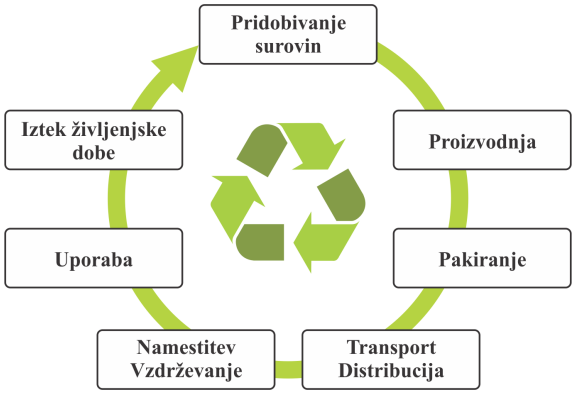
Slika 2: Proces načrtovanja izdelka v primerjavi z okoljskim vplivom [5]

okoljske in gospodarske vidike celotnega življenjskega cikla izdelka (slika 2).

Hkrati z vpeljavo ekološko osveščenega načrtovanja v prve faze načrtovanja izdelkov in v kasnejšo fazo uporabe se je pričel tudi razvoj postopkov, orodij in pristopov k primernejšemu ravnanju z izdelki ob končni življenjski dobi (*End-of-Life* – EoL). Na ta način se olajša razstavljanje, izboljša možnost reciklaže ter izbor materialov, ki optimizirajo uporabo naravnih virov, saj jih je možno kasneje reciklirati, pouporabiti ali obnoviti (t.i. *design for X*).

Zaradi tega se v zadnjem obdobju izpopolnjujejo različni postopki in pristopi k reciklaži, katerih končno uspešnost pa lahko nato upoštevamo že pri izbiri osnovnih metod načrtovanja.

Poleg klasične metode načrtovanja za reciklažo (*design for recyclability*) pa so na voljo tudi številne druge metode, kot npr. načrtovanje razstavljanja (*design for disassembly*), načrtovanje nadgradnje (*design for upgra-*



Slika 3: Umestitev načrtovanja izdelka v življenjski krog izdelka (LCA)

dability), načrtovanje vzdrževanja in popravil (*design for maintainability and serviceability*), načrtovanje ponovne uporabe (*design for reuse*), načrtovanje ponovne izdelave (*design for remanufacturing*) in načrtovanje izbora materialov (*design for material selection*).

Omenjene metode načrtovanja so soodvisne; le njihova kombinacija in interakcija med posameznimi sklopi zajema vse faze, optimizira kakovost in poveča učinkovitost ekološko osveščenega načrtovanja.

3 Umestitev načrtovanja izdelka v življenjski krog

Osnovni namen ekološko osveščenega načrtovanja je zmanjšati ekološko breme skozi celoten življenjski krog izdelka (*Life Cycle Assesment* – LCA). Posamezne faze življenjskega kroga (slika 3) so:

- **pridobivanje surovin**: uporaba vrste in količine surovin na način, ki zmanjšuje okoljsko breme ob pridobivanju (ne glede na njihov izvor), za izredno redke surovine pa minimiziranje njihove količine;
- **proizvodnja**: vključitev proizvodnih procesov na način, ki omogoča nadzor nad potekom, zmanjša odpadke ter porabljeno energijo;
- **pakiranje**: načrtovanje pakiranja na način, da so izdelki čim manjši, s čimer se zmanjša količina pakirnega materiala in zniža strošek skladiščenja ter transporta (ozirma celotna logistika);
- **transport in distribucija**: načrtovanje krajših razdalj med proizvodnim obratom in končnim uporabnikom;
- **namestitev in vzdrževanje**: načrtovanje izdelka s poenostavljeno montažo in nezahtevnim vzdrževanjem;
- **uporaba**: načrtovanje večstransko uporabnega izdelka s čim večjim krogom uporabnikov in visoko uporabno vrednostjo;
- **iztek življenjske dobe**: načrtovanje izdelka na način, ko je možno izdelek delno pouporabiti ali reciklirati s čim manjšim vplivom na okolje, s čimer je cilj pridobiti čim več delov ali materialov iz že odsluženih izdelkov.

4 Direktive in standardizacija

Na področju EU je določen okvir, ki zajema splošno politiko o izdelkih (*Integrated Product Policy* – IPP). Ta določa splošno zakonodajo na podlagi okoljskih direktiv, ki urejajo posamezna področja. Z njihovo uvedbo so podjetja primorana uvajati izboljšave v vse faze proizvodnih procesov.

Glavne skupine direktiv, ki urejajo področje ekološko osveščenega načrtovanja v EU, so:

- **Poraba energije v življenjskem ciklu**: Direktiva EU za izdelke, ki so povezani z električno energijo (*Energy-related Products* – ErP), Direktiva EU za izdelke, ki uporabljajo električno energijo (*Energy-using Products* – EuP);
- **Uporaba nevarnih in prepovedanih snovi**: Direktiva EU o evalvaciji in autorizaciji kemikalij (*Registration, Evaluation & Authorisation of CHemicals* – REACH), Direktiva EU o omejevanju nevarnih snovi (*Restriction of Hazardous Substances* – RoHS);
- **Uporaba baterij**: Direktiva EU o baterijah;
- **Strategija ravnanja z izdelki ob zaključku življenjske dobe**: Direktiva EU o odpadni električni in elektronski opremi (*Waste Electrical and Electronic Equipment* – WEEE), Direktiva EU o končni življenjski dobi vozil (*End-of-Life Vehicles* – ELV).

Poleg direktiv so podjetjem, kot vodilo, na voljo tudi skupine standardov ISO 14000, ki zajemajo različna področja in pristope k ekološko osveščenem načrtovanju. Pomembnejši med njimi so:

- **ISO 14001** – načrtovanje ravnanja z okoljem;
- **ISO 14006** – smernice za vključitev okoljsko primerne zasnove;
- **ISO 14010** – smernice za okoljsko presojo;
- **ISO 14020** – splošna načela okoljskega označevanja;
- **ISO 14040** – načrtovanje ravnanja z okoljem – smernice za ocenjevanje življenjskega cikla;
- **ISO 14050** – izrazi in definicije;
- **ISO 14060** – vodič vključevanja okoljskih vidikov v proizvodne standarde.

Z razvojem in dopolnitvijo standardov se nakažejo smernice pričakovanih sprememb zakonodaje in direktiv, na podlagi katerih se lahko podjetja dolgoročno pripravijo za optimalen nastop in konkurenčnost na tržišču.

5 Implementacija konceptov ekološko osveščenega načrtovanja na področju elektrotehnike

Reciklaža elektronskih delov in naprav je težavna zaradi majhnosti in/ali kompleksnosti, prav tako pa zaradi uporabe raznovrstnih materialov. Tipični primeri polizdelkov in naprav so: raznovrstni MEMS, integrirana vezja, nanostrukture, tiskana vezja, zasloni ter medicinska oprema, nadzorni in kontrolni sistemi, majhni in veliki gospodinjski aparati, uporabna elektronika, svetila, električna orodja, igrače in naprave ter IT in telekomunikacijska oprema.

Ob primernem načrtovanju postopkov izdelave naprav lahko že v zgodnjih fazah predvidevamo kasnejše pridobivanje raznovrstnih elementov, snovi in komponent iz iztrošenih naprav, kot so npr. platina, zlato, srebro, baker, raznovrstne plastike in plastične mase, steklo, moduli tiskanih vezij, integrirana vezja, motorji, zasloni, baterije ipd.

Za zmanjšanje vpliva na okolje se poslužujemo tudi različnih prijemov pri načrtovanju, kot npr. zasnova modularnih komponent v napravah, miniaturalizacija (kar ima lahko tudi negativen vpliv, saj otežuje postopek reciklaže), optimizacija zgradbe naprav ipd.

V poteku proizvodnega procesa izdelave elektronskih naprav so uporabljene in se kot njegov stranski produkt ustvarjajo kemikalije, ki so za okolje toksične, zato je eden od temeljnih ciljev ekološko osveščenega načrtovanja le-te zamenjati s primernejšimi ali vsaj delno zmanjšati njihovo uporabljeno/proizvedeno količino. Direktiva RoHS trenutno omejuje uporabo živega srebra, kadmija, šestvalentnega kroma, polibromiranega bifenila (PBB), polibromiranega difenil etra (PBDE); prav tako pa so v fazi sprejema predlogi, da naj bi se na seznam uvrstila prepoved uporabe nanosrebra in uvedlo označevanje uporabljenih nanotehnologij v izdelkih [6].

6 Sklep

Zaradi povečevanja osveščenosti uporabnikov so podjetja posledično vedno bolj spodbujena k razmišljanju o družbenih in okoljskih vplivih njihovega delovanja. Poleg tega so zakonodajni organi primorani ob očitnih vplivih na okolje zaostrovati zakonodajo in s tem urejati dopustno delovanje podjetij.

Podjetja se približujejo trem ciljem – zmanjševanju porabe surovin in vplivov na okolje (poraba virov in energije, škodljivi stranski produkti), ponovni uporabi elementov, polizdelkov in izdelkov, ki niso več uporabni na tržišču ter čim bolj učinkoviti reciklaži preostankov.

Glavnina odgovornosti sloni na načrtovalcih standardov, direktiv in zakonodaje na eni strani ter v čim bolj učinkoviti vpeljavi le-teh prek načrtovalcev izdelkov v podjetjih na drugi strani, saj bo le tako lahko družba kot celota zagotovila pogoje čistejšega okolja naslednjim generacijam in razvoju v njem.

Literatura

[1] R. Gheorghe in K. Ishii: "Eco-Design Value Alignment – Keys to Success. STANFORD School of Engineering Publication, pp. 1–32. Dosegljivo: https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/other-eco-design-value-aligment.pdf

[2] EN ISO 14006 (2011) Environmental Management Systems – Guidelines for Incorporating Ecodesign. July 2011, Brussels.

[3] J. B. Milojković, V. B. Litovski: "Eco-design in Electronics - The State of the Art". FACTA UNIVERSITATIS, Series: Working and Living Environmental Protection, vol. 2, no. 2, 2002, pp. 87–100, July 2002. Dosegljivo: http://facta.junis.ni.ac.rs/walep/walep2002/walep2002-01.pdf

[4] K. Schischke, M. Hagelüken in G. Steffenhagen: "An Introduction to EcoDesign Strategies – Why, What and How?". July 2018. Dosegljivo: https://www.researchgate.net/publication/268001428_An_Introduction_to_EcoDesign_Strategies_-_Why_what_and_how

[5] J. O'Hare, E. Cope in S. Warde: "Five Steps to Eco Design – Improving the Environmental Performance of Products through Design". GRANTA Design, pp. 1–16, 2015. Dosegljivo: http://www.grantadesign.com/download/pdf/FiveStepsToEcoDesign.pdf

[6] M. Keskinen: "End-of-Life Options for Printed Electronics. V: Waste Electrical and Electronic Equipment (WEEE) Handbook, 2012.

[7] T. V. Tran, S. Brisset in **P**. **B**rochet: "Approaches for the Ecodesign in Electrical Engineering Application to a Safety Transformer". 2009 6th International Multi-Conference on Systems, Signals and Devices, 23–26 March 2009, Djerba, Tunisia. Dosegljivo: https://ieeexplore.ieee.org/document/4956695/

[8] Ecodesign. The European Commission. Dosegljivo: http://ec.europa.eu/growth/industry/sustainability/ecodesign_en

[9] F. Zhang, M. Rio in P. Zwolinski: "Dynamic Ecodesign Strategic Options for Electric-Electronic Industry". ED2E 2016 – International Conference on Eco-Design in Electrical Engineering, March 2016, Grenoble, France. Dosegljivo: https://hal.archives-ouvertes.fr/hal-01357212/document

[10] V. Goodship in A. Stevels, eds.: "Waste Electrical and Electronic Equipment (WEEE) Handbook. A Volume in Woodhead Publishing Series in Electronic and Optical Materials". 1st Edition, Woodhead Publishing Limited, 2012. Dosegljivo: https://www.elsevier.com/books/waste-electrical-and-electronic-equipment-weee-handbook/goodship/978-0-85709-089-8

Analysis of a power supply bypassing approaches

Jakob Bevc
University of Ljubljana, Faculty of Electrical Engineering
E-mail: jakob.bevc@gmail.com

Abstract. *In this paper analysis is presented how different values, arrangements and placement of bypass capacitors affects stability of the power supply voltage. Analysis is supported by measurements on a designed and manufactured printed circuit board and compared with results of simulations made in LT Spice. Simulations are developed, step by step from simplest one 100nF decoupling capacitor, to optimal bypass solution. Voltage fluctuations were reduced for 90 mV after adding second (equal) bypass capacitor in parallel to existing. Adding more than two capacitors in parallel did not show significant improvement. Using damp capacitor in addition to existing bypass capacitor reduced voltage noise for additional 80 mV down to 180 mV, being our final solution. Measurements on the PCB are well supported by simulations.*

1 INTRODUCTION

Every electronic circuit has power supply connections that supply constant DC voltage to all loads. One of the problems that has to be dealt with is that many high-speed electronic components are commonly used, which sink high-frequency currents. Power supply connections will always introduce impedance for this high-frequency currents that consequently cause power supply voltage noise. With constantly improving technology, the trend for high-performance integrated circuits is toward higher operating frequency and lower power supply voltages. For that reasons voltage noise must be kept at a minimum value, to ensure correct operation of the design [1], [2].

Voltage noise can be reduced by minimizing the impedance of the power supply circuitry, where the most common technique is to use bypass capacitors, which are maintaining low impedance paths for high-frequency currents. Impedance must be kept below the maximum impedance in whole range of the current transients. For lower frequencies up to 1 MHz, the bulk (electrolytic) capacitor is commonly sufficient [2]. However, at higher frequencies, ceramic bypass capacitors are used. In addition, they provide a localized source of DC power and are particularly useful in reducing peak current surges propagated across the board [1]. In ideal case, a capacitor has no losses and no series inductance, but in reality, every capacitor has its parasitic equivalent series resistance (ESR) and equivalent series inductance (ESL) [3]. These two parameters determine the frequency range, where the impedance is low enough, to serve its purpose. The impedance of capacitor drops with frequency up to the self-resonance frequency (1), where the minimal impedance is achieved [4].

$$f_0 = \frac{1}{2\pi\sqrt{ESL \cdot C}} \quad (1)$$

From that point on impedance increases with frequency due to ESL, which is influenced also by the connection traces and vias of the capacitor, as shown in the Figure 1.

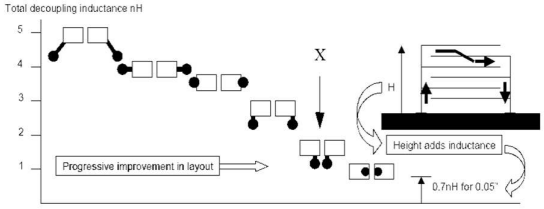


Figure 1: Inductance influenced by the connections.

To decrease the parasitic effects more capacitors can be put in parallel. If different capacitors are used, this causes unwanted parallel resonances, which lead to high impedance at this resonance frequencies. Between the self-resonant frequencies of two different capacitors, the impedance of the larger capacitor is essentially inductive, whereas the impedance of the smaller capacitor is capacitive. In this frequency range there exists a parallel resonant LC circuit. Around this resonant point, the impedance of the parallel combination is larger than the impedance of either isolated capacitor [3], [5].

For proper operation, capacitor values must be calculated for a specific load type. Furthermore, the dielectric material of the capacitor must be properly selected together with the case size. The bypass capacitors are often chosen according to past usage experiences and rules-of-thumb. In this paper, we will try to show, how all these properties influence in reality and compare that with simulations.

2 EXPERIMENT

2.1 Printed circuit board (PCB)

As shown in the Figure 2, we have designed a printed circuit board (PCB) with the MIC 4605-2YM chip, a power MOSFET gate driver with two outputs, each providing 1A of output current [6]. Its outputs are connected to resistor and a capacitor, connected in series to simulate the MOSFET gate impedance. When switching, the capacitors charge and discharge and the charging current is supplied from the power supply. With this configuration, the amplitude of current pulses can be controlled (resistor) and the frequency with input PWM signal. Resistor values determine the amplitudes of current pulses.

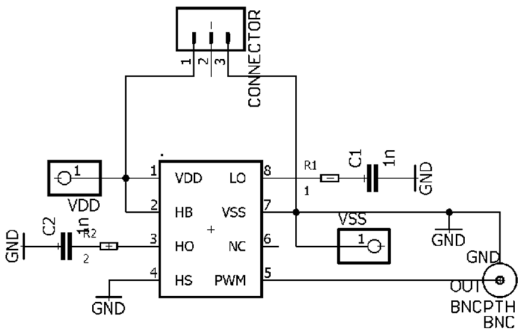


Figure 2: Detailed schematics of the chip

The PCB was designed using EAGLE PCB design software [7] in a way that it offers various positions for placement of decoupling capacitors and that the power supply traces are wide enough to allow different case sizes of capacitors to be soldered. Screw-terminal connector is positioned on the left side of the board for DC power supply. 2 mm wide power supply trace is routed from the connector to the chip MIC 4605-2YM on the right side of the board.

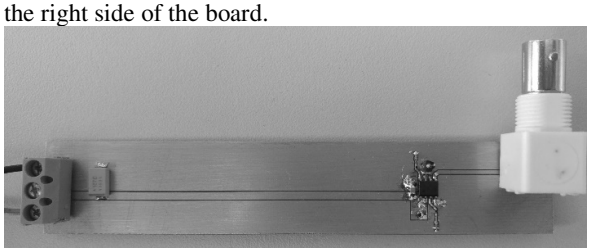


Figure 3: PCB design

For signal source connection to drive the chip, a BNC connector is provided. All ground is routed using ground plane. For measurements purposes, a pin header was put on VDD and GND pins of the chip. Bottom side of the PCB has a 2 mm wide trace directly under the chip (via at VDD pin), offering another option for positioning a decoupling capacitor. Figure 4 shows the detailed routing of the chip.

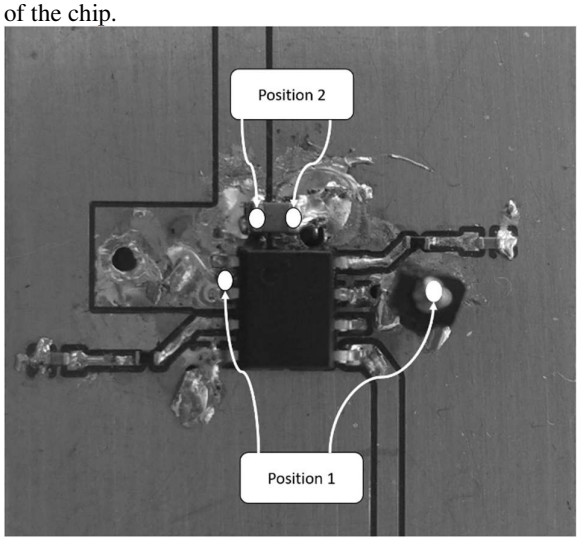


Figure 4: Chip routing details and probe positions

2.2 Measurement setup

The board was connected to a DC laboratory power supply, providing constant 12V DC voltage, controlled with multimeter, while a function generator was used to generate PWM signal. Voltage fluctuations of power supply were measured with an oscilloscope. Most of the measurements were made on pin-headers and also on some different spots on the PCB. Data from oscilloscope was later processed in Matlab.

2.3 Spice simulation model

Simulations were made with LT SPICE simulator from Linear Technology [8]. Since the LT spice library does not include a model for the MIC 4605-2YM, a model had to be added. A very useful source of information was the manufacturer's website, where they provide a model for another simulator: MPLAB Mindi [9]. This simulator was initially used for the transient simulation of the circuit. Value of each resistor was selected to get maximum amplitude of 1 A from both outputs. This simulation gave us enough information to make a model in LT Spice. To model the outputs of the chip, MOSFET transistors (included in LT Spice library, Si1555DL_P from Siliconix) were chosen that provide equal current pulses, as simulated in the MPLAB Mindi simulator. Correct MOSFET driving including invert and dead time was solved with two phase shifted square signal sources.

The chip model was included in a transient simulation with all necessary parasitic for the AC model of the power supply circuit, as shown in the Figure below.

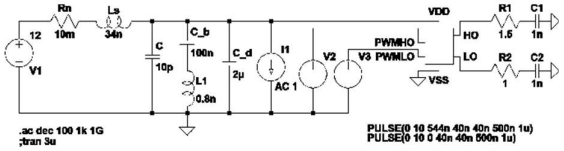


Figure 5: Basic simulation design

For accurate simulations, impedance of the traces was taken into consideration and included in simulation. We have used Saturn PCB design software [10] using basic trace parameters: trace width/height material and copper thickness, giving a value of capacitance and inductivity per centimeter. Trace parameters for our design are: 2 mm width, 1.6 mm height, material is FR-4 STD and copper thickness 35 µm. This results in inductance of 3.78 nH/cm and capacitance of 0.97 pF/cm. Multiplying this with trace length (9.5 cm) gives a good approximation for L_0 = 35.91 nH and C_0 = 9.21 pH and these results were used in the trace AC model.

3 RESULTS

3.1 Power supply bypassing design using simulations

Simulations of power supply impedance were made using different bypassing arrangements starting from simple to more advanced.

Simulation results using realistic 100 nF bypass capacitor with ESR set to 10 mΩ and ESL to 0.5 nF, are shown in the Figure 6. Due to ESL a series resonance appears at around 10.5 MHz, where impedance is limited

by ESR. Further on, the impedance starts to increase again as capacitor exhibits inductive behavior.

Putting more capacitors in parallel can improve impedance trough whole frequency range. Figure 6 presents simulation results, where two 100 nF capacitors are connected in parallel. Since both capacitors are the same, the resulting capacitance increases by two, while the inductance and resistance drops by a factor of two.
The increase of capacitance decreases the parallel resonance frequency. The impedance at frequencies above parallel resonance decreases significantly.

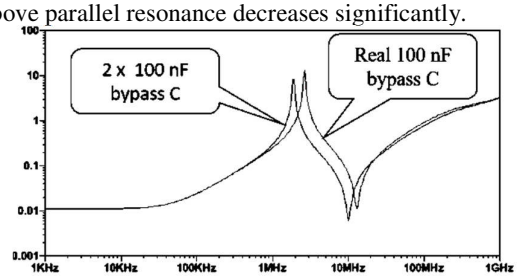


Figure 6: Impedance graphs of different bypass arrangements using a realistic capacitor model

Figure 7 presents bypassing arrangement using a 100 nF capacitor (same as in previous simulations) connected in parallel with a 1 nF capacitor with better parasitic values (ESR = 1 mΩ and ESL = 0.25 nF). Since another (better) capacitor was used, lower impedance is generally expected. But in the frequency range between self-resonances of capacitors additional parallel resonance of capacitors is introduced, where the impedance is actually larger than previous, where no additional capacitor was used.

It turns out, that using equal capacitors in parallel is the best option. The only problem that has to be solved is the first parallel resonance. Usually this is accomplished by introducing a damp capacitor with right capacitance and ESR to form a critical damped resonance circuit [11].

$$C_{damp} = 2.2\ uF, \quad ESR_{damp} = 0.235\ \Omega \tag{2}$$

With this solution, the power supply impedance is lower than 0.23 Ω in the whole frequency range of interest (0-50 MHz), as show in the Figure below.

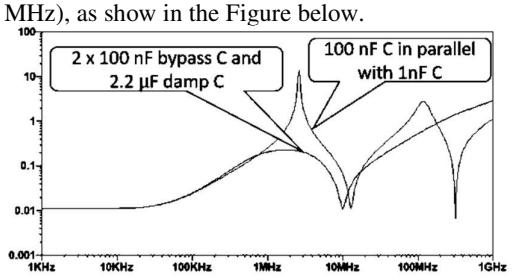


Figure 7: Impedance graph

3.2 Measurement of voltage ripple at different positions

This measurement shows importance of bypass capacitor placement. Figure 8 and 9 present two simulations and measurements results with equal bypass capacitors and different probe positions 1 and 2 (labeled in the Figure 3). Measurement closer to the capacitor (position 2) gives lower voltage fluctuations without spikes. Although both positions are very close to each other, the difference in voltage ripple is 100 mV. In simulation, the extra line length to the position 1 was modeled as additional impedance in series with the bypass capacitor. Impedance was estimated with Saturn PCB design program to be 1 nF. The good match between simulation and measurement validates the model parameters.

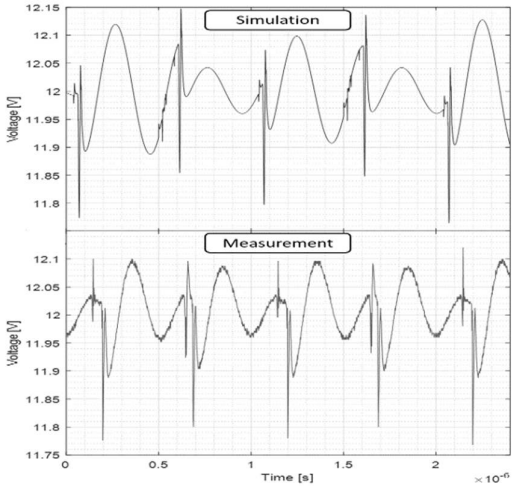


Figure 8: Simulation and measurement on PCB, position 1

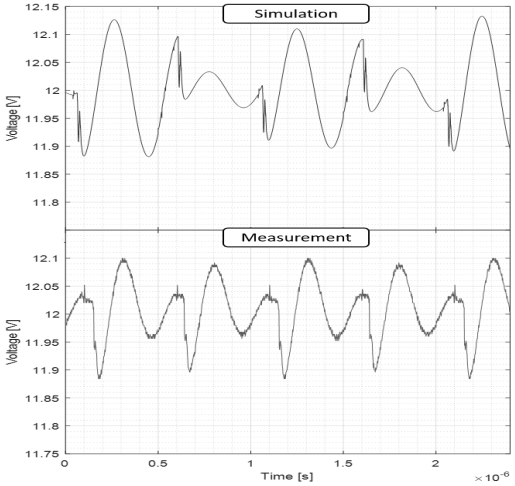


Figure 9: Simulation and measurement on PCB, position 2

3.3 Bypass effectiveness of parallel connected capacitors

We started with two equal 100 nF bypass capacitors and compared that with previous results. Figure 10 presents usage of one 100 nF bypass capacitor, $U_{pp} \approx 260$ mV. With two capacitors, voltage fluctuations were reduced to $U_{pp} \approx 170$ mV. Additional measurements on the PCB indicated that there is no significant decrease in voltage fluctuation after adding third bypass capacitor. This is mainly due to the fact, that more capacitors require more place and consequently more trace length. Thus, at one-point adding more capacitors in parallel is no more effective.

The best results are made with capacitors stacked on top of each other, this minimizes trace length and takes least amount of space on PCB.

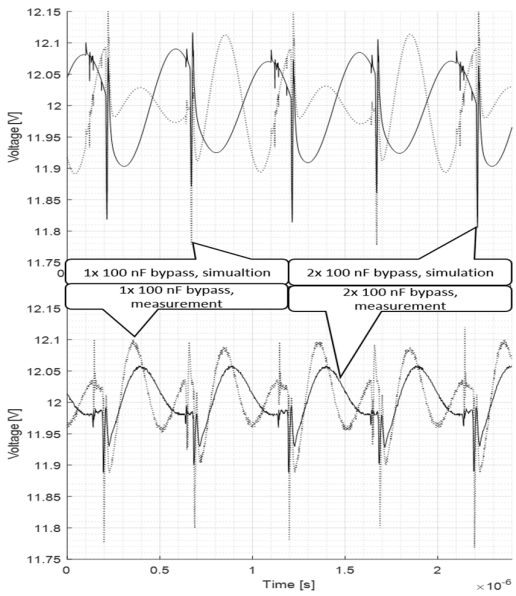


Figure 10: Comparison of parallel connected bypass capacitors

3.4 Optimizing bypassing with a damp capacitor

To reduce parallel resonance, damp capacitor is added in parallel to the existing bypass capacitor. According to the [11], capacitance and ESR values of damp capacitor were calculated.

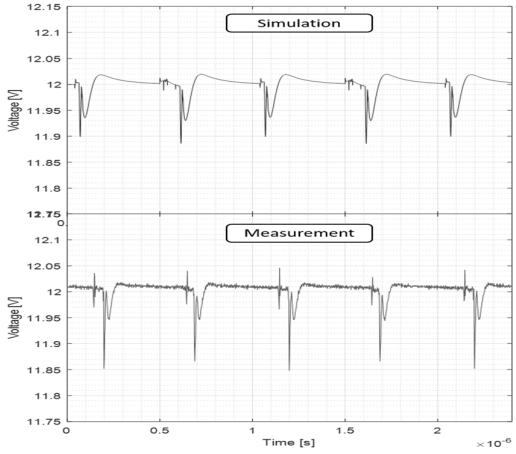


Figure 11: Simulation and measurement with bypass and damp capacitor, position 1

Required values were realized with ceramic capacitor, its parameters: 2.2 uF capacitance in series with 0.235 Ω resistor, to get ESR in desired range. Simulation AC analysis confirmed decrease of impedance at parallel resonance frequency, shown in Figure 7. After adding damp capacitor voltage fluctuations are limited to 180 mV, which is an 80 mV less, compared to usage of bypass capacitor alone. Simulations give somewhat lover values. This is expected, due to parasitic values, which cannot be approximated with high accuracy.

4 CONCLUSION

This paper presented different bypassing approaches. Analysis showed some simple and cost-efficient solutions and pointed out common mistakes one could do. Development of simulation lead to bypassing solution later tested on a real PCB. Usage of parallel bypass capacitors and damp capacitor showed significant improvement in voltage stability. Position of bypass capacitor is also very important. Every additional trace length causes noticeable increase of noise. Good match between simulation and measurement proved, that with accurate simulation we can predict bypass capacitors arrangement early in the development process of electronic circuit and avoid later malfunction. For further improvement of this research more testing could be done concentrating on capacitor type, manufacturer and materials.

ACKNOWLEDGEMENTS

I would like to express my very great appreciation to doc. dr. Marko Jankovec for his help during the development of this research work. I would also like to thank to every member of LPVO.

REFERENCES

[1] T. Williams, EMC for Product Designers, 2006, pp. 300-312.

[2] L. D. Smith, R. E. Anderson, D. W. Forehand, T. J. Pelc and T. Roy, "Power Distribution System Design Methodology and Capacitor Selection for Modern CMOS Technology," 3 august 1999.

[3] M. Montrose, Printed Circuit Board Design Techniques for EMC Compliance, New York: IEEE, 2000, pp. 77-109.

[4] R. Tanmoy, L. Smith and J. Prymak, "ESR and ESL of Ceramic Capacitor Applied to Decoupling Applications," in *Electrical Performance of Electronic Packaging, 1998*, New York, 1998.

[5] A. Mediano, "Using Capacitors in Parallel: Dangerous?," 29 december 2016. [Online]. Available: https://incompliancemag.com/article/using-capacitors-in-parallel-dangerous/.

[6] MICROCHIP, „MIC 4605-2YM datasheet," 2018. [Elektronski]. Available: http://ww1.microchip.com/downloads/en/DeviceDoc/20005853A.pdf. [Poskus dostopa 9 8 2018].

[7] „Autodesk EAGLE," [Elektronski]. Available: https://www.autodesk.com/products/eagle/overview.

[8] A. devices, „LTspice," [Elektronski]. Available: http://www.analog.com/en/design-center/design-tools-and-calculators.html.

[9] MICROCHIP, „MICROCHIP MPLAB Mindi," [Elektronski]. Available: http://www.microchip.com/mplab/mplab-mindi.

[10] S. P. Design, „Saturn PCB Design," 2017. [Elektronski]. Available: https://www.saturnpcb.com/#page=page-1.

[11] K. Kundert, "The Designer's Guide," january 2004. [Online]. Available: http://www.designers-guide.org/Design/bypassing.pdf. [Accessed 9 may 2018].

Telekomunikacije

Telecommunications

Celovit sistem za upravljanje voznega parka v potniškem prometu

Dalibor Igrec, Amor Chowdhury, Gregor Felzer

MARGENTO R&D d.o.o., Gosposvetska cesta 84, 2000 Maribor, Slovenija
E-pošta: dalibor.igrec@margento.com

A comprehensive fleet management system for passenger transport

Abstract. *In order to increase or at least sustain the quality of live for citizens and visitors, improving management capabilities of the services is a necessary task of everyone involved in urban planning. With the rising city population and urbanization smart managing of city services becomes more and more important. Public transit, as one of the most important city services, stands to gain the most from the introduction of the smart management system. This article presents solution called comprehensive fleet management system for passenger transport. Featured example is based on management of bus transit services. Presented solution is comprised of mobile units, installed in bus vehicles and platform acting as central back-office processing system. Key tasks of the platform are managing communication with mobile units and performing data processing. Platform also contains web applications and interfaces for operational management, communicating with drivers and supporting services as well as interfaces for third party integration, such as time schedules, mobile planners etc.*

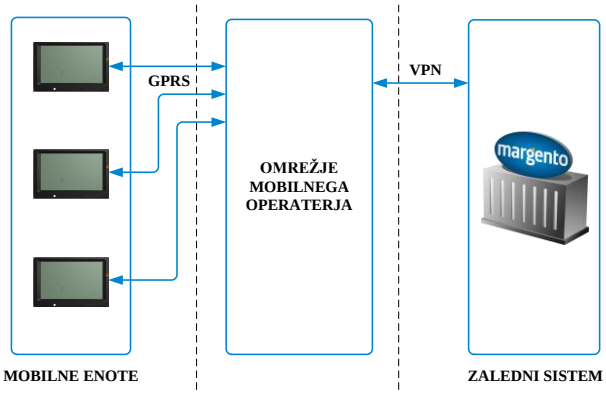
1 Uvod

Upravljanje voznega parka brez informacijskega sistema je v današnjem času nepredstavljivo. Takšni sistemi za upravljanje voznega parka omogočajo sledenje vozil in voznikov, beleženje voženj in delovnega časa voznika, prikaz podatkov o trenutni lokaciji in preteklih vožnjah na zemljevidu, omogočajo enostavno komunikacijo med vozniki in upravljavcem voznega parka, navigacijo voznika, beleženje dogodkov in telemetrije vozila ter optimizacijo voženj, stroškov goriva in ostalih stroškov voznega parka.

Sistemi za upravljanje voznega parka v potniškem prometu poleg navedenih funkcionalnosti vključujejo še nekaj dodatnih. Mednje spada urejanje postaj in linij, urejanje urnikov voženj s časi pričetka linije in časi prihodov na postaje, sledenje avtobusnih vozil na liniji, beleženje voženj po urnikih z ugotavljanjem razlik med urnikom in dejanskim prihodom na postajo ter izračunavanje predvidenega časa prihoda na naslednjo postajo s pomočjo analize podatkov preteklih voženj. Takšen sistem vključuje dodatne senzorje za nadzor avtobusnega vozila, senzorje za nadzor števila potnikov ter dodatne sisteme za obveščanje potnikov, kot so mobilne in spletne aplikacije, informacijske table ter avtomatske in/ali ročne plačilne sisteme za obračunavanje prevozov.

2 Komponente sistema

Celovit sistem za upravljanje voznega parka v potniškem prometu (slika 1) zajema mobilne enote in zaledni sistem. Mobilne enote se namestijo v avtobusna vozila in preko GPRS mobilnega omrežja komunicirajo z zalednim sistemom, ki je nameščen v ustreznem strežniškem prostoru in dosegljiv preko interneta.



Slika 1: Komponente sistema

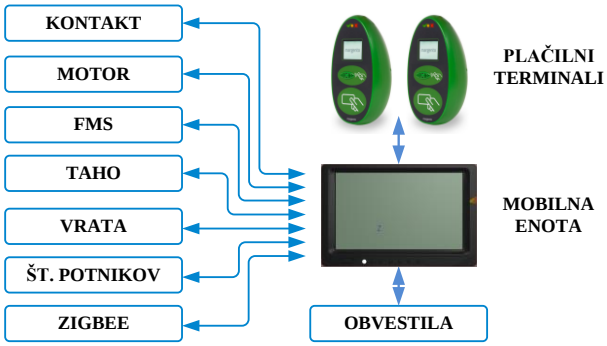
Za zaščito prenosa podatkov med mobilnimi enotami in zalednim sistemom se uporablja kombinacija zasebni APN in virtualno zasebno omrežje VPN. Zasebni APN predstavlja zaprto mobilno omrežje na relaciji mobilna enota in mobilni operater. Povezava med mobilnim operaterjem in zalednim sistemom poteka preko internetne povezave zaščitene z VPN standardom.

Podatki, ki se prenašajo med mobilno enoto in zalednim sistemom, se sestavljajo v sporočila zapisana v JSON formatu. Vsako sporočilo se shrani v čakalno vrsto, za katere prenos v zaledni sistem oz. v mobilno enoto poskrbi komunikacijski vmesnik. Vsako sporočilo vsebuje parametra tip in podtip, ki določata strukturo in vsebino. Na podlagi teh parametrov procesna logika pravilno obdela sporočilo in nanj ustrezno odgovori, v kolikor je to potrebno. Sporočila so lahko tudi binarni podatki, ki se uporabljajo za nadgradnjo programske opreme mobilne enote oz. za prenos drugih strnjenih podatkov. Takšen prenos sporočil omogoča zelo raznolik način komunikacije na visokem programskem nivoju.

3 Mobilna enota

Mobilno enoto [1] sestavljajo mobilni računalnik in ostale komponente združene v ohišje z LCD zaslonom občutljivim na dotik. Ostale komponente zajemajo GPRS vmesnik, GPS vmesnik, pospeškometer, digitalni vhodi in komunikacijski vmesniki (RS232, RS485, USB).

LCD zaslon občutljiv na dotik služi za prijavo voznika in vnos podatkov. Omogoča enostavno izbiro menijev za prikaz in delo z različnimi moduli mobilne enote. GPRS vmesnik skrbi za mobilno dvosmerno komunikacijo med mobilno enoto in zalednim sistemom. GPS vmesnik se uporablja za zajemanje trenutne lokacije. Na podlagi lokacije se beležijo prevožene poti, podatek o lokaciji se uporablja za detekcijo avtobusnih postaj, obenem pa se podatek o lokaciji beleži tudi na druge dogodke (npr. podatki senzorjev). Pospeškometer je uporabljen za beleženje pospeškov in pojemkov, ki so vir podatkov za ugotavljanje kvalitete vožnje voznika.



Slika 2: Vse možnosti uporabe digitalnih vhodov in komunikacijskih vmesnikov mobilne enote v avtobusnem vozilu

V nadaljevanju sledi prikaz funkcionalnosti glede na posamezne sklope mobilne enote.

3.1 LCD zaslon občutljiv na dotik

Predstavlja robustno izvedbo LCD zaslona, preko katerega voznik v času opravljanja dejavnosti izvaja različne naloge: vnaša število vozovnic pri validaciji z mestno kartico, opazuje dejansko stanje glede prihoda na naslednjo avtobusno postajo (v kolikor prehiteva ali zamuja), vnaša manjše okvare posameznih sklopov avtobusa, ki jih zasledi med vožnjo, pošilja vnaprej pripravljena sporočila, prejema sporočila informativne narave, …

3.2 GPRS modul

Omogoča stalno komunikacijsko povezavo s sistemom za upravljanje in s tem prenos podatkov v obe smeri v realnem času.

3.3 GPS modul

Omogoča zajemanje trenutne lokacije v vsakem trenutku in s tem računanje prevožene razdalje od začetne postaje.

3.4 3-osni pospeškometer

Omogoča detekcijo tipa vožnje (pospeške in pojemke), na podlagi katerih se izvaja enostaven prikaz vozniku na zaslonu (zeleno, rumeno in rdeče področje) ter podrobnejša analiza v sistemu za upravljanje.

3.5 ZigBee modul

Omogoča komunikacijo z napovedno tablo na avtobusni postaji. Namen te komunikacije je dodatna detekcija približevanja avtobusa k avtobusni postaji in dodatna korekcija prihoda na sami prikazovalni tabli. Domet komunikacije je odvisen od okolice in se lahko giblje do nekaj 100 metrov.

3.6 Digitalni vhodi

Na digitalni vhod je prvenstveno priključen kontakt motorja avtobusa. Na podlagi te informacije se sistem mobilne enote odloča, kdaj bo zaključil delovanje in se postavil v stanje nizke porabe oz. se izklopil iz napajanja. Na ostale proste digitalne vhode so priključena posamezna vrata, na podlagi katerega se izvaja detekcija odprtosti posameznih vrat.

3.7 FMS vmesnik

FMS vmesnik omogoča varno povezavo na CAN vodilo avtobusa in/ali digitalnega tahografa. Na ta način pridobimo natančne informacije, kot so nivo goriva v rezervoarju, trenutna in povprečna poraba goriva, ekonomičnost vožnje, števec kilometrov ipd. Pri izvedbah z digitalnim tahografom je dodatno omogočeno oddaljeno prenašanje šifriranih DDD tipa tahografskih datotek direktno do sistema za njihovo nadaljnjo obdelavo.

3.8 Serijski vmesniki

Posamezni vmesnik se glede na potrebe nastavi na RS232 ali RS485 tip delovanja. V praksi so na serijske vmesnike mobilne enote priključene sledeče enote:

- en, dva ali trije terminali za validacijo elektronskih vozovnic (odvisno od izvedbe),
- tiskalnik za tiskanje vozovnic v primestnem prometu,
- senzor za štetje potnikov (IBIS IRMA protokol),
- zunanji ZigBee vmesnik,
- sprednja in/ali stranska tabla z oznako linije po kateri avtobus trenutno vozi,
- sistem za oglaševanje v avtobusu,
- sistem za zvočno napoved naslednje avtobusne postaje.

Glavne funkcionalnosti mobilne enote so prijava voznika, ugotavljanje lokacije vozila, beleženje in javljanje teh podatkov v zaledni sistem. Vse glavne in dodatne funkcionalnosti so prikazane na sliki 3.



Slika 3: Funkcionalnosti mobilne enote

Izbira statusa voznika upravljavcu voznega parka pove ali je voznik na liniji ali na poti v garažo. Vnos službe voznika služi pridobivanju podatkov, na podlagi katere mobilna enota s centra pridobi vse podatke o pričetkih linij, dolžini linije, postajah, njihovih lokacijah ter razdaljah med njimi in časih prihoda na posamezno postajo. To vozniku omogoča enostaven nadzor nad opravljenim delom, obenem pa s temi podatki mobilna enota izvaja detekcijo vožnje po liniji in detekcijo avtobusnih postaj. Te podatke mobilna enota sproti pošilja v zaledni sistem, kjer podatki upravljavcu voznega parka služijo za nadzor nad opravljenimi vožnjami v realnem času. Ti podatki se v zalednem sistemu beležijo za nadaljnjo obdelavo, poročila in analizo, obenem pa se nad temi podatkih izvaja napovedni izračun prihoda na naslednje postaje za obveščanje potnikov preko spletnih portalov, mobilnih aplikacij in prikazovalnikov na avtobusnih postajah. Prodaja vozovnic omogoča izbiro števila potnikov, izbiro tarif, beleženje gotovine in drugih plačil. Vnos okvar omogoča enostavno sprotno beleženje morebitnih okvar in opažanj na vozilu. Ti podatki se prenašajo v zaledni sistem v realnem času, kar ustreznim službam omogoča vnaprejšnjo pripravo razporeda popravil in morebitno nabavo rezervnih delov. Sporočanje predstavlja klasično pošiljanje sporočil med upravljavcem voznega parka in vozniki oz. mobilnimi enotami, ki služi za obveščanje o stanju na cestah, iskanju izgubljenih predmetov, menjavi služb, vozil, itd.

V nadaljevanju sta prikazana primera dveh zaslonskih slik aplikacije na mobilni enoti.

Slika 4 prikazuje voznikov osnovni meni, ki podaja na izbiro 6 podmenijev: Bus, Status, Vozovnica, Tarife, Sporočila in Pregled.



Slika 4: Prikaz osnovnega menija

Slika 5 prikazuje trenutno, prejšnjo in naslednjo postajo na trenutni liniji.

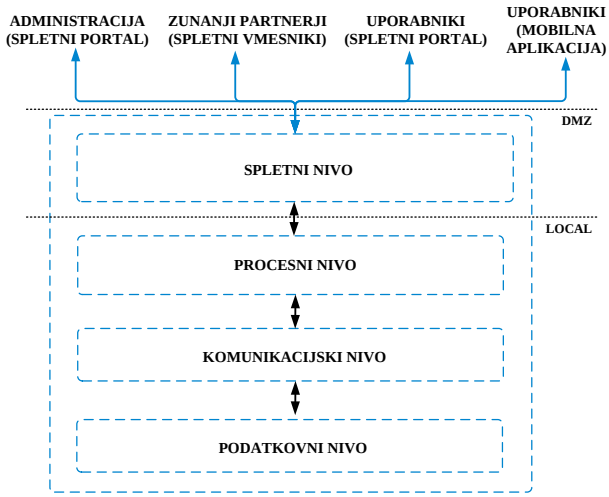


Slika 5: Prikaz podmenija Bus

4 Zaledni sistem

Oprema, ki sestavlja zaledni sistem [2], zajema strežnike, diskovno polje, sistem za arhiviranje podatkov in mrežno opremo. Oprema je postavljena v visoko razpoložljivo konfiguracijo in je nameščena v ustrezen strežniški prostor, kjer sta zagotovljena varnost in nadzor, zagotovljene so hitre in visoko razpoložljive internetne povezave, visoko razpoložljiv vir električne energije s sistemom za neprekinjeno napajanje ter visoko razpoložljiv sistem hlajenja. Celoten sistem je postavljen v primernem strežniškem virtualnem okolju oz strežniškem oblaku.

Strežniki zalednega sistema (slika 6) so postavljeni v več logičnih nivojev med katerimi je zagotovljena varnost na podlagi seznama dostopa in kontrole prometa (ang. access lists, intrusion prevention system).



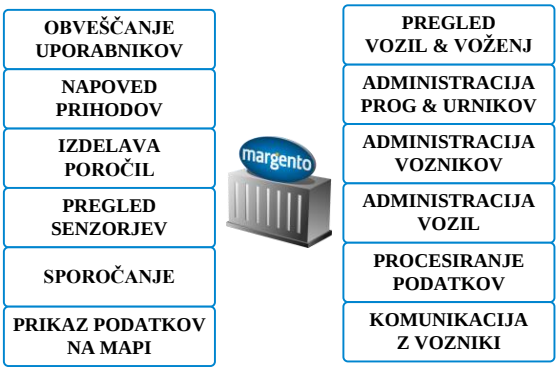
Slika 6: Struktura zalednega sistema

Podatkovni nivo zajema podatkovno bazo, kamor se beležijo in hranijo vsi podatki potrebni za upravljanje. Prav tako so v njej shranjeni vsi podatki mobilnih enot, ki se arhivirajo za kasnejšo analizo ter izdelavo poročil oz. so potrebni za obdelave v realnem času. Komunikacijski nivo skrbi za povezavo z mobilnimi enotami in prenose sporočil. Procesni nivo zajema sistemske servise za obdelavo sporočil mobilnih enot.

Procesni nivo zajema ostalo procesiranje podatkov kot so izdelava poročil, analize, izračun napovedi prihodov na postaje na podlagi preteklih podatkov ter drugo. Komunikacijski in procesni nivo sta dimenzionirana na način, da lahko obdelata večje število sporočil in drugih obdelav hkrati. To omogoča, da je v zaledni sistem lahko povezano večje število mobilnih enot, kjer vse komunicirajo istočasno in v realnem času.

Spletni nivo zajema spletne portale za upravljanje ter spletne portale in mobilne aplikacije za obveščanje oz. prikaz urnikov voženj končnim uporabnikom. Ena ključnih komponent spletnega nivoja so tudi spletni vmesniki, ki se uporabljajo za prenos podatkov iz oz. v zunanje sisteme. V zunanje sisteme se prenašajo podatki o prihodih na postaje (prikazovalne table na postajah), trenutno stanje vozil za različne obdelave ter linije, postaje in urniki v druge sisteme, ki se uporabljajo za usmerjanje oz. navigacijo potnikov (npr. Google Transit). Iz zunanjih sistemov se prenašajo podatki o urnikih, seznami voznikov, vozil in drugo.

Funkcionalnosti zalednega sistema povzema slika 7. Osnovne funkcionalnosti zalednega sistema so komunikacija z mobilnimi enotami, procesiranje podatkov, ki se generirajo na mobilnih enotah oz. pri uporabnikih ter priprava in obdelava teh podatkov za prikaz rezultatov upravljavcem voznega parka kot tudi končnim uporabnikom.



Slika 7: Funkcionalnosti zalednega sistema

5 Spletna aplikacija za upravljanje

Spletna aplikacija za upravljanje zajema administracijo vozil, mobilnih enot, voznikov, linij, postaj in urnikov. Te nadalje zajemajo spletne forme, ki omogočajo pregled, urejanje, dodajanje in brisanje podatkov. Pomemben segment za delo upravljavca voznega parka je pregled trenutnega stanja vozil, kjer se preverja javljanje vozil, prijava voznika in njegov status. Iz pregleda trenutnega stanja je možem pregled vozila oz. več vozil na mapi. Za zemljevid se uporablja GoogleMaps storitev, brezplačna OpenStreetMap storitev ali drugi podobni servisi; odvisno od naročnika.

Zraven trenutnega stanja vozil je možen pregled poti vozil, ki zajema podatke o vozniku, podatke o času prijave in odjave voznika, čas gibanja in mirovanja vozila, trajanje poti in prevoženi kilometri. Možen je pregled celotne poti enega ali več vozil na mapi. Pregled senzorjev omogoča pregled in podrobno analizo nad podatki, ki se zbirajo preko senzorjev mobilne enote. Sporočanje omogoča pošiljanje sporočil, kjer je možen pregled poslanih in prejetih sporočil ter pregled združenih sporočil posameznega vozila (potek pogovora med vozilom in upravljavcem voznega parka). Pri poslanih sporočilih je na voljo informacija, kdaj je bilo sporočilo prejeto na mobilni enoti in kdaj prebrano s strani voznika.

Napoved prihodov zajema algoritem, ki na podlagi povprečnih časov voženj avtobusov med posameznimi postajami na liniji, izračuna predviden čas prihoda na naslednje postaje. Povprečni časi voženj med posameznimi postajami na liniji se izračunavajo na podlagi preteklih podatkov. Upoštevajo se primerni časovni intervali, da se zajamejo prometne konice, delovni dnevi, vikendi, prazniki in različnimi urniki (poletni, zimski, šolski, počitnice, …). Napovedni časi izboljšajo točnost prihodov avtobusov na posamezne postaje in s tem končnim uporabnikom omogočajo boljše časovno planiranje ter manj čakanja na avtobusni postaji. Napovedni časi in urniki voženj so končnim uporabnikom na voljo preko spletne in mobilne aplikacije, preko spletnih vmesnikov pa se čas prihoda avtobusa za posamezno postajo prikazuje na prikazovalnikih, ki so nameščeni na nadstreške avtobusnih postaj. Pomembna funkcionalnost zalednega sistema je izdelava poročil, ki omogoča izvoz podatkov v različnih formatih za poročanje in nadaljnjo obdelavo ter analizo.

6 Zaključek

Ključni cilj modernizacije sistema za upravljanje voznega parka predstavlja razbremenitev spremnega osebja, saj ni več potrebna prodaja in/ali validacija vozovnic, temveč le kontrola potnikov. Modernizacija sistema hkrati predstavlja večji korak dostopnosti in prilagodljivosti sistema za potnika, saj mu ves čas in v realnem času podaja informacije o zamudah in sprotnih spremembah voznega reda.

Razvoj v prihodnje zajema širitev funkcionalnosti mobilne enote in zalednega sistema s ciljem poenostavitve nalog in opravil upravljavcev voznega parka ter voznikov, izboljšave načina optimizacije voznega parka in stroškov, kakor tudi izboljšave uporabniške izkušnje upravljavcem in končnim uporabnikom. Razvoj se nadaljuje v smeri izdelave modulov za enostavno razporejanje voznikov in vozil ter širitvijo spletnih vmesnikov za integracijo zalednega sistema naročnika s sistemi drugih ponudnikov.

Literatura

[1] Mobilna enota – interna dokumentacija, Margento R&D d.o.o., Maribor 2018

[2] Zaledni sistem – interna dokumentacija, Margento R&D d.o.o., Maribor 2018

Prenos alarmnih sporočil preko interneta v sistemu Intervencije.net

Tomaž Dežman[1], Grega Jakus[2]

[1]*Aldia, d.o.o., informacijske storitve, Pot za Brdom 100, 1000 Ljubljana*
[2] *Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, Ljubljana, Slovenija*
E-pošta: tomaz.dezman@aldia.si, grega.jakus@fe.uni-lj.si

Transferring alarm messages in Intervencije.net system over internet

Abstract. *Intervencije.net is a system that activates rescue services, like fire departments or emergency medical service teams. The system currently supports collecting information about an incident from various communication channels, including paging system, telephone network, text messages, and forwarding information to members of intervention teams.*

It is now necessary to upgrade the system with internet-based message transfer in the event of an automatic incident detection. Required is a definition of a new protocol, based on TCP protocol and upgraded with client identification and authentication mechanisms, data encryption and communication path availability control.

Clients are simple IoT devices, collecting data from their environment and forwarding alarms to the server. The server part processes alarm data and, when required, notifies the rescue services.

Currently, upgrading of the system is in its testing phase, aimed at discovering errors and troubleshooting as well as establishing, if the system meets the needs of its users. Once the testing phase is concluded, the devices will be introduced into real-world environments. Initially, we plan to install the IntP devices in fire stations and health centres, later however, they will also be offered to individual members of rescue services, so they can stay connected in their flats and houses.

1 Uvod

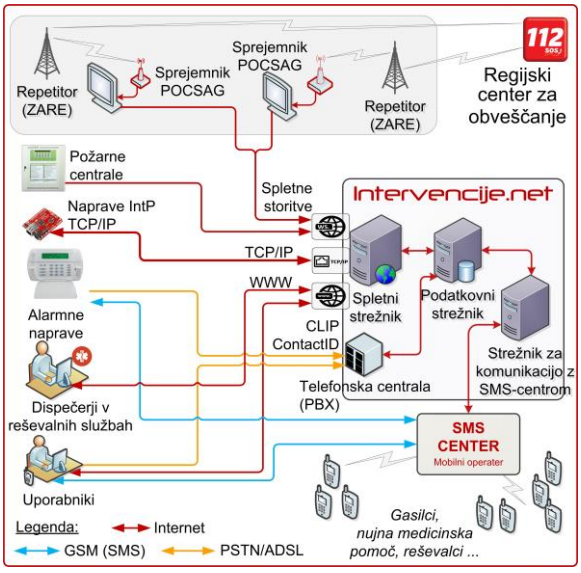
Vpeljava sodobnih informacijsko-komunikacijskih tehnologij na področju zaščite in reševanja pomembno vpliva na učinkovitost reševalnih skupin. S takojšnjim posredovanjem natančnih informacij o nesreči reševalni službi lahko namreč prepolovimo njen odzivni čas, čemur pritrjuje tudi analiza sistema eCall, ki v primeru prometne nesreče samodejno sporoči mesto nesreče pristojnim službam [1]. Poleg tega lahko intervencijo organiziramo učinkoviteje, če imamo informacijo o tem, kdo se je nanjo odzval in kakšno opremo ima na voljo.

V Sloveniji imamo zelo dobro razvite in razvejane službe za zaščito in reševanje. Če bi po zgledu sistema eCall vpeljali samodejno obveščanje pristojnih služb tudi v primeru vseh ostalih nesreč v naravnem in urbanem okolju ter v zasebnih prostorih in javnih ustanovah, bi s tem omilili posledice teh nesreč in tako pomembno pripomogli k zaščiti ljudi in njihovega imetja.

O aktualnosti problematike priča množica projektov, ki so v teku, izmed katerih se večina osredotoča na vodenje intervencij na nacionalni ravni. Primera sta projekta NEXES [2], ki se ukvarja predvsem z geografskim usmerjanjem klicev v regijske centre za obveščanje (112), in 6InACtion [3], ki je namenjen vzpostavitvi intervencijskega prenosnega omrežja v primeru obsežnih naravnih in drugih nesreč.

V Sloveniji sicer od leta 2010 deluje sistem Intervencije.net, ki je za razliko od navedenih sistemov namenjen podpori posameznim (lokalnim) intervencijskim službam. Sistem Intervencije.net je bil razvit z namenom izboljšati obveščanje prostovoljnih gasilcev v primeru intervencije, kmalu po njegovi uvedbi pa se je izkazalo, da je uporaben tudi za druge službe, kot so nujna medicinska pomoč, bolnišnice, zdravstveni domovi, gorska in jamarska reševalna služba, kinološka zveza in druge. V zadnjem času so se v sistem začela vključevati tudi nekatera podjetja, ki imajo organizirane lastne reševalne službe. Danes tako v sistemu Intervencije.net deluje že več kot 930 reševalnih služb z več kot 35 tisoč reševalci.

Čeprav je sistem Intervencije.net namenjen predvsem podpori lokalnih intervencijskih skupin, pa omogoča tudi integracijo s sistemi za vodenje intervencij na nacionalni ravni, saj lahko zanje zbira podatke iz lokalnih okolij.



Slika 1. Arhitektura sistema Intervencije.net [4]

Na Sliki 1 je prikazana arhitektura sistema Intervencije.net. Jedrni del sistema sestavljajo

- *podatkovni strežnik*, ki shranjuje prejeta in poslana sporočila, podatke o uporabnikih itd.,
- *spletni strežnik*; namenjen gostovanju spletnih strani in storitev,
- *strežnik PBX*, ki prestreže klice, jih obdela in posreduje podatkovnemu strežniku, ter
- *strežnik za komunikacijo z SMS-centri* mobilnih operaterjev.

Informacije o nesrečah so zbrane preko

- *spletnih storitev*, namenjenih sprejemanju alarmov iz požarnih central in sporočil, ki jih regijski centri za obveščanje sicer pošiljajo pozivnikom članov reševalnih služb,
- *spleta*, z uporabo katerega lahko obveščanje članov sproži končni uporabnik sistema ali pa dispečer preko spletne strani;
- *klica na predpisano stacionarno telefonsko številko*, pri čemer kombinacija kličoče in klicane številke sproži aktivacijo neke enote;
- *protokola ContactID*, ki ga uporabljajo alarmne naprave za javljanje alarmov varnostno-nadzornim centrom;
- *SMS sporočil*.

Sistem posreduje informacije o nesreči članom reševalnih enot ali drugim napravam preko SMS-sporočil in elektronske pošte. Obveščeni se lahko odzove na intervencijsko sporočilo s klicem na določeno telefonsko številko. Pregled odzivov je dostopen preko spletne strani, tako da je vodja intervencije seznanjen, kdo se je namerava udeležiti.

2 Nadgradnja sistema

Čeprav so Intervencije.net že uveljavljen in preverjeno uporaben sistem, ima ta še vedno nekaj omejitev. Ena izmed njih je zmožnost posredovanja informacij samo o tistih nesrečah, o katerih je nekdo že obvestil center za obveščanje, samodejno obveščanje pa povečini ni mogoče.

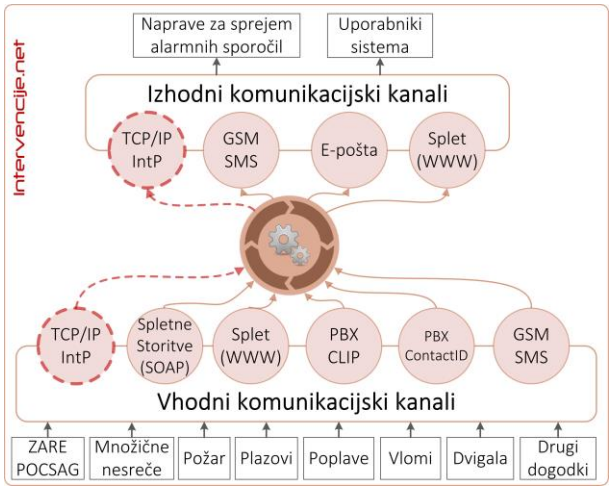
Poleg tega določene reševalne službe niso vključene v sistem tihega alarmiranja (ne uporabljajo pozivnikov), v takih primerih pa je obveščanje lastnega osebja, ko reševalna služba prejme poziv na intervencijo, pogosto zamudno in nepraktično. Takšno obveščanje je trenutno mogoče s prijavo v spletno aplikacijo in vpisom intervencijskega sporočila ali pošiljanjem SMS-sporočila oziroma klicem s točno določene naprave na določeno številko. Obveščanje bi bilo veliko lažje, če bi obstajala naprava, ki bi ob pritisku na tipko samodejno sprožila obveščanje osebja.

Po drugi strani nekatere skupine, na primer gasilci, želijo določeno avtomatizacijo postopkov ob pozivu na intervencijo, kot na primer

- obveščanje članov gasilskega društva z opravljenim strokovnim izpitom iz nudenja prve pomoči, ko je zaznana uporaba samodejnega zunanjega defibrilatorja (AED),
- odpiranje garažnih vrat gasilskega doma in
- obveščanje ljudi v okolici s svetlobnim in zvočnim signalom, da se ti umaknejo gasilcem.

Poleg omenjenega mora biti nadgradnja sistema cenovno sprejemljiva, brez dodatnih stroškov, kot je naročnina pri mobilnem operaterju in tako dosegljiva čim širšemu krogu uporabnikov.

Za izpolnitev omenjenih zahtev smo se odločili za dopolnitev obstoječega sistema z dvosmernim komunikacijskim kanalom, ki bo temeljil na internetnem omrežju in protokolnem skladu TCP/IP (Slika 2, rdeča črtkana črta).



Slika 2. Vhodno-izhodni kanali strežniškega dela sistema

V ta namen smo razvili:

- avtonomno napravo za (samodejno ali ročno) obveščanje strežnika o alarmnih dogodkih in odziv na strežnikove ukaze (za krmiljenje določene naprave, npr. za odpiranje vrat itd.);
- strežnik, ki bo obdeloval podatke z odjemalcev in jih posredoval drugim napravam oziroma uporabnikom sistema Intervencije.net;
- komunikacijski protokol za komunikacijo med odjemalci in strežniki.

Omenjeni elementi so podrobneje opisani v nadaljevanju.

3 Protokol za prenos alarmnih sporočil

Ker zagotavlja zanesljiv prenos, smo protokol za prenos alarmnih sporočil osnovali na transportnem protokolu TCP. Slednjega smo nadgradili z lastnim aplikacijskim protokolom, ki smo ga poimenovali IntP (*Intervencije.net Protocol*). Protokol IntP nadgrajuje protokol TCP z mehanizmi za

- identifikacijo in overjanje odjemalca,
- šifriranje podatkov ter
- nadzor razpoložljivosti komunikacijske poti.

3.1 Identifikacija in overjanje odjemalca

IntP je povezavno naravnan protokol. Vzpostavitev povezave med odjemalcem in strežnikom je namenjena predvsem identifikaciji in overjanju odjemalca.

Prvi korak pri identifikaciji in overjanju odjemalca je preverjanje, če je odjemalčev IP-naslov na seznamu dovoljenih naslovov. Če odjemalčevega naslova ni na omenjenemu seznamu, strežnik odjemalca zavrne.

Identifikacija in overjanje odjemalca se nadaljujeta po različici postopka *izziv in odgovor* (ang. CHAP – Challenge and Response Protocol). Po vzpostavitvi povezave TCP, se odjemalec strežniku predstavi s sporočilom CHAP0, strežnik pa mu odgovori z naključnim besedilom v sporočilu CHAP1. Odjemalec nanj odgovori s sporočilom CHAP2 z izvlečkom prejetega besedila, ki ga ustvari z zgostitveno funkcijo SHA-1 z uporabo lastnega ključa. Strežnik prejeti izvleček primerja z lastno izračunanim izvlečkom. Če se izvlečka ujemata, pošlje odjemalcu sporočilo CHAP3 z vsebino 'OK', v nasprotnem primeru pa 'ERR, wrong HASH'. Ko je odjemalec overjen, strežniški protokolni osebek o tem obvesti svoj uporabniški proces.

3.2 Izmenjava uporabniških sporočil in šifriranje podatkov

Uspelemu overjanju sledi izmenjava podatkovnih sporočil DATA in ukaznih sporočil COMMAND. Obe vrsti sporočil sta sestavljeni iz

- glave, v kateri sta določeni vrsta sporočila ('DA|' za DATA in 'CM|' za COMMAND) in njegova zaporedna številka, ter
- vsebine, ki lahko vključuje
  - alarm, meritev ali ukaz za vklop ali izklop izhoda na odjemalcu pri sporočilih DATA in
  - ukaze za upravljanje strežnika in odjemalca v primeru sporočil COMMAND.

Vsebina obeh vrst sporočil je šifrirana s simetričnim šifrirnim postopkom AES (ang. Advanced Encryption Standard), ki je dovolj enostaven tudi za manj zmogljive mikrokrmilnike.

Obe vrsti sporočil zahtevata potrjevanje sprejema z uporabo pozitivnih potrditev ACK ali negativnih potrditev NACK. Obe potrditveni sporočili vsebujeta glavo ('AY|' oziroma 'AN|') in zaporedno številko sporočila, na katerega se nanašata.

Sporočilo ACK potrjuje, da je sprejemni osebek sporočilo uspešno dešifriral, razbral vsebino, jo obdelal in, na strežniški strani, predal svojemu uporabniškemu procesu. Sporočilo NACK po drugi strani pomeni zavrnitev prejetega sporočila zaradi napake pri enem izmed omenjenih procesov in posredno zahtevo za ponovno pošiljanje sporočila, na katerega se nanaša.

3.3 Nadzor nad komunikacijskim kanalom

Ker protokol TCP nima primernega mehanizma za nadzor razpoložljivosti komunikacijske poti, tega v protokolu IntP zagotavljamo s sporočilom HB (ang. HeartBeat, slov. srčni utrip). Sporočilo HB se lahko prenaša od strežnika k odjemalcu ali v nasprotni smeri. Uspešno prejeto sporočilo HB pomeni prehodno pot in v sprejemnem osebku povzroči ponastavitev časovnikov za preverjanje razpoložljivosti komunikacijske poti do svojega partnerja.

Sporočilo PING in odgovor PONG uporabljamo za ugotavljanje lastnosti prenosne poti in odzivnosti odjemalca oziroma strežnika. S pomočjo omenjenih sporočil lahko strežnik izračuna hitrost odziva posameznega odjemalca in temu primerno prilagaja časovnike za nadzor komunikacijske poti.

4 Odjemalec

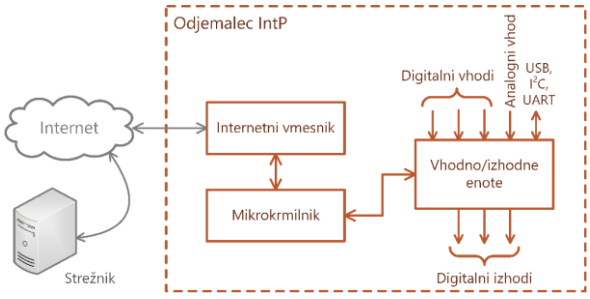
Odjemalec je naprava, ki zaznava izredne dogodke v svoji okolici in jih posreduje strežniku v skladu s protokolom IntP. Tipični dogodki, ki jih odjemalec zaznava, so:

- pritisk na gumb,
- sklenjene vhodne priključne sponke,
- izmerjena vrednost na analognem vhodu,
- podatki s komunikacijskih vodil (npr. I2C ali UART za komunikacijo z drugimi napravami in senzorji, na primer za zaznavanje temperature).

Poleg tega lahko odjemalec alarme in ukaze s strežnika tudi sprejema in nanje reagira. Podatki in ukazi, ki jih odjemalec prejme od strežnika, se lahko posredujejo na odjemalca priključenim napravam preko digitalnih izhodov ali komunikacijskih vodil.

Odjemalec je lahko bistveno preprostejši od strežnika, saj ga lahko izdelamo z uporabo cenovno ugodnega mikrokrmilnika (npr. ESP8266, ESP32 ali ATMEGA328) z dodanim omrežni vmesnikom in vhodno-izhodnimi enotami (Slika 3).

Tudi odjemalčev uporabniški vmesnik je lahko enostaven, saj mora vključevati le gumbe oziroma priključne sponke za proženje alarmov in LED-diode za prikaz stanja odjemalca.



Slika 3. Shema odjemalca IntP

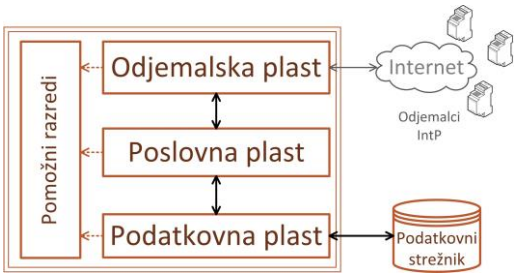


Slika 4. Odjemalec IntP

Na podlagi predstavljene zasnove smo izdelali odjemalca IntP (Slika 4), ki temelji na modulu WEMOS D1 mini [5]. Ta med drugim vključuje mikrokrmilnik ESP8266 z vgrajenim čipom za povezljivost z Wi-Fi omrežji in komunikacijo po protokolnem skladu TCP/IP, anteno WiFi, napetostni stabilizator, 11 vhodno-izhodnih povezav, analogno-digitalni pretvornik ter povezavo USB.

5 Strežnik

IntP strežnik je strežniški proces, ki komunicira z odjemalci: odziva se na njihove zahteve, obdeluje prejete podatke in jih posreduje drugim napravam oziroma komponentam sistema. Zaradi boljše preglednosti in preprostejšega vzdrževanja smo strežniški del izdelali v skladu s tri-nivojsko arhitekturo. Strežniški del tako sestavljajo odjemalska, poslovna in podatkovna plast (Slika 5).



Slika 5. Arhitektura strežnika IntP

Odjemalska plast predstavlja izvedbo protokola IntP na strežniku. Naloge plasti vključujejo

- komunikacijo z odjemalci z uporabo vseh mehanizmov, ki jih predvideva protokol IntP (razdelek 3),
- predajo sporočil iz odjemalcev poslovni plasti,
- sprejem sporočil s poslovne plasti in njihovo pošiljanje odjemalcem.

Poslovna plast je uporabnik protokola IntP, ki izvaja vsebinsko komunikacijo z odjemalci, obenem pa uporablja storitve podatkovne plasti. Naloge poslovne plasti vključujejo

- vodenje zbirke aktivnih odjemalcev,
- sprejem in potrditev sprejema sporočil z odjemalske plasti ter posredovanje prejetih sporočil podatkovni plasti,
- sprejem sporočil iz podatkovne plasti in njihovo posredovanje odjemalski plasti in
- odziv na izredne dogodke, kot je na primer neprehodna povezava do nekega odjemalca.

Naloga *podatkovne plasti* je zagotoviti poenostavljen in centraliziran dostop do podatkov, ki so zapisani v relacijski podatkovni zbirki sistema Intervencije.net.

Strežnik IntP temelji na Microsoftovih tehnologijah .Net Framework (aplikacijski del) in SQL Server (podatkovni del) in se izvaja kot storitev na operacijskem sistemu Windows.

6 Zaključek

Predstavljena nadgradnja sistema Intervencije.net je trenutno v postopku preizkušanja. Prvo fazo preizkusa smo opravili s tremi odjemalci IntP, ki so se povezovali z enim strežnikom. Namen preizkusa v tej fazi je bil predvsem odkrivanje in odpravljanje napak.

Trenutno poteka druga faza preizkušanja, namen katere je ugotoviti, če naprave zadoščajo potrebam uporabnikov, zato preizkusi potekajo v gasilskih domovih. V preizkuse je vključenih deset odjemalcev in dva strežnika.

Opravljenemu preizkušanju bo sledila vpeljava odjemalcev IntP v stvarna okolja. Najprej načrtujemo nameščanje naprav v gasilskih in zdravstvenih domovih, kasneje pa jih bomo ponudili tudi posameznim članom reševalnih služb za priključitev v njihovih stanovanjih in hišah.

Čeprav v okviru dosedanjega preizkušanja nismo odkrili večjih pomanjkljivosti, pa je odjemalce IntP in strežniški del sistema mogoče še izboljšati. Načrtovane so predvsem naslednje izboljšave:

- samodejno posodabljanje programske opreme v odjemalcih, kar je predpogoj za njihovo vpeljavo v stvarna okolja,
- dodajanje rezervnega komunikacijskega kanala, kot so sporočila SMS, če pride do prekinitve internetne povezave,
- izdelava grafičnega uporabniškega vmesnika za lažje upravljanje sistema Intervencije.net,
- vgradnja LED-prikazovalnika v napravo IntP namesto statusnih LED-diod, ki bo omogočal dodatne nastavitve naprave,
- uporaba zmogljivejšega mikrokrmilnika ESP32,
- povečanje števila relejskih izhodov in digitalnih vhodov,
- izdelava spletnega vmesnika, s pomočjo katerega bo lahko uporabnik sam upravljal določene nastavitve naprave.

Ostale funkcionalnosti bomo dodajali postopoma, glede na potrebe prihodnjih uporabnikov.

Literatura

[1] ETSC - Europan Transport Safety Council: New Pan-European Emergency Call System, sep. 2013, http://www.roadsafetyobservatory.com

[2] NEXES - NEXt generation Emergency Services, http://nexes.eu/

[3] Gen 6, 6 in Action – Smart communications solution in emergency situations, http://www.6inaction.net/

[4] Aldia, d.o.o.: Predstavitev sistema Intervencije.net., sep. 2015, https://www.intervencije.net

[5] WEMOS Electronics: D1 mini [WEMOS Electronics], https://wiki.wemos.cc/products:d1:d1_mini

Overjanje uporabnikov pametnega doma prek biometričnih značilnosti na pametnih telefonih

Miha Marolt, Mark Umberger, Iztok Humar

Fakulteta za elektrotehniko, Univerza v Ljubljani
iztok.humar@fe.uni-lj.si

Smart-home user authentication through biometric parameter on smartphones

Abstract. *This paper presents possibilities for biometrical authentication methods of users on smartphones. A practical example provides its usage for a smart-home application. We first provide an overview of smart-home functionality with an emphasize of functions allowing entering of a building, due to possible security vulnerabilities, such as unauthorized access. The second part introduces biometrical authentication methods, offered by modern smartphones, such as fingerprint scanner, iris scanner and face recognition. These three methods are compared for a security level, duration of authentication and reliance. A practical example demonstrates the replacement of current username and password login for a smart-home application with a biometrical authentication method based on fingerprint scanning. It has been proven, that fingerprint scanning type of authentication is adequate for this type of smartphone applications, because it allows quick authentication and does not compromise safety level.*

1 Uvod

Poleg upravljanja in nadzora razsvetljave in temperature pametni domovi danes omogočajo še dodatno funkcionalnost za zagotavljanje varnosti, kot je npr. nadzor kakovosti zraka, zaznavanje izliva vode in uhajanja plina, video nadzor, obveščanje o nenavadnih dogodkih in raznovrstne protivlomne alarme. To narekuje potrebo po centralnem nadzoru nad funkcijami pametnega doma. Sprava je bila možnost dostopa do centralnega nadzora vzpostavljena prek osebnih računalnikov, kasneje pa prek za to posebej prilagojenih zaslonov na dotik oziroma stikalnih panelov, vgrajenih v stavbi. Danes pa so se ob porastu uporabe pametnih mobilnih telefonov uveljavile aplikacije, ki omogočajo nadzor nad funkcijami stavbe.

Pogosto med funkcije pametnega doma spada tudi možnost oddaljenega odklepanja in zaklepanja vhodnih vrat s pomočjo magnetne ključavnice. Če že ne zaradi drugih funkcij, pa zaradi možnosti odklepanja vhodnih vrat uporabnika pogosto skrbi, kdo lahko dostopa do centralnega nadzora oziroma kakšna je njegova varnost. Ta mora biti vsaj ekvivalentna varnosti klasičnega pristopa - torej ključavnice in ključa. To pa vzbudi vprašanje učinkovitosti overjanja uporabnika v sistem centralnega nadzora. Do nedavnega je bila najbolj razširjena rešitev z uporabniškim imenom in geslom, danes pa na pametnih telefonih v ospredje vse bolj prihajajo biometrični načini, izmed katerih se je najbolj uveljavilo prepoznavanje prstnega odtisa. Povod za to je izboljšana zanesljivost in hitrost delovanja strojne opreme, ki omogoča biometrične načine overjanja in posledično vse pogostejšo vgradnjo tovrstne strojne opreme v pametne telefone vseh cenovnih razredov. Poleg čitalnika prstnega odtisa v uporabo prodira še prepoznavanje očesne šarenice in prepoznavanje obraza.

Biometrični načini overjanja uporabnika so sicer že prisotni v klasičnih stavbah, predvsem v industrijskih objektih in poslovnih stavbah, kjer so bili sestavni del tehničnega varovanja. V zadnjem obdobju pa so se uveljavili tudi v stanovanjskih objektih. Najbolj razširjena oblika je čitalnik prstnega odtisa, vgrajen v vhodna vrata. Vendar pa se ob zavedanju, da na trg končno masovno prodirajo pametni telefoni z že vgrajenim čitalnikom prstnega odtisa, pojavi vprašanje o smiselnosti njegove vgradnje tudi na vhodna vrata stanovanjskega objekta, saj lahko strošek njegove vgradnje celo preseže strošek nakupa pametnega telefona, ki že vsebuje čitalnik prstnega odtisa. S tem pa se tudi oddaljujemo od željene energetske učinkovitosti, saj beležimo prisotnost dveh naprav z isto funkcijo. Tako lahko predvidevamo, da se bo v prihodnosti čitalnik prstnega odtisa z vhodnih vrat umaknil na pametne telefone, tako kot so se v preteklosti stikalni paneli in na dotik občutljivi zasloni za upravljanje s funkcijami pametnega doma s sten umaknili na aplikacije pametnih telefonov oziroma tabličnih računalnikov.

Poleg tega pa se čitalnik prstnega odtisa kot način overjanja uporabnika uveljavlja tudi v drugih strokah, ki so morda še bolj občutljive glede varnostnih zahtev. Na tem mestu velja izpostaviti spletno bančništvo in spletne trgovine, pri katerih transakcije že lahko potrjujemo prav s prstnimi odtisom.

Nesporno lahko torej ugotavljamo, da se biometrični načini overjanja uveljavljajo tudi na področju pametnih domov, zato v nadaljevanju sledi pregled oblik biometričnih načinov overjanja in ustreznosti implementacije tovrstne tehnologije v aplikacije, ki zagotavljajo centralni nadzor nad stavbo. Kasneje pa demonstriramo implementacijo biometričnega načina overjanja uporabnika prek prstnega odtisa v konkreten primer aplikacije za pametni dom. Na koncu izpostavimo možnosti za nadaljni razvoj oziroma možnosti nadgradnje in oceno zanesljivosti biometričnih načinov overjanja uporabnika pri tovrstnih sistemih.

2 Biometrično overjanje pri telefonih

Eno izmed področij, kjer so se biometrični načini overjanja uporabnika v zadnjem obdobju močno uveljavili, je področje pametnih telefonov. Ti namreč omogočajo dostop do najrazličnejših vsebin, izmed katerih so nekatere z varnostnega vidika še posebej občutljive, zato je potreba po varnosti naprave precej visoka. Obenem pa se je potrebno zavedati, da uporabnik odklene svoj telefon tudi več stokrat na dan, zato je ključno, da je identifikacija uporabnika izvedena kar najhitreje. Biometrični načini tako izpolnjujejo obe zahtevi. Za branje prstnega odtisa je mogoče trditi, da se je že popolnoma uveljavilo, saj ima danes že več kot dve tretjini novih telefonov vgrajen čitalec prstnega odtisa. Poleg prstnega odtisa pa se pri najdražjih telefonih kot biometrična načina overjanja uporabnika pojavljata tudi prepoznavanje šarenice in prepoznavanje obraza.

2.1 Pregled načinov biometričnega overjanja

Predstavimo biometrične načine overjanja za tiste človeške lastnosti, ki jih pametni telefoni najpogosteje koristijo in na kratko razložimo postopke zajema vzorca izbrane lastnosti.

2.1.1 Prstni odtis

Prstni odtis ustvari koža blazinice človeškega prsta. Na svetu do danes še niso našli dveh ljudi, ki bi imela identični prstni odtis, niti pri enojajčnih dvojčkih. Prstni odtis nastane še pred rojstvom in se s staranjem ne spreminja, ampak raste, ne da bi spreminjal svojo obliko. Pri odraslem človeku ostane v svoji velikosti nespremenjen. Tudi v primeru poškodbe kože blazinice, se prstni odtis obnovi v prvotno obliko. Zato se ta človeška lastnost lahko uporablja za dokazovanje istovetnosti osebe in je trenutno ena izmed najzanesljivejših načinov preverjanja identitete oseb.

Gledano s fiziološkega vidika je prstni odtis struktura grebenov s porami, ki ležijo neposredno pod kožo na ožilju, grebene pa delijo doline. Prstnega odtisa pa ne določa le oblika, pač pa tudi specifične električne in toplotne značilnosti kože. Za zajem prstnega odtisa tako lahko uporabljamo posamično lastnost prstnega odtisa ali pa kombinacijo večih.

Vzorec prstnega odtisa pri postopku overjanja služi kot ključ, s katerim se šifrirajo podatki o uporabniku, preden se shranijo v napravo. Tako shranjene podatke je nemogoče dešifrirati brez vzorca prstnega odtisa.

Za zajem prstnega odtisa mora telefon imeti posebej v ta namen prilagojeno strojno opremo - čitalnik prstnega odtisa. Prvi so bili večinoma optični, vendar so jih kmalu nadomestili učinkovitejši kapacitivni. Danes pa je prisoten še tretji tip, ki prstni odtis prepozna s pomočjo ultrazvočnih valov.

Optični čitalnik

Optični čitalnik je najstarejša strojna oprema za zajemanje in primerjanje prstnih odtisov. Princip delovanja sloni na zajemu digitalne slike, preko za to specializirane digitalne kamere (iz CCD ali CMOS senzorjev, ki na izhodih vračajo napetostne impulze). Poleg tega potrebujemo še vir svetlobe, ki osvetli površino, na katero prislonimo prst in lečo, ki odbite svetlobne žarke usmeri na polje.

Optični čitalnik prstnih odtisov je cenovno ugoden, vendar ima nekaj slabosti. V primeru, da je površina, na katero polagamo prst, umazana, to vpliva na zajeto sliko, ki tako odstopa od referenčne in identifikacija je neuspešna. Enako se zgodi, če ima uporabnik poškodovano kožo na blazinici prsta. Druga slabost pa je, da preverja le dvodimenzionalno sliko, kar pomeni, da bi se lahko na primer identificirali tudi s pomočjo dovolj kvalitetne slike prstnega odtisa, natisnjene na papir, kar omogoča dokaj preprosto zlorabo. Optični čitalnik tako ne zagotavlja visoke stopnje zaščite in se zato umika iz uporabe.

Kapacitivni čitalnik

Danes najpogosteje uporabljen tip čitalnikov prstnih odtisov je kapacitivni čitalnik. Za svoje delovanje uporablja princip kapacitivnosti kondenzatorja. Pod površino, kamor prislonimo prst, se nahaja množica majhnih kondenzatorjev. Z večanjem njihovega števila se izboljšuje ločljivost senzorja, kar povečuje stopnjo varnosti, obenem pa so čitalniki z večjim številom kondenzatorjev dražji, zato so v preteklosti vpeljali drsne čitalnike, ki so za zbiranje informacij potrebovali manj kondenzatorjev, saj uporabnik prsta ni prislonil na površino, pač pa je z njim drsel preko ozkega pasu kondenzatorjev. Kljub sposobnosti hitrega posodabljanja izhodnih vrednosti kondenzatorjev je bil odstotek uspešnih identifikacij dokaj nizek, zato so tovrstne kapacitivne čitalnike umaknili s trga.

Kapacitivni čitalnik prstnega odtisa je sicer zelo zanesljiv, stopnja varnosti je visoka. Napake pri delovanju bi se lahko pojavile v območjih z močnimi radiofekvenčnimi motnjami.

Ultrazvočni čitalnik

Ultrazvočni čitalnik prstnega odtisa pri svojem delovanju posnema medicinsko tehniko. Sliko prstnega odtisa tako generira s pomočjo visokofrekvenčnih zvočnih valov, ki potujejo od oddajnika do kože, kjer se odbijejo nazaj proti sprejemniku. Medtem se beleži čas, ki je potreben za celotno pot valov, kar nam ob že poznani hitrosti valov omogoča izračun razdalje.

Za delovanje čitalnika potrebujemo oddajnik visokofrekvenčnih zvočnih valov in sprejemnik. Oddajnik električno energijo pretvarja v zvočne valove, ki se nahajajo v ultrazvočnem območju nad 18 kHz. Sprejemnik pa sprejema signale, ki se od kože odbijajo nazaj in jih ponovno pretvori v električni signal. Za sprejemnik se ne uporablja mikrofon, pač pa tipalo, ki beleži mehanske tresljaje na površini senzorja. Običajno je tipalo narejeno iz piezoelektričnih materialov, ki ustvarjajo napetost, ko nanje pritiska mehanska sila, ki jo povzročajo odbiti ultrazvočni valovi.

Iz povedanega lahko zaključimo, da je ultrazvočni senzor najučinkovitejši čitalnik prstnega odtisa. Za razliko od kapacitivnega in optičnega, ki generirata le 2D sliko prstnega odtisa, ultrazvočni generira sliko v treh dimenzijah, kar je mnogo težje zlorabiti. Zato lahko trdimo, da je stopnja varnosti višja kot pri ostalih dveh različicah. Poleg tega na delovanje ultrazvočnega čitalnika ne vpliva umazanija na površini senzorja ali kože. Prav tako senzor deluje ne glede na to, iz

kakšnega materiala je narejena njegova zgornja površina (plastika, aluminij, steklo), kar omogoča bolj diskretno umestitev tovrstne strojne opreme v napravo.

2.1.2 Očesna šarenica

Očesna šarenica je pigmentiran obroč vezivnožilnega tkiva in mišičnih vlaken, ki se nahaja v sprednjem segmentu očesa. Poleg šarenice so v tem segmentu še zenica, ciliarnik in leča. Naloga šarenice je, da z ustreznim širjenjem in oženjem zenice zagotavlja čim bolj konstanten nivo svetlobe v očesu.

Z vidika biometrije je zanimiva zaradi edinstvenosti svojega vzorca in barve. Šarenica se do popolnosti izrazi že v zgodnem otroštvu in ostane nespremenjena skozi celo življenje. Očesna šarenica kot biometrični način overjanja se je sprva uveljavila v farmacevtski industriji, kjer so dostopi do laboratorijev kar najostreje omejeni. Kasneje se je ta način pojavil še na letališčih in nekaterih mejnih kontrolah, v zadnjem času pa se podobna tehnologija pojavlja tudi na pametnih mobilnih telefonih in tabličnih računalnikih.

Sam postopek šifriranja je enak kot pri prstnem odtisu, le da se tokrat za ključ šifriranja podatkov o uporabniku (na primer uporabniško ime in geslo) uporablja očesna šarenica in ne prstni odtis, podatki pa se nato v šifrirani obliki shranijo v napravo. Hitrost overjanja uporabnika je s pomočjo šarenice še višja, saj uporabnik le pogleda telefon in postopek overjanja steče. Za zajem slike očesne šarenice mora telefon vsebovati dve namenski komponenti, in sicer sta to kamera, ki uporablja poseben slikovni filter za zajem slik očesne šarenice ob infrardeči svetlobi, ter IR-LED dioda. Obe komponenti sta nameščeni na sprednji strani naprave, saj to napravi omogoča, da v slabih svetlobnih pogojih izkorišča svetlobo, ki jo oddaja zaslon naprave. Zajem slike pod infrardečo svetlobo se izvaja zato, ker se na ta način izničijo vplivi zunanje svetlobe in motečih odsevov. Takšna svetloba ima nizko refleksijo, zato vrača jasne vzorce v vseh pogojih. Seveda mora IR-LED dioda ustrezati nekaterim standardom, da ne pride do poškodbe oči, poleg tega pa ima večina naprav vgrajene posebne senzorje, ki diodo izklopijo, če zaznajo, da je naprava preblizu uporabnikovega očesa. V napravo je mogoče shraniti le en vzorec šarenice, s čimer se nivo varnosti, v primerjavi s čitalnikom prstnih odtisov, kamor je mogoče shraniti več prstnih odtisov, še izboljša. Poleg tega pa je tudi število neuspešnih poizkusov overjanja manjše.

2.1.3 Prepoznavanje obraza

Tretji in tudi najnovejši način biometričnega overjanja pri pametnih telefonih pa je prepoznavanje obraza. Tudi v tem primeru se iz slike obraza uporabnika generira simetrični ključ, s katerim se šifrirajo podatki. Zajeta slika se analizira s številnimi algoritmi, ki jih lahko razdelimo na geometrijske in fotometrične. Geometrijski algoritmi obraz primerjajo na osnovi razdalj, velikosti in položaja oči, nosu, ličnic in čeljusti. Fotometrični algoritmi pa sliko pretvarjajo v vrednosti in istovetnost obraza določijo glede na odstopanje vrednosti med sliko v bazi in sliko, zajeto s kamero.

Prepoznavanje obraza se izvede še hitreje kot metoda očesne šarenice, vendar pa je na začetku uporabe število zavrnitev dokaj visoko, ker pa je sistem samoučeč, se število zavrnitev v daljšem intervalu uporabe zmanjšuje. Trenutno tovrsten način biometričnega overjanja uporabnika na področju pametnih telefonov uporablja le malo proizvajalcev.

2.2 Primerjava načinov overjanja uporabnika

Izpostavljene so bile prednosti in slabosti posameznih biometričnih načinov overjanja uporabnika na pametnih telefonih, na podlagi česar lahko izberemo naši aplikaciji najustreznejšo obliko biometričnega overjanja. V tem trenutku za skoraj enako varni in zanesljivi metodi biometričnega overjanja veljata primerjava prstnega odtisa in primerjava vzorca očesne šarenice.

Primerjavo prstnega odtisa lahko sem uvrščamo samo, če je za njegov zajem uporabljen ustrezen čitalnik, torej ultrazvočni ali kapacitivni. Proizvajalec Apple za svojo obliko biometričnega overjanja uporabnika z vzorcem prstnega odtisa, imenovano touch ID, navaja možnost ujemanja vzorca prstnega odtisa dveh uporabnikov 1:50000, medtem ko štirimestna PIN koda ponuja le deset tisoč kombinacij, kar pomeni, da je možnost naključnega poizkušanja 1:10000. S te perspektive je overjanje z vzorcem prstnega odtisa petkrat varnejše, vendar se moramo zavedati, da je prstni odtis možno tudi rekonstruirati.

Primerjava očesne šarenice v tem trenutku velja za najvarnejšo, njena rekonstrukcija skoraj ni mogoča. Hitrosti overjanja so vsaj enake kot pri primerjavi vzorca prstnega odtisa, v nekaterih primerih pa celo višje. Zajem vzorca je možen tudi v izredno slabih svetlobnih pogojih, težava nastane le ob direktni sončni svetlobi, saj takrat uporabnik zelo težko odpre oči do te mere, da bi čitalnik lahko zajel vzorec zenice. Edina vidnejša slabost je v tem, da je trenutno na trgu zelo malo naprav z ustrezno strojno opremo, zato se pojavlja vprašanje o smiselnosti implementacije tovrstnega načina biometričnega overjanja uporabnika v aplikacije, vendar se bo število naprav, ki omogočajo tovrstno overjanje v bodoče zagotovo povečalo.

Prepoznavanje obraza je hitro in ob ustreznih svetlobnih pogojih (v temi ne deluje) tudi zanesljivo, vendar pa trenutno še ne dosega najvišje stopnje varnosti, kar na nek način priznava tudi edini ponudnik tega načina overjanja uporabnika, Samsung, ki prepoznavanja obraza ni ponudil kot možnost vpisa v svojo aplikacijo Samsung Pay.

3 Praktični del

Namen praktičnega dela je nadgraditi konkretno aplikacijo za pametni dom (ENTIAliving), ki je doslej uporabljala zgolj klasičen postopek overjanja (z uporabniškim imenom in geslom) ter skrajšan postopek overjanja (s štirimestnim PIN) z biometričnim načinom overjanja uporabnika pri vpisu v aplikacijo.

Najprej predstavimo konkretno implementacijo nadgradje z biometričnim načinom overjanja uporabnika. Nato predstavimo, kako na podlagi nje

poteka postopek overjanja oziroma kako si sledijo koraki v primeru prvega vpisa ali nadaljnjih overjanj.

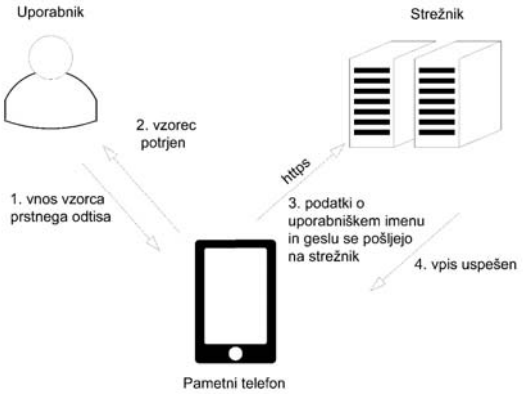
Zasnova aplikacije ENTIAliving [4] za operacijski sistem Android temelji na spletni strani, zato je ključen ukaz v programski kodi aplikacije WebView, ki spletno stran prilagaja mobilnemu telefonu. Za obdelavo prstnega odtisa smo uvozili ustrezno knjižnico in spisati dve funkciji v programskem jeziku Java. Prva funkcija omogoča branje in shranjevanje podatkov, ki jih je uporabnik vnesel v prijavni obrazec aplikacije. Druga funkcija shranjene podatke vpiše nazaj v polja uporabniškega imena in gesla.

Implementirali smo postopek generiranja simetričnega ključa na podlagi prstnega odtisa, s katerim se šifrirata uporabniško ime in geslo, preden se shranita v napravo.

V nadaljevanju iz uporabniškega vidika predstavimo še postopke overjanja pri prvem in nadaljnjih vpisih.

Pri postopku prvega vpisa uporabnika v aplikacijo, pogovorno okno od njega zahteva uporabniško ime in geslo, nato pa steče vpis na strežnik. Če sta vpisana uporabniško ime in geslo pravilna, pogovorno okno od uporabnika zahteva še vzorec prstnega odtisa, prebrani avtentikacijski podatki se zašifrirajo s pomočjo simetričnega ključa na podlagi vzorca prstnega odtisa uporabnika in shranijo v napravo.

V primeru nadaljnjih vpisov, so podatki že shranjeni v napravi, in sicer v šifrirani obliki. Pogovorno okno od uporabnika najprej zahteva vzorec prstnega odtisa, ki predstavlja simetrični šifrirni ključ. V primeru, da je vzorec prstnega odtisa pravilen, se shranjeni avtentikacijski podatki dešifrirajo in steče postopek overjanja na strežniku; overjanje je uspešno pod pogojem, da so shranjeni avtentikacijski podatki pravilni. Postopek je prikazan na sliki 1. Glede na zgoraj predstavljeni postopek prvega vpisa, napačnih podatkov preko telefona ni mogoče shraniti, saj je uspešno overjanje uporabnika na strežnik pogoj za shrambo podatkov. Lahko pa pride do spremembe avtentikacijskih podatkov npr. zaradi spremembe gesla za pametni dom na osebnem računalniku.



Slika 1. Overjanje uporabnika s podatki predhodno shranjenimi v napravi

Vzorec prstnega odtisa, ki se uporablja kot simetrični ključ za šifriranje podatkov, ki jih prebere funkcija, preden se podatki shranijo v napravo, je vezan na telefon. Vzorec je lokalno shranjen na telefonu in se uporablja samo za to, da uporabnik lahko dostopa do svojega gesla in se overi na sistemu Entia. Vzorec prstnega odtisa torej ni neposredno povezan z uporabniškim računom, kar tudi pomeni, da podjetje Entia ne hrani vzorcev prstnih odtisov svojih uporabnikov in tako ne more priti v navzkrižje z zakonom o varstvu osebnih podatkov.

4 Zaključek

Število aplikacij, pri katerih je potrebna hitrost in visoka stopnja varnosti pri overjanju uporabnika, v zadnjem obdobju močno narašča. Biometrični načini overjanja uporabnika zagotavljajo prav to dvoje, zato je razumljivo, da je njihova uporaba vse bolj množična. Trenutno najbolj razširjen način biometričnega overjanja uporabnika je vzorec prstnega odtisa, saj ima največ naprav temu načinu ustrezno strojno opremo, vendar pa se bo v prihodnosti močno uveljavilo prepoznavanje obraza, ki je trenutno še zelo redko v uporabi. Tako kot pri ostalih biometričnih načinih je tudi pri prepoznavanju obraza glavna prednost hitrost in zanesljivost, poleg tega pa še enostavnost uporabe z uporabniškega vidika, saj ni treba polagati prstov na senzorje, pač pa samo pogledamo prednjo stran naprave in postopek overjanja steče. Poleg tega lahko tovrsten način overjanja uporabljamo, ne da bi se uporabniki tega zavedali.

Zato se prepoznavanje obraza uveljavlja povsod tam, kjer je prisotno večje število ljudi (navajamo nekaj konkretnih primerov): nakup vozovnice na železniških postajah, plačevanje v restavraciji s hitro hrano, e-bančništvo, razvijajo pa se tudi »pametni ključi« za nadzor nad vozilom.

V obstoječem prispevku smo predstavili možnost overjanja na podlagi prstnega odtisa pri aplikaciji za upravljanje pametnega doma ter prikazali pilotsko implementacijo tovrstne rešitve. Ker bo v prihodnje vse več pametnih mobilnih telefonov imelo vgrajeno tudi ustrezno strojno opremo za prepoznavanje obraza in očesne šarenice, je zato smiselno razmišljati tudi o nadgradnji aplikacije v tej smeri.

Literatura

[1] R. Triggs »How fingerprint scanners work: optical, capacitive, and ultrasonic variants explained« Dosegljivo: http://www.androidauthority.com/how-fingerprint-scanners-work-670934/ [Dostopano: 10.8.2017]

[2] »Keeping an Eye on Security: The Iris Scanner of the Galaxy Note7« Dosegljivo: https://news.samsung.com-/global/in-depth-look-keeping-an-eye-on-security-the-iris-scanner-of-the-galaxy-note7 [Dostopano 9.9.2017]

[3] M. Marolt, Overjanje uporabnikov pametnega doma preko čitalnikov biometričnih značilnosti na pametnih telefonih. Diplomsko delo, 2017, FE

[4] ENTIAliving. http://dostop.entia.si [Dostopano 4.6.2018]

Bluetooth based sensor networks for wireless EEG monitoring

Tamara Jakovljevic[1], Tomaz Javornik[2], Gregor Papa[3]

[1]*Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana*
[2]*Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana*

E-pošta: tamara.jakovljevic@hotmail.com

Abstract*—Wireless sensor networking has a crucial role to play in the progressive healthcare industry – the health monitoring systems in particular can gain a lot from this type of technology. The focus of this paper is to discuss the innovative wireless sensor network technology applied to electroencephalogram (EEG) devices alongside the paradigm used for mobile sensor networks in healthcare monitoring. Another focus of this paper is to present the technological challenges and benefits of the application of wireless EEG devices. The foundation of this method is based on the use of EEG sensors that are both miniaturized and wireless. As a result, we are able to perform ambulatory EEG recordings, which opens the door to real-time seizure detection during everyday activities. This in turn leads to a completely new stage in both medicine and research systems in general.*

1 Introduction

Wireless sensor networks are a technology that is gaining a lot of attention lately. The broad spectrum of these networks of small devices equipped with sensors, microprocessor and wireless communication interfaces is already finding use in military applications, environmental monitoring as well as healthcare [1]. One of the main focuses in this field is the introduction of smart sensor nodes. These small, low-cost and low-power devices are the product of MEMS (micro-electrical-mechanical systems) technology, combined with electric components and wireless communication. They are capable of performing various tasks related to sensoring, data processing and wireless communication [2]. When put together, these devices can make up a well-developed platform that enables many uses in terms of security systems, health monitoring, detection of aerial and hydro chemical agents etc. Medical monitoring as well as data access – are the benefits of the wireless sensor network. With it, one can determine any emergency conditions of patients. All one has to apply transducers embedded in clothing [3] or a network of body sensors [4].

Electroencephalography (EEG) is a very precise and insightful method to measure brain activity and function when exposed to certain factors. And since it is non-invasive, it is used today to diagnose and treat certain clinical disorders which can cause and present themselves as abnormalities in EEG readings. The most basic examples include epilepsy and encephalopathy, but it can be used to diagnose other neurological diseases [5].

The communication in sensor networks can be applied using various technologies such as infrared, UWB, Bluetooth, WiFi or even any legacy short range RF technology.

2 Bluetooth based sensor network

Bluetooth found quite the application in mobile devices such as mobile phones, laptops, cameras PDAs etc. So, the original idea to remove the need for cables was resolved with the introduction of low-cost, short-range wireless technology that not only has a small footprint but also minimal power consumption, making it perfect for mobile devices.

The concept has since spread and developed from point-to-point to becoming a network technology. Sensor network related Bluetooth operates in the 2.4 GHz frequency band and uses frequency hopping spread spectrum technique. There are 79 channels, each 1MHz wide, available for hopping [6].

Due to the low energy consumption and ease of use Bluetooth Low Energy (BLE) is a communication technology developed for low power consumption and low cost applications. The BLE technology is used in various devices, ranging from health care, fitness, home automation, toys, transportation and industry. Because of its low power consumption design, it is especially suitable for coin cell battery communication devices like wireless sensors [7].

In order to be able to communicate with other devices, a Bluetooth device has to be member of a piconet. A piconet represents a combination of 8 devices that hop together on the same frequency. Out of those 8 connected devices, one is the master device, whereas the remaining 7 are so-called slave devices. To establish a piconet, the device has to discover other Bluetooth devices in range using the Inquiry option (on other devices performing at the same time). Once they've found one another, the Inquiry option makes sure that both devices are on the same frequency at the same time. Once it comes to that, the Bluetooth address and the clock of the master device are exchanged in order to establish a working connection. If it happens that more than 7 devices try to communicate with each other, there is one of the ways to make this happen. The first option relies on the so-called Park mode of the device (disassociates from the piconet, but still maintains timing synchronization with it). Sniff, Hold and Park – these are the 3 low-power modes defined by Bluetooth.

The second option is called Scatternet – a network of several piconets, where devices can connect to multiple

piconets [8]. These devices can either be slaves in all piconets or master in one and slave in others [9].
Data from distant nodes cannot be transferred using a direct connection, the basis of multi-hop communication mechanism, according to [7] there is ability of each node to temporary assume the master role and to acquire the data from not already processed nodes, and relay acquired data to its master node along with its sensor data.

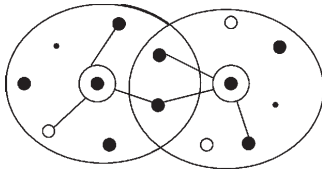


Fig. 1. Scatternet

The main aim of Bluetooth is the wireless communication between different devices that leads unification and harmony.

3 EEG signal transmission

EEG was developed in 1875 and marks the electric activity of various parts of the brain with the use of electrodes placed on the head. In addition, discoveries were made for measuring the magnetic field of the heart and brain, also known as magnetocardiography (MCG) and magnetoencephalography (MEG).

3.1 Mechanism

EEG is based on electrophysiological monitoring of electrical activity of the brain. It is graphic display of a difference in voltages from two sites of brain function recorded over time. The brain's electrical charge is maintained by billions of neurons. The synaptic potentials are the most important source of the extracellular current flow that produces potentials in the EEG. Most routine EEGs recorded at the surface of the scalp represent pooled electrical activity generated by large numbers of neurons [10].

3.2 Measurements

Measurements are conducted by setting up a system of electrodes on the skin surface of the head with previous preparations by adding electrode gel that is used for acceleration of electrode contact with the skin.
The distance between the electrodes is very important because it directly influences the sensitivity of the measurement – closer electrodes provide more accurate results [11].

3.3 Information process

Electroencephalography is the simple method of recording and measuring electrical activity of the brain. The analysis of EEG signal plays very important parts for detecting and predicting various brain diseases. The EEG signals are first amplified, filtered and then converted to digital output. This will be used by digital system to record the information.

After the signal is recorded, next phase of measurement is to differentiate the signals by using mathematical models that classifies the signals from different parts of the brain and to use that data for a specific research.

4 WSN AND EEG MONITORING

In order to monitor both the physical and physiological condition of patients, one can find various sensors, upgraded with medical equipment, installed at homes, in clinics and hospitals [12,13,14].
Previous user interfaces were isolated from these medical sensors. However, with the introduction of wireless short-range connections (such as Bluetooth and Zigbee) to close-vicinity PDAs, Smartphones or PCs, these obstacles have been overcome. The embedded platform that present technology offers combines data processing, storing, sensing and computing into one integrated unit.
According to IEEE, Wireless personal area networks (WPANs) are used to convey information over short distances among a private, intimate group of participant devices. IEEE 802.15.1 also defines a Bluetooth as used WPAN technology. While many sensors connect to controllers and processing stations directly (e.g., using local area networks), an increasing number of sensors communicate the collected data wirelessly to a centralized processing station. Therefore, a wireless sensor has not only a sensing component, but also on-board processing, communication, and storage capabilities. When many sensors cooperatively monitor large physical environments, they form a wireless sensor network (WSN) [15].
The term Wireless Body Sensor Network (WBSN) it is also common in use. WBSN can be defined as a self-directed, self-sufficient and self-governing i.e. autonomous system which is used to monitor the daily life activities of a person [16] [17].
The EEG monitoring wireless system is comprised of multiple brain electrodes that have the capability to conduct and carry electrical signals to and from the brain.

4.1 Case Study

Mbrain Train is a Belgrade-based start-up company which developed SMARTING - a small, simple and mobile EEG device.
The appeal of this EEG device is found in the real-time brain activity monitoring and high-quality recordings in real life conditions. SMARTING offers superior data quality and high time-precision, no matter if paired with a smartphone or PC.
SMARTING can be used for cognitive group studies because it supports synchronization with other sensors and simultaneous multi-amplifier streaming via labstreaming layer.
The advantages of this device are: Lighter than 60g; Sampling 250-500Hz; Recording cap have 24ch; Bluetooth range 10m; Up to 5 hours recordings.

We analyze two studies in which the SMARTING system was used:
1. EEG Recording and Online Signal Processing on Android: A Multiapp Framework for Brain-Computer Interfaces on Smartphone [18].
2. Cortical Sensitivity to Guitar Note Patterns: EEG Entrainment to Repetition and Key [19].

1) Android-based processing

In the first one, authors wanted to present a fully smartphone-operated, modular closed-loop Brain Computer Interface (BCI) system that can be combined with different EEG amplifiers and can easily implement other paradigms. The combination of unobtrusive EEG sensors [20], wireless EEG amplifiers, and smartphone-based signal acquisition and stimulus presentations (which we call transparent EEG [21]) opens up a plethora of possibilities for research, diagnostics, and therapy. The focus on smartphone-operated wearable devices for healthcare [22] allows for home-based applications with a high usability [23].
In previous studies, they have demonstrated the feasibility of Android smartphone-based EEG recordings [24, 25] as well as stimulus presentation on the phone or on a tablet [26]. However, while the signal quality achieved on handheld devices may be comparable to previous desktop computer-recorded EEG signals, all signal processing and classification routines were applied offline on desktop computers, after signal acquisition was concluded.
More specifically, the SMARTING Android application was used for EEG acquisition and storage [27]. The Smarting Android application receives EEG data via Bluetooth from a small, wireless head-mounted 24-channel EEG amplifier and streams signals continuously over the local network via the Labstreaming Layer (LSL) [28]. There is an option to correct the transmission delays on the receiving device with the help of the time-stamped EEG samples on the amplifier before they are sent via Bluetooth.

2) Cortical sensitivity to guitar

The second research focused on demonstrating the brain's ability to process a series of repeated complex acoustic patterns to showcase the cause for the appreciation of music. The way the study was conducted by using a series of guitar notes presented with a musical pattern or without a pattern, during which the cortical responses were measured.
While using a 24-electrode Bluetooth mobile EEG system (Smarting mBrainTrain), the ERP (Event-related potential) responses of 13 healthy non-musicians to individual notes were measured. They have used the SMARTING Streamer (the software interface for mBrainTrain's Smarting EEG amplifier) to collect data. The amplifier is connected to the computer with Bluetooth manager BlueSoleil.
The primary role of the BlueSoleil is found in the ability to enable the wireless integration of multiple Bluetooth devices such as: mobile phones, headsets, keyboards and mouses – all on one computer screen, and it can tranfer files from and to mobile phones.

5 CONCLUSION

The multitude of potential applications for wireless sensor networks is endless and make for interesting research. Bluetooth is a possible choice for data communication in sensor networks. The advantages of Bluetooth can be found in good throughput, low-power, low-cost, standardized specification and hardware availability. When it comes to disadvantage, one has to note slow connection establishment and lack of scatternet support as the more essential ones. Implemented bluetooth platform presents a good environment for further research in field od EEG monitoring.
SMARTING team choose Bluetooth technologie for EEG especially because it can support 24 channels - 3 bytes each, which is total 83bytes of data for every 2 ms (which is 41.5kbyte per second, or 332 kbit per sec). This is roughly the upper limit of what it can go through the Bluetooth Serial Port Profile (SPP) protocol. They use Bluetooth because it needs less power than wifi, so the battery is slower. Also, the Bluetooth SPP protocol is more "real time" because there is less delay compared to Wifi. Some manufacturers also use wireless profiles - a radio connection that is not Bluetooth or WiFi, but the problem is that they must then have one more device loaded onto the computer, and this is not standard communication. In that case, they can not communicate with a mobile phone.
Since most of the devices work on BLE, and SPP has been much reduced, so it cannot be close to 2 Mbit/sec, but it's about 350kbit. It's possible that it will change in some future bluetooth profiles.
With the existing design, it is possible to integrate Bluetooth wireless technology into a deep pool of new products, hence following the trend of connected information appliances. In addition, the introduction of Bluetooth technology has a double benefit – it reduces expenses for wiring and transforms industrial into smart environment.
There is some research that presents the solution for replacing cables with a wireless interface in bridge-sensor measurement applications. It is also mentioned that medical applications are more oriented towards Bluetooth tehnologies. The main idea of that research was to investigate the possibility of implementing a wireless replacement for a sensor-bridge cable. They were able to prove the concept of an inverted topology. The concept was tested using several receiver/transmitter pairs, and the results were always the same. This confirmed the idea of an inverted topology and validated it for mass production. That solution is currently integrated within a serial production for medical application [29].
Smartphones used as medical devices [30] and as scientific instruments [31, 32] will play a crucial role in the future, the main reason for this being the high usability and the large number of integrated sensors.
And even though EEG is still not quite there, it's just a matter of time before it becomes completely mobile and wireless. EEG hardware is available for low cost and, in

the near future, it may be sufficiently user-friendly and small enough to be taken out of the lab and into real-life situations. In addition, wearable EEG sensor technology seems within reach. In the future, a stable online EEG solution, requiring little more hardware than already available may support several use cases in the growing field of healthcare.

References

[1] S. Krco: Bluetooth Based Wireless Sensor Networks – Implementation Issues and Solutions, Applied Research Lab, Ericsson Ireland Invited paper

[2] M. Bandyapadhyay: Health monitoring system in Wireless sensor networks, IT, MCKV Institute of Engineering/ MAKAUT, India

[3] G. Cho, S.K. Yoo1: Wearable ECG Monitoring System Using Conductive Fabrics and Active Electrodes, Proceedings of the 13th International Conference on Human-Computer Interaction, 2009, Berlin, Heidelberg

[4] A. Darwish , A. E. Hassanien: Wearable and Implantable Wireless Sensor Network Solutions for Healthcare Monitoring, 2012, Sensors 12: 12375-12376.

[5] J. Chiang, K. Rabab: Energy-Efficient Data Reduction Techniques for Wireless Seizure Detection Systems, Ward Department of Electrical and Computer Engineering, The University of British Columbia, 2332 Main Mall, Vancouver, BC V6T 1Z4, Canada

[6] Bluetooth specification, http://www.bluetooth.com

[7] B. Skočir, G. Papa, A. Biasizzo: Multi-hop communication in Bluetooth Low Energy ad-hoc wireless sensor network, Informacije MIDEM, 2018, Vol. 48, No. 2, pages: 85-95

[8] T. Salonidis, P. Bhagwat, L. Tassiulas, R. LaMaire: Distributed Topology Construction of Bluetooth Personal Area Networks

[9] N. Johansson, F. Alriksson, U. Jönsson: JUMP Mode - A Dynamic Window-based Scheduling Framework for Bluetooth Scatternets, Proceeding of the Mobicom 2001, Rome, Italy, July 2001

[10] W.O. Tatum, A. M. Husain, S. R. Benbadis: Handbook of EEG Interpretation, Demos Medical Publishing, 2008

[11] K. V. Harpale, K. V. Bairagi: Mathematical Modeling and Data Analysis of Electroencephalographic Signals: A Review. ISSN 2250-2459, 2015, ISO 9001:2008 Certified Journal, Volume 5, Issue 3

[12] P. Yager, T. Edwards, E. Fu, K. Helton, K. Nelson, M. R. Tam, B. H. Weigl: Microfluidic diagnostic technologies for global public health, 2006, Nature, 442(7101):412–8.

[13] J. Burrell, T. Brooke, R. Beckwith (2004) Vineyard computing: sensor networks in agricultural production, IEEE Pervasive Computing, vol 03(1), pp 38-45

[14] P. A. Aberg, Tatsuo Togawa, and Francis A. Spelman, editors, 2002, Sensors in Medicine and Healthcare, Wiley-VCH.

[15] W. Dargie, C. Poellabauer: Fundamentals of Wireless Sensor Networks: Theory and Practice, 2011, ISBN:9780470997659, John Wiley & Sons, Ltd

[16] G. Fortino, S. Galzarano: Programming Wireless Body Sensor Network Applications through Agents, 2015, Department of Electronics, Informatics and Systems (DEIS) University of Calabria (UNICAL), 1-8.

[17] B. O. Sadiq, A. E Adedokun, Z.M Abubakar: The Impact of Mobility Model in the Optimal Placement of Sensor Nodes in Wireless Body Sensor Network, Department of Electrical and Computer Engineering, Ahmadu Bello University, Zaria Nigeria

[18] S. Blum, S.Debener, R.Emkes, N. Volkening, S. Fudickar, M. G. Bleichner: EEG Recording and Online Signal Processing on Android: A Multiapp Framework for Brain-Computer Interfaces on Smartphone, 2017, BioMed Research International, Article ID 3072870, 12 pages

[19] D. Bridwell, E. Leslie, D. Q. McCoy, S.M. Plis, V.D. Calhoun: Cortical Sensitivity to Guitar Note Patterns: EEG Entrainment to Repetition and Key, 2017, Front Hum Neurosci

[20] S. Debener, R. Emkes, M. De Vos, M. Bleichner: Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear, 2015, Scientific Reports, vol. 5, Article ID 16743

[21] M. G. Bleichner and S. Debener: Concealed, Unobtrusive Ear-Centered EEG Acquisition: cEEGrids for Transparent EEG, 2017, Frontiers in Human Neuroscience, vol. 11, p. 163

[22] C. Gretton and M. Honeyman: The digital revolution: eight technologies that will change health and care , The King's

[23]Fund," https://www.kingsfund.org.uk/publications/articles /eight-technologies-will-change-health-and-care. Google Research: "TensorFlow: Large-scale machine learning on heterogeneous systems." pp. 19, 2015.

[24] S. Debener, F. Minow, R. Emkes, K. Gandras, M. de Vos: How about taking a low-cost, small, and wireless EEG for a walk?, 2012, Psychophysiology, vol. 49, no. 11, pp. 1617–1621

[25] M. De Vos, S. Debener: Mobile eeg: towards brain activity monitoring during natural action and cognition, 2014, International Journal of Psychophysiology, vol. 91, no. 1, pp. 1-2

[26] B. Griffiths, A. Mazaheri, S. Debener, S. Hanslmayr: Brain oscillations track the formation of episodic memories in the real world, 2016, NeuroImage, vol. 143, pp. 256–266

[27] mBrainTrain | Fully Mobile EEG Devices, https://mbraintrain.com/

[28] Swartz Center for Computational Neuroscience and C. Kothe: Lab Streaming Layer (lsl), https://github.com/sccn/labstreaminglayer.

[29] M. Pavlin, F. Novak: A wireless interface for replacing the cables in bridge-sensor applications, Sensors, 2012, Vol. 12, No. 8, pages: 10014-10033

[30] W. H. Frist: Connected health and the rise of the patient-consumer, Health Affairs, 2014, vol. 33, no. 2, pp. 191–193

[31] R. Kwok: Personal technology: Phoning in data, 2009, Nature, vol. 458, no. 7241, pp. 959–961, 2009.

[32] J. Cartwright: Smarthphone science: researchers are learning how to convert devices into global laboratories, 2016, Nature, vol. 531, pp. 669–671

What is a FabVillage? Introducing new concepts to make rural areas smarter

Rebeka Škrjanc, Veronika Zavratnik, Andrej Kos, Emilija Stojmenova Duh
Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana, Slovenija
rebeka.skrjanc@ltfe.org, veronika.zavratnik@ltfe.org, andrej.kos@fe.uni-lj.si, emilija.stojmenova@fe.uni-lj.si

Abstract

Fast evolving technology is often at the center of social changes and challenges which communities are tackling in the last decades. In this paper, we are focusing on solutions of smart development in both urban and rural areas. We discuss them in the context of wider european policies, initiatives and examples. In the area of smart rural development we are also introducing a new concept - FabVillages.

Keywords: Smart City, Smart Village, FabLab, FabCity, FabVillage

Uvod

Mesta se ne soočajo z enakimi družbeno-ekonomskimi problemi kot podeželje. Resursi, optimizacija prometa, količina proizvedenih smeti, uvoz, izvoz,... Vse to so problemi, s katerimi se ukvarjajo mestne oblasti. Podeželje ima drugačne probleme, kot so izseljevanje mladih, odrezanost od večjih mest in dostopnost javnih ustanov, ki ponujajo storitve, kot so izobraževanje, oskrba, itd. V nadaljevanju bomo opisali pameten razvoj mest, ki je bil mogoč z uporabo novodobnih digitalnih tehnologij (IKT). Predstavili bomo kako so ustvarjalni laboratoriji (ang. FabLabi) prispevali k razvoju pametnih mest in opisali koncept FabMesta (ang. FabCity), ki ga je zastavila svetovna FabLab mreža. Še posebej želimo izpostaviti pojem in koncept FabVasi (ang. FabVillages), ki smo ga ustvarili v Laboratoriju za telekomunikacije, FE UL, in ga uvajamo tudi v svojih projektih.

Pametna mesta in skupnosti

Hiter tehnološki razvoj narekuje spremembe, ki drastično preoblikujejo naša življenja. Velik vpliv ima tudi na gospodarstvo, kar je v zadnjih letih pripeljalo do naglega družbenega razvoja. Najbolj je to opazno v urbanih območjih, ki thenološko napredujejo z večjo hitrostjo kot ruralna. V zadnjih letih se je dobro uveljavil pojem Pametna mesta (ang. Smart cities), s katerim opisujemo urbana naselja, ki s pomočjo informacijskih komunikacijskih tehnologij (IKT) skrbijo za boljšo kvaliteto življenja na določenem območju. S pomočjo senzorike in konceptov interneta stvari (ang. IoT) lahko nadzorujejo porabo energije in vode, proizvodnjo smeti, izboljšajo mobilnost, zmanjšajo emisije, itd. [4] Z analizo pridobljenih podatkov lahko dosegajo boljše rezultate na teh pomembnih področjih. Posledica tega sta učinkovitost in trajnost, ki jo mesta lahko dosegajo z boljšim načrtovanjem porabe sredstev, ter posledično manjšanjem proizvedenih odpadkov. Na tak način mesta postajajo "pametna", kar pozitivno vpliva na ekonomsko rast, okolje ter zadovoljstvo prebivalstva. [1, 2]

Vsako (pametno) mesto je drugačno in se sooča z drugačnimi problemi. Osredotočajo se na področja, ki so ključna za njihov razvoj, zato njihove napredke (med sabo) težko primerjamo. Za vsa mesta pa velja, da bodo napredovala bistveno hitreje, če bodo v proces razvoja vključeni tudi meščani. Na primer, interaktiven prikaz podatkov dostopen javnosti, lahko pripomore k ozaveščenosti prebivalcev in jih motivira, da tudi sami prispevajo k izboljšavam.
Pametna mesta se trudijo z meščani navezati tesnejši stik in s pomočjo novodobnih tehnologij vzpostaviti nov komunikacijski kanal. Ta jim omogoča, da prebivalci s predstavniki lokalnih oblasti delijo svoja mnenja glede aktualnih dogodkov v mestu, hkrati pa postanejo del procesa odločanja. Tako lahko župan ves čas spremlja njihovo (ne)zadovoljstvo, meščani pa se tako počutijo bolj vključene.
S spremljanjem podatkov pridobljenih s pomočjo senzorjev ter odziva/mnenja meščanov skozi čas, dobimo pregled nad razvojem mesta.

Kljub temu, da koncept pametnih mest rešuje veliko problemov, se z vsemi ne more soočiti. Digitalizacija je velik del pametnega razvoja, obstajajo pa tudi druge pomembne komponente, katerim pametna mesta ne namenjajo veliko pozornosti. Kot primer si lahko pogledamo eno večjih pametnih mest na svetu, Rio de Janeiro. S pomočjo sodobne tehnologije uspešno napovedujejo bližajoče se, nevarne vremenske pojave in s tem poskrbijo za hitrejšo ter bolj učinkovito evakuacijo. Na dolgi rok bi ta problem bolje rešili, če bi raziskali zakaj ljudje gradijo na tako nevarnih območjih in kako bi jih lahko preusmerili drugam. [3]

Ko govorimo o pametnem razvoju, so urbana območja definitivno v prednosti, pred ruralnimi. Vendar pa se razvoj podeželja v sklopu globalnih iniciativ vedno bolj spodbuja po celem svetu. Kot rezultat tega, se je uveljavil tudi pojem Pametne vasi (ang. Smart villages). Kljub temu, da si na evropski ravni Evropska unija že dolgo prizadeva za razvoj podeželja, se je pri nas o pametnih vaseh eksplicitno začelo govoriti šele leta 2016. Takrat je bila zasnovana Deklaracija Cork 2.0 pod imenom "A Better Life in Rural Areas", v kateri so opisane problematike s katerimi se soočajo prebivalci podeželja. Deklaracija je postala vodilo za nadaljne iniciative za pametne vasi po Evropi. Deklaracija je vodila do zaključka, da bo Evropska unija v podeželja morala vložiti več denarja, da bodo le-ta privlačila vse generacije in da bi izboljšali ekonomsko rast. Leto kasneje je evropska komisija zagnala iniciativo/akcijo EU Action for Smart Villages. Aprila, leta 2018 je bila na Bledu sprejeta blejska deklaracija, ki se zavzema za pametnejšo prihodnost podeželskih območij EU. V njej so zapisani cilji, ki naj bi jih s pametnim razvojem

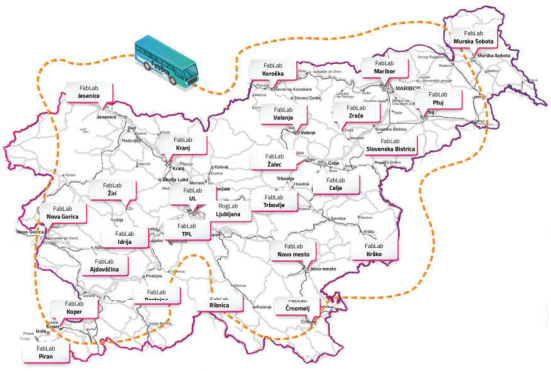
dosegli, da bi čim bolj spodbudili tako digitalno, kot družbeno preobrazbo podeželja v EU. [5, 6, 13]

Ustvarjalni laboratoriji - FabLabi

FabLab (fabrication laboratory) je prostor v katerem lahko kdorkoli, s pomočjo sodobnih orodij, digitalnih tehnologij in mentorjev, razvija svoje ideje. Prvi FabLab je bil ustanovljen v Centru za bite in atome v Tehnološkem inštitutu Massachusetts (MIT). Do danes se je FabLab mreža že močno razširila. Po celem svetu jih je že več kot 1200, najdemo pa jih kar v 78 državah sveta. [8] S širitvijo mreže se želi k inovativnemu in ustvarjalnemu razmišljanju spodbuditi tudi manj razvita območja. Delavnice spodbujajo razvoj lokalne industrije in podjetništva, veliko vlogo pa imajo tudi pri izobraževanju in spodbujanju kreativnosti.

FabLab je odprto okolje, v katerem je dobrodošel vsak, ki ima inovativno idejo ali željo po znanju. Starost in predznanje nista pomembna. Gre za prostor, ki spodbuja medsebojno povezovanje, podjetništvo in pridobivanje novih znanj ter izkušenj. Razvijalsko okolje, sodobna oprema in osebe na katere se lahko ustvarjalci obrnejo, nudijo podporo in spodbujajo okoliške prebivalce k inovativnosti ter sodelovanju. Napredna tehnološka oprema, ki jo laboratoriji nudijo, ponavadi zajema 3D tiskalnik, laserske stroje za rezanje in graviranje ter CNC stroje. [7, 11]

V Sloveniji smo prvi FabLab dobili v letu 2012. Imenuje se RogLab, nahaja pa se v Ljubljani. Posvečen je dejavnostim s področij arhitekture, oblikovanja in sodobne umetnosti, njihovemu medsebojnemu povezovanju in sodelovanju z drugimi sektorji ter mednarodnemu mreženju. [17] Na Fakulteti za elektrotehniko je bil 2014 ustanovljen MakerLab, ki študentom nudi dostop do orodij in znanj, ki jih potrebujejo za razvijanje svojih inovativnih spretnosti. [18] Že v letu 2017 smo dobili nacionalno referenčno FabLab mrežo, katere pobudnica in koordinatorica je Fakulteta za elektrotehniko Univerze v Ljubljani v sodelovanju z Direktoratom za informacijsko družbo, Ministrstva za javno upravo.



Slika 1: FabLab Zemljevid

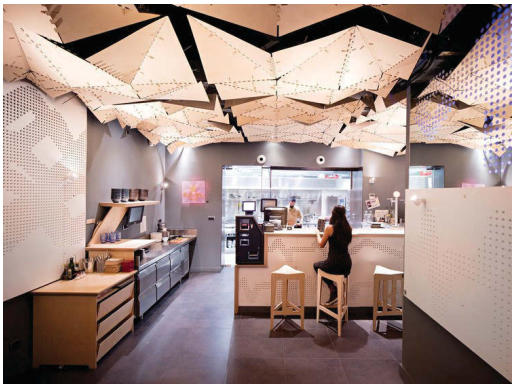
FabLab mreža Slovenije si močno prizadeva za razvoj lokalnega gospodarstva, zato spodbuja nastanek laboratorijev v vseh regijah in širitev mreže tudi na podeželje. Mreža želi tehnologije približali tudi tehnološko manj razvitim, ruralnim območjem in s tem omogočiti, da tamkajšnji prebivalci pridobijo novodobna znanja in kompetence. Kar nekaj ustvarjalnih laboratorijev je bilo zato ustanovljenih na podeželju, vsi pa so nastali na podlagi obstoječih lokalnih iniciativ. [7]

FabMesta

Model FabMest (ang. FabCity) se je razvil pod okriljem globalne FabLab mreže. Pojem temelji na transformaciji in oblikovanju novih načinov za lokalno proizvodnjo in pametnejšo porabo sredstev. Vizija FabLab mreže je, da bi mesta v prihodnosti postala samozadostna. Da bi to dosegli, morajo najti način, da bi prebivalci znotraj mesta proizvedli vse kar potrebujejo za svoje delovanje in obstoj. To bi lahko dosegli s postavitvijo več FabLab delavnic, ki bi bile usmerjene v lokalno proizvodnjo. Da bi bili pri tem čim bolj učinkoviti, je potrebno premisliti in reorganizirati tudi samo porabo virov, saj želijo, da so ti čim bolj trajni. [9, 10]

FabMesta slonijo na modelu "from PITO to DIDO". PITO predstavlja kratico za "Product In, Trash Out". Gre za princip po katerem so mesta delovala do sedaj. Produkte uvažajo, jih uporabijo in pri tem proizvedejo odpadke, ki jih nato izvozijo. V prihodnosti pa naj bi se uveljavil nov princip; DITO-"Data In, Data Out". Nov model mesta tako ne bo več izvažal odpadkov, saj bodo ti sami po sebi predstavljali resurse, ki jih bo mesto ponovno uporabilo. FabMesta bodo temeljila na trajnosti in pametni uporabi virov. [11, 12]

Vicente Guallart - arhitekt in vodja FabLab mreže v Barceloni je močen podpornik FabMest. "Pripravljeni smo na samozadostna mesta in na to, da od FabLabov preidemo na FabMesta. Na tej točki, bi morali biti posamezniki in lokalne skupnosti zmožni proizvesti skoraj vse, kar v mestu potrebujejo". [11] V Barceloni si prizadeva, da bi postavili več FabLab delavnic, ki bi bile med seboj oddaljene zgolj en kvadratni kilometer. Tako bi okoliškim prebivalcem omogočili, da se povežejo in postanejo del skupnosti, ki bi stremela k samozadostnosti znotraj lastnega območja. Vicente si je zadal cilj, da bo Barcelona postala samozadostno mesto do leta 2054, k sodelovanju pa je pozval tudi ostala svetovna mesta.



Slika 2: Restavracija Leka v Barceloni

Lep primer povezovanja ustvarjalnih laboratorijev in mest je restavracija Leka, ki se nahaja v Barceloni. Vsa oprema je bila narejena v njihovem lokalnem FabLabu. Vsi načrti za opremo in recepti, ki jih restavracija uporablja, so dostopni javnosti. Najdemo jih lahko na njihovi spletni strani.

FabVasi

V Sloveniji na podeželju živi približno polovica prebivalstva. Podeželje pa se tehnološko razvija bistveno bolj počasi, kot mesta. Srečuje se z drugačnimi problemi, zato je razvoj ruralnih območij potrebno zasnovati malo drugače. Velik problem predstavlja pomanjkljiva infrastruktura, ki omejuje možnosti delovnih mest izven kmetijskega sektorja. Posledica je odseljevanje mladih v urbana območja, odraz česar je staranje prebivalstva na podeželju. Ogromno regij v Sloveniji se srečuje tako z demografskimi kot gospodarskimi problemi. V zadnjih letih je bilo na tem področju narejenih kar nekaj v razvoj usmerjenih sprememb; izboljšanje prometnih povezav, izboljšani dostop do širokopasovnih povezav ipd. Pomemben del pametnega razvoja predstavlja digitalizacija. Nanjo se osredotočajo predvsem sodobna mesta, medtem, ko se na podeželju osredotočajo na reševanje drugačnih problemov, kot so razvoj zelenega turizma, ohranjanje mladih na podeželju, ustvarjanje novih delovnih mest ter krepitev podjetniškega potenciala. Kljub temu pa je vidik digitalizacije za pametni razvoj podeželja zelo pomemben.

V raziskavah, ki se osredotočajo na razvoj podeželja, smo v Laboratoriju za telekomunikacije, FE UL, opazili veliko že obstoječih lokalnih iniciativ, ki delujejo v smeri lokalnega, krožnega in trajnostnega razvoja gospodarstva. Kot povezavo med temi iniciativami, Fab mesti in pametnimi vasmi, smo začeli uveljavljati koncept FabVasi (ang. FabVillages). Koncept je tesno povezan s FabMesti. Utemeljen je na spodbujanju krožne ekonomije, uporabi lokalnih dobrin in materialov, razvoju lokalnih naravnih/kulturnih/družbenih virov in gospodarstva. Prav tako stremimo k temu, da bi podeželje postalo čim bolj trajnostno in "prožno". FabVasi slonijo na »grass roots« iniciativah, za katere je značilno, da izvirajo iz potreb ljudstva. Razvoj podeželja se tako začne pri krajanih samih in ne pri zakonodaji ali oblasti.

V Sloveniji je prisotnih že kar nekaj iniciativ, ki spodbujajo trajnostno naravnan razvoj podeželja. Omenili smo že FabLab delavnice, ki jih lahko najdemo v vseh regijah Slovenije. Še ena pomembna iniciativa pa so razpršeni hoteli. Gre za inovativno obliko organiziranosti v turizmu, ki se osredotoča na izvajanje namestitvene dejavnosti in poslovnega povezovanja v turističnih kooperativah. [15] Na podeželju je mnogo neizkoriščenega potenciala, ki ga razpršeni hoteli združujejo v celovito turistično ponudbo. Prazni prostori v hišah, dejavnosti ter pridelki so razpršeni po celem kraju. V razpršenem hotelu si namestitveni prostori delijo recepcijo, njihovim gostom pa je na voljo vrsta dejavnosti, ki jih kraj ponuja. Na tak način lahko v turistično obrt vključimo posameznike, ne da bi morali sami ustanoviti svoje podjetje in razviti celo vrsto dejavnosti v ta namen. [14, 16] V Sloveniji je prvi razpršeni hotel začel obratovati v Slovenskih Konjicah. Upravlja ga Kooperativa Konjice, obratovati pa je začel 1.1.2018. Sedež hotela in recepcija se nahajata v Dvorcu Trebnik. V njem so poleg prenočišč tudi konferenčna dvorana, kavarna in restavracija Mali Grof, v kateri za pripravo obrokov uporabljajo lokalne sestavine. Gostom, ki so nastanjeni v hotelu so v tej restavraciji na voljo zajtrki, malice in kosila.
Občina delovanje hotela odobrava in spodbuja. Hotel tesno sodeluje s ponudniki ostalih storitev v mestu in regiji. Vso notranjo opremo razen kopalnic jim je oblikovalo podjetje iz Mokronoga, v lokalnem FabLabu (U-Lab) pa so naredil tablice s številkami sob. S srednješolskim centrom sodelujejo tako, da v hotelu nastanijo gostujoče profesorje, sodelujejo pa tudi s smučiščem Rogla. Svojim gostom namreč nudijo ugodnosti na smučišču. Hotel ima 3 zvezdice, približno 80% njegovih gostov pa je tujcev.



Slika 3: Ročno narejena tablica za označevanje številke sobe v razpršenem hotelu v Slovenskih Konjicah

Nekaj podobnih iniciativ povezanih z razpršenimi hoteli, ki še niso dokončno formalizirane pa najdemo tudi drugod po Sloveniji: v Padni, Robidišču, Blatni Brezovici, na Krasu.

Koncept FabVasi, ki ga predlagamo na Fakulteti za elektrotehniko bomo v sklopu različnih projektov tudi poskušali prenesti v prakso, na slovensko podeželje. Bistveno je, da gre za povezovanje iniciativ, ki že obstajajo, kot so na primer razpršeni hoteli in FabLab delavnice. FabLabi se lahko povežejo tudi z ostalimi lokalnimi ponudniki storitev; za restavracije, kavarne in ponudnike storitev lahko izdelajo pohištvo prav tako pa poskrbijo tudi za izboljšanje povezave z internetom. Še eden izmed načinov sodelovanja je oblikovanje spletnih strani ali razvoj tematskih aplikacij. Vse to so načini, na katere se lahko FabLab delavnice še bolj vključijo v lokalna in regionalna okolja in doprinesejo k temu, da se (razvojni)potenciali, ki jih ponuja slovensko podeželje kar najbolje izkoristijo.

Zaključek

Tehnološki razvoj in digitalizacija sta močno vplivala na razvoj mest in podeželja skozi leta. Pametni razvoj je mestom in urbanim skupnostim omogočil boljši

vpogled ter načrtovanje porabe resursov. Mesta so tako postajala pametnejša, kar je imelo dober vpliv na ekonomsko rast, ekološki vidik in na njihove prebivalce. Čeprav digitalizacija predstavlja rešitev mnogih problemov s katerimi se danes soočajo tako mesta kot podeželje, se morajo skupnosti razvijati tudi v drugih smereh. Na podlagi obstoječih primerov, ki so usmerjenih v trajnostni in pametni razvoj podeželja predstavljamo nov koncept Fab Vasi. Koncept, ki je bil primarno uveljavljen v urbanih, gosto poseljenih območjih smo na podlagi obstoječih primerov iz prakse prenesli v ruralna območja. V sklopu projektov, ki jih izvajamo v Laboratoriju za telekomunikacije, FE UL, predlagani koncept tudi prenašamo v prakso.

Viri

[1] IoT Agenda, https://internetofthingsagenda.techtarget.com/definition/smart-city
[2] European Smart Cities, http://www.smart-cities.eu
[3] A. Glasmeier, S. Cristopherson: Thinking about smart cities, 2015, US
[4] European Commission, https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en
[5] EU Action for Smart Villages, https://ec.europa.eu/agriculture/sites/agriculture/files/rural-development-2014-2020/looking-ahead/rur-dev-small-villages_en.pdf
[6] V. Zavratnik, A. Kos, E. Stojmenova Duh: Smart Villages: Comprehensive Review of Initiatives and Practices, 2018, Slovenia
[7] FabLab Mreža Slovenija, http://fablab.si
[8] Fab Foundation, http://fabfoundation.org/index.php/what-is-a-fab-lab/index.html
[9] Summit, https://summit.fabcity.paris
[10] World Economic Forum, https://www.weforum.org/agenda/2018/03/makerspaces-smart-sustainable-cities-thomas-ermacora/
[11] Waag technology & society, https://waag.org/en/article/fab-labs-fab-cities
[12] Tomas Diez: Personal Fabrication: Fab Labs as Platforms for Citizen-Based Innovation, from Microcontrollers to Cities, 2012, Barcelona
[13] Blejska deklaracija, http://pametne-vasi.info/wp-content/uploads/2018/04/Blejska-Deklaracije-Za-za-pametnejšo-prihodnost-podeželskih-obmocij-v-EU.pdf
[14] Dovolj za vse, http://dovoljzavse.si/wp-content/uploads/2017/10/prirocnik_razprseni-hotel-1.pdf
[15] Razpršeni hotel Konjice, https://razprsenihotel.si
[16] Dovolj za vse, http://dovoljzavse.si/kako-v-sloveniji-vzpostaviti-zadruzni-razprseni-hotel-in-urediti-skupnostni-vrt/
[17] RogLab, http://fablab.si/fablabs/roglab/
[18] MakerLab, http://fablab.si/fablabs/makerlab/

Storitev označevanja nogometne tekme z informacijami o igralcih

Danilo Zimšek[1], **Luka Banfi**[2]

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,
Koroška cesta 46, 2000 Maribor
E-pošta: danilo.zimsek@um.si, luka.banfi@student.um.si

Football match annotation service with player information

Our paper describes an implementation of a football match annotation service with player information. Use case for the service is discussed with its overall implementation.

In this paper the colors of jerseys are analysed and limit color values for process of binarization are determined for 14 jerseys of randomly selected teams. Then the process of localisation and extraction of numbers is described. Special consideration is put on cases with same color of upper and lower part of jerseys. Other problematic frames are discussed, where there is colour similarity between jersey and background. Afterwards the algorithm for number recognition is described. The use of described approach is explained and annotation process described. The results of the process of localisation and recognition of numbers are also given.

In conclusion, future work and possible improvements of the process are discussed.

1 Uvod

Nogomet je že od nekdaj eden od najbolj popularnih športov predvsem v Evropi in Južni Ameriki, svojo popularnost pa pridobiva tudi v regijah, kjer je slabše poznan. Zanimanje za ta šport skozi zgodovino raste. Zaradi velikega zanimanja je za ponudnike telekomunikacijskih storitev ta šport posebej zanimiv za ponujanje novih storitev ogleda vsebine. Storitvi kot sta označevanje vsebine in skupinski ogled vsebine z videokonferenco lahko omogočita uporabnikom interaktivno spremljanje tekme v mikro socialnem omrežju, kjer lahko uporabniki v izbranem časovnem trenutku določenemu igralcu pridajo izbran čustveni simbol in tako na ta način izrazijo svoje mnenje o potezi igralca.

Takšna storitev, zasnovana na okolju IP televizije, ponudi uporabniku dodatno možnost socializacije in je iz tega vidika posebej pomembna za starejše in invalide, saj jim lahko uporaba tehnologij omogoči izkušnjo, ki je blizu izkušnji skupinskega spremljanja nogometne tekme, ki je dokaj pogost pojav.

Storitev skupinskega spremljanja nogometne tekme je zgrajena iz naslednjih modulov:

- storitev video konference;
- sistem povabljanja uporabnikov;
- storitev označevanja nogometne tekme.

Sistem povabljanja uporabniku omogoča, da drugega uporabnika povabi k ogledu nogometne tekme. Drug uporabnik lahko povabilo sprejme ali zavrne. V primeru, da vabilo sprejme vsaj en uporabnik, storitev video konference ob času začetka predvajanja nogometne tekme tvori konferenčno sobo v katero vključi uporabnike, ki so sprejeli vabilo. Video konferenca omogoča, da uporabniki v manjših delih zaslona spremljavo video portrete drugih udeležencev in se med seboj pogovarjajo. Hkrati se uporabnikom predvaja obogaten video tok nogometne tekme, ki ga tvori storitev označevanja nogometne tekme. Interakcija z vsebino je omogočena preko razširjene funkcionalnosti storitve video konference.

Eden glavnih gradnikov pri implementaciji zgoraj opisanih storitev je označevanje vsebine z informacijo o trenutno prikazanem igralcu. V nadaljevanju bomo opisali pri tem uporabljene pristope.

Zaznava in razpoznava igralca nogometne tekme je pogosto obravnavana tematika med postopki označevanja video vsebin za razpršeno oddajo. Pristop v [2] uporabi v procesu zaznave pri empiričnem določanju mejnih vrednosti za izločanje dresov igralcev na podlagi barv metodo odvisno od svetlosti, kar povzroči, da se pojavljajo težave pri video okvirjih, ki so delno ali v celoti senčni. Za zmanjšanje vpliva osvetljenosti na algoritem zaznave dresa igralca, smo uporabili barvni model HSV. Ta se je po testiranju 5 primerov video okvirjev v barvnih prostorih RGB, BGR, HSV in YCbCr izkazal kot najprimernejši, saj smo empirično ocenili najboljše ločevanje med dresom in ozadjem okvirja.

Poleg odvisnosti od osvetljenosti, se v procesu zaznave pojavljajo omejitve vezane na podobnost med barvami zastopanimi na sceni in dresom igralca ter težave pri popačitvah predvsem zaradi umazanih dresov. Posebno težavo predstavljajo za do sedaj znane algoritme dresi, ki so v celoti v eni barvi, ki je enaka barvi nogavic. Te situacije so v zadnjem času pogostejše.

Naš cilj je z uporabljenim pristopom zmanjšati vpliv naštetih omejitev. V literaturi se za razpoznavo igralca uporabljajo različne kombinacije naslednjih tehnik: razpoznava številke in razpoznava obraza. V našem pristopu se omejimo na razpoznavo številke. Omejitev našega pristopa je delovanje na bližnjih posnetkih, saj lahko le tako zagotovimo dovolj natančno razpoznavo številk, kar pa



Slika 1: Vhodni video okvir.



Slika 2: Binariziran video okvir.

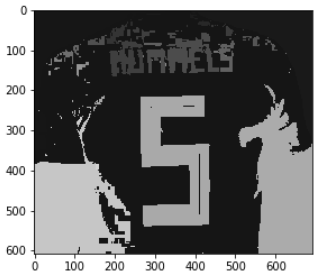
je smiselna omejitev, saj se bližnji posnetki navadno pojavljajo po pomembnem dogodku, katerega glavni akter je prikazan igralec.

V naslednjem poglavju opišemo postopek lokalizacije dresa, način določanja mejnih vrednosti barv za ta namen in zaznavo številke, pri čemer izpostavimo zahtevnejše primere. V tretjem poglavju obravnavamo postopek razpoznave številke pri čemer uporabimo deskriptor in mero sorodnosti z referenčno sliko številke. V četrtem poglavju opišemo označevanje vsebine. V petem predstavimo in komentiramo rezultate zaznave in razpoznave ter v zaključku podamo predloge za nadaljnje delo in možnosti za izboljšave.

2 Lokalizacija dresa

Prvi korak v postopku označevanja nogometne tekme z informacijami o igralcih predstavlja lokalizacija dresa, to je določanje pikslov v okvirju, ki pripadajo igralcu na igrišču. Predpostavljeno je bilo, da so določene barve povezane z določenimi nogometnimi ekipami. Za pravilno delovanje metode lokalizacije dresa, ta potrebuje podatek o barvnem razponu, ki predstavlja veliko večino barv zastopanih v dresih posamezne ekipe. Razpon vrednosti smo za vsako ekipo doličili eksperimentalno v barvnem prostoru YUV. Proces vzpostavljanja primernih razponov vrednosti znotraj določenega barvnega modela prikazujejo enačbe 1, 4 in 3.

$$\Delta Y = \max_n Y_n - \min_n Y_n \tag{1}$$



Slika 3: Pridobljeni objekti izločenega območja video okvirja okvirja.

$$\Delta U = \max_n U_n - \min_n U_n \tag{2}$$

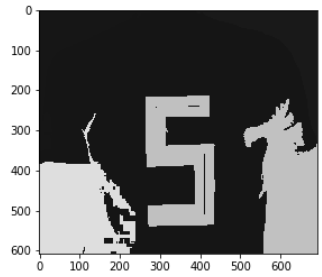
$$\Delta V = \max_n V_n - \min_n V_n \tag{3}$$

pri čemer je $n \in \{1, ..., št.okvirjev\}$ in Δ označuje uporabljen barvni razpon, n pa indeks analiziranega okvirja. Skupno število analiziranih okvirjev za posamezno barvo dresa je bilo odvisno od razpršenosti trenutno pridobljenih podatkov. Večja kot je bila razpršenost, več okvirjev smo analizirali.

Proces lokalizacije dresa sestoji iz binarizacije slike, izločanja povezanih komponent in odstranjevanja madežev. Binarizacijo smo izvedli z uporabo empirično določenih mejnih vrednosti za vsako od komponent barvnega prostora. Slika 1 prikazuje primer vhodne, slika 2 pa primer binarizirane slike.

V nadaljevanju znotraj binarizirane poiščemo največje povezano področje bele barve. Izbrano področje uokvirimo in v nadaljevanju procesiramo samo to območje. Znotraj izločenega področja označimo vsa povezana področja (slika 3). Dobimo predstavitev večih objektov, med katerimi je tudi številka.

Opazimo lahko, da so nekateri objekti zelo majhni, zato odstranimo objekte, ki so manjši od 5% celotne površine izbranega območja. S to omejitvijo izločimo napise imen igralcev na dresih in druge nepravilnosti v video okvirju. Dobljeno prikazuje slika 4. Ker je v nekaterih primerih barva majice enaka barvi hlač igralca, barvi nogavic ter ozadju (golovi vratnici), dobimo več objektov, iz katerih želimo izločiti samo številko. Do sedaj znani pristopi[1] za ta namen uporabljajo omejitev glede minimalne širine objekta ali omejitve glede razmerja stranic. Minimalna širina objekta v takem zahtevnejšem primeru ni zadosten pogoj, zato uporabimo omejitev glede razmerja stranic. Za večino pisav je razmerje med višino in širino števk med $2 : 1$ in $1,5 : 1$, zato smo uporabili ta pogoj. Poleg tega smo dodali pogoj o maksimalni višini objekta glede na višino izločenega območja in o maksimalni površini objekta glede na površino izločenega območja. Višina objekta ne sme presegati dveh tretjin višine območja. Površina objekta pa ne sme presegati ene polovice izbranega območja. Omejitve so bile določene



Slika 4: Problematični objekti video okvirja po odstranitvi majhnih objektov.

empirično glede na dobljene objekte 10 naključno izbranih testnih okvirjev. Če upoštevamo vse naštete pogoje se število napačno izločenih objektov zmanjša na zanemarljivo število, tudi pri težavnih primerih. Do napačnih zaznav pride v manj kot 5% primerov.

3 Razpoznava številke

Eden od pristopov razpoznave števil vključujejo deskriptorje oblike. Eden od bolj primernih, zlasti zaradi neznane velikosti in usmerjenosti številke, je generični Fourierjev deskriptor (angl. Generic Fourier Descriptor), v nadaljevanju GFD [3].

Deskriptor obliko opiše tako, da se koordinate pikslov, ki pripadajo področju slike z obliko, transformirajo v polarne koordinate in prenesejo v pravokotno kartezijsko sliko. Nad podatki je uporabljena dvodimenzionalna Fourierjeva transformacija. Med dobljenima predstavitvama se izračuna Evklidova razdalja, ki predstavlja mero podobnosti med referenčnim objektom in kandidatom. V postopek GFD damo binarno sliko, ki ima en objekt v središču. Deskriptor izračuna stopnjo ujemanja vhodnega objekta z referenčnim objektom. V našem primeru so referenčni objekti slike števk pridobljene iz številk zapisanih v pisavi Arial v krepkem načinu. Algoritem izračuna s pomočjo GFD podobnost za vsak izločen objekt do vseh referenčnih objektov. Najvišjo mero podobnosti primerja z eksperimentalno določeno mejo. Če je ta višja od empirično določene meje, gre za ujemanje in objekt na sliki predstavlja obravnavano številko. V tem primeru steče proces označevanja razpoznave na video tok.

Eksperimentalno določanje meje je potekalo na podlagi naključnega izbora dveh vzorcev video okvirjev vsake ekipe. Vzeli smo povprečno vrednost rezultata GFD iz vsake kategorije, izračunali povprečno vrednost nad njimi in zmanjšali za 20%. Opisan postopek prikazuje enačba 4.

$$m = \frac{\sum_{k=1}^{št.\ kategorij} f_k}{št.\ kategorij} \times 0.8 \tag{4}$$

pri čemer m označuje mejno vrednost in f_k povprečje mejnih vrednosti video okvirjev posamezne kategorije.

4 Označevanje vsebine

Za potrebe opisane storitve je potrebno uporabniku posredovati informacijo o prepoznanem igralcu. Slednje smo izvedli z uporabo informacij iz elektronskega programskega vodiča (ang. electronic program guide) in z dodajanjem delno transparentne vsebine na video sled predvajanega video streama.

Pridobivanje informacij iz EPG je izvedeno s skripto, ki pregleda EPG v zapisu XML in na podlagi zapisa naslova vsebine izloči vse nogometne tekme in informacije o ekipah. Na podlagi teh informacij pridobi podatke o sestavi ekipe in preslikavami med imeni igralcev in številko dresa. Ti podatki se uporabijo v procesu označevanja video toka.

Označevanje video vsebine izvedemo z uporabo programske knjižnice ffmpeg [6]. Ta ponuja funkcionalnosti označevanja vsebin v realnem času, kar omogoča prikaz poljubne tudi delno transparentne vsebine na vsakem video okvirju. Po izvedenem označevanju se obogatena vsebina struja do uporabniškega klienta.

Storitev smo izvedli z uporabo programskega jezika Python in uporabo knjižnic scikit-image, opencv in scipy. Storitev smo poganjali na štiri jedrnem sodobnem procesorju intel. Pri tem je sistem za delovanje uporabljal dve jedri, saj vzporedno tečeta dva procesa, za vsako od dveh ekip nogometne takme za isti vhodni video okvir. Ob predpostavki, da ima video tok 25 sličic na sekundo, smo uspeli storitev na navedeni strojni poganjati za vsak deseti video okvir, torej je celoten postopek potreboval manj kot 0.4 sekunde.

5 Rezultati

Za evalvacijo opisane metode je bilo najprej potrebno zgraditi bazo. Baza video okvirjev nogometnih tekem sestoji iz 14 kategorij. V vsaki kategoriji je od 10 do 15 video okvirjev iz bližnjih posnetkov z igralci istih ekip. Zgradbo baze in uspešnost uporabljenega pristopa prikazuje tabela 1.

Iz dobljenih rezultatov je razvidno, da je omenjen pristop z uporabo barvnega modela YUV in določanje mejnih vrednosti za binarizacijo iz izbranega modela uspešnejši od pristopa opisanega v [1]. Tak rezultat smo pripisali predvsem določanju mejnih vrednosti za vsako ekipo posebej v barvnem prostoru YUV. Ker pa rezultati niso testirani na isti bazi in ker je pridobljena baza relativno majhna, je težko z gotovostjo trditi, da je to edini faktor, ki vpliva na izboljšanje.

K nepravilnim razpoznavam pa so prispevali predvsem okvirji, pri katerih so igralci z dresi in nogavicami v isti barvi kot ozadje ter okvirji slabe kakovosti.

Trenutni pristop je mogoče uporabiti za implementacijo opisanih storitev, a se pri tem pojavlja težava pri razpoznavi igralcev s številko dresa sestavljeno iz dveh števil. To težavo rešujemo na način, da na podlagi koordinat skrajnih vrednosti mejnega polja števke in njegove oddaljenosti od druge zaznane števke določimo, ali

Tabela 1: Uspešnost uporabljenega pristopa po kategorijah video okvirjev.

Razred	št. okvirjev s števili	št. pravilnih razpoznav	uspešnost[%]
1.	15	13	87
2.	12	10	83
3.	11	10	91
4.	14	13	93
5.	13	12	92
6.	10	8	80
7.	12	10	83
8.	13	12	92
9.	11	9	82
10.	15	13	87
11.	12	10	83
12.	10	9	90
13.	12	11	92
14.	11	10	91
skupaj			**88**

števki sestavljata število. Če je njuna medsebojna oddaljenost manjša od polovice širine števke, predvidevamo, da števki pripadata istemu številu.

6 Zaključek

V sklopu nadaljnjega dela načrtujemo vključitev opisane storitve označevanja vsebin v storitev videokonference. Za dosego tega cilja je potrebno avtomatizirati procese znotraj celotne storitve, ki trenutno potrebujejo za uspešno delovanje vnos informacij s strani človeka. Tak je predvsem proces pridobivanja podatkov o igralcih vključenih v nogometno tekmo. V nadaljevanju bomo procese pridobivanja poskušali izvesti na enega od naslednjih načinov:

- razpoznava imen igralcev in številk dresov iz začetnih zaslonov prenosa tekme;
- pridobivanje podatkov iz dostopnih aplikacijskih vmesnikov.

Nadalje bomo omogočili interaktivno rokovanje z vsebino s strani uporabnika na način, da bo uporabnik s funkcijskimi tipkami daljinskega upravljalnika imel možnost dodajanja čustvenih simbolov ob času razpoznave igralca in določeno časovno obdobje po koncu bližnjega posnetka.

Želeli bi doseči tudi izboljšanje natančnosti razpoznave števil. Za ta namen bi lahko generični Fourierjev deskriptor nadomestili z drugim [5] ali bi za ta namen uporabili pristop z nevronskimi mrežami [4]. Celoten postopek je kritičen predvsem zaradi zahteve po izvajanju tekom predvajanja vsebine, zaradi česar bomo pred menjavo pristopa preučili časovno zahtevnost posameznih alternativnih deskriptovjev in pristopov z nevronskimi mrežami.

Literatura

[1] Frejlichowski, Dariusz: Identification of Football Players based on Generic Fourier Descriptor Applied for the Recognition of Numbers, Image Processing & Communications. 2016

[2] Frejlichowski, Dariusz: A Method for Data Extraction from Video Sequences for Automatic Identification of Football Players Based on Their Numbers, Image Analysis and Processing – ICIAP, 356–364, 2011

[3] Dengsheng Zhang, Guojun Lu: Generic Fourier descriptor for shape-based image retrieval, IEEE International Conference on Multimedia and Expo, 425–428, 2002

[4] Yin, Wu, Zheng Ncfm: Accurate handwritten digits recognition using Convolutional Neural Networks, International Joint Conference on Neural Networks (IJCNN), 525–531, 2016

[5] Dengsheng Zhang, Guojun Lu: Enhanced Generic Fourier Descriptors for object-based image retrieval, IEEE International Conference on Acoustics, Speech, and Signal Processing, 3668–3671, 2002

[6] Ffmpeg. https://www.ffmpeg.org/. Dostopno 13.07.2018.

Analiza kakovosti načrtovanja digitalnega radijskega omrežja

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

Analisis of the Digital Radio Network Planning

Abstract. *Planning and construction of a radio network is always followed by a process that never ends - optimization. Optimization is a process that eliminates shortcomings of insufficient or wrong input information, the differences between chosen model in planning tools and the actual behavior of the signal on field, not enough detailed digital terrain maps or inaccuracies in them, incorrect decisions of the planner or incorrect installation settings, new elements in capacity or functionality needs during network operation and others.*

In this article we analyze planning performance of a digital mobile radio network (DMR) [1, 2, 3, 4] that is operating in the frequency range from 146 to 174 MHz. This technology is used also for the Public Protection and Disaster Relief (PPDR) purposes [5, 6].

For the planning of PPDS network, the TAP tool with the Mapper module [7] was used, and field measurements with the Mototrbo Site Survey tool [8] were performed.

1 Uvod

Načrtovanju in izgradnji radijskega omrežja vedno sledi proces, ki se nikoli ne konča – optimizacija. Z optimizacijo skušamo odpraviti izzive, ki se pojavijo zaradi nezadostnih ali slabih vhodnih informacij ob načrtovanju, razlik med modeli v planerskih orodjih in dejanskim obnašanjem signala, ne dovolj podrobnih digitalnih kart terena ali njihovih netočnosti, napačnih odločitev planerja ali napačnih nastavitev ob montaži, spremenjenih potrebah po zmogljivosti ali funkcionalnosti med delovanjem omrežja in drugih.

V tem članku analiziramo uspešnost načrtovanja digitalnega mobilnega radijskega omrežja (DMR) [1, 2, 3, 4], ki deluje na frekvenčnem območju od 146 do 174 MHz. To omrežje se uporablja tudi za potrebe javne varnosti, zaščite in reševanja (ang. Public Protection and Disaster Relief - PPDR) [5, 6].

Za potrebe načrtovanja pokrivanja je uporabljeno programsko orodje Terrain Analysis Package (TAP) z modulom Mapper [7], za analizo dejanskega pokrivanja so se izvedle terenske meritve z orodjem Mototrbo Site Survey [8].

2 Načrtovanje

Omrežja PPDR so omrežja varnosti in zaščite, zdravstvene pomoči, nadzora in vzdrževanja elektroenergetskih omrežij, avtocest ali drugih, ki služijo varovanju človeških življenj ter nadzoru in vzdrževanju javne infrastrukture.

Zanesljivost delovanja takšnega omrežja je izjemnega pomena, zato mora biti tudi načrtovano tako, da je njegova razpoložljivost vsaj 99,99 %, imeti mora dovolj veliko zmogljivost in rezervo, pokrivanje s signalom mora biti skoraj popolno, vzpostavljanje komunikacije mora biti hitro in z majhnimi zakasnitvami.

Način izgradnje omrežja PPDR se razlikuje od komercialnega javnega mobilnega omrežja, sploh če je grajeno namensko. Zanj so običajno namenjena javna sredstva, ki jih ni veliko, prav tako ni niti časa niti sredstev za kasnejšo optimizacijo. Zato je zelo pomembno, da so rezultati teoretičnega načrtovanja, izdelanega s programskim orodjem in po katerem gradimo omrežje, čim bliže tistemu, kar bomo dobili v realnosti.

2.1 Izbira modela

Za načrtovanje pokrivanja z radijskim signalom se uporablja veliko modelov, ki upoštevajo parametre, kot so klimatske razmere, vrsta in oblika terena, prevodnost in dielektrična konstanta tal in podobno, od katerih so nekateri zajeti v tabeli 1, pogosteje uporabljeni modeli pa so podrobneje opisani v tabeli 2.

Tabela 1. Modeli za računanje razširjanja signala nad terenom

Model razširjanja	Primeren za frekvenčno območje
Okumura Hata	150 MHz – 1,5 GHz
Longley-Rice	30 MHz – 40 GHz
Za areonavtiko	30 MHz – 30 GHz
Model ravne Zemlje	30 MHz – 10 GHz
Egli,-urbana naselja	30 MHz – 10 GHz
HCM	30 MHz – 2 GHz
ITU 370	30 MHz – 1 GHz
ITU 567	30 MHz – 1 GHz
CEPT	30 MHz – 250 MHz

Tabela 2. Pogosteje uporabljeni modeli razširjanja signala

Model razširjanja	Opi modela
Model Okomura	Pri tem modelu se izračuna mediana jakosti polja, ki se potem popravi zaradi vpliva razgibanosti terena, vrste terena (urbano, podeželje,…), nagiba terena, usmerjenosti ulic in podobno

	Ta model je najbolj uporaben za računanje v urbanih območjih (osnovni model je bil narejen za Tokio), dodane so korekcije za podeželje in odprt teren. Krivulje v tem modelu so primerne za razdalje od 1 do 100 kilometrov.
Model Hata-Davidson	Ta model uporablja modificirane empirične formule, dobljene iz družine krivulj Okomura. Izračuni so odvisni od vrste okolja (urbano, veliko mesto, majhno mesto, primestje, podeželje). Ne vsebuje nobenih posebnih korekcij, odvisnih od poti širjenja signala med oddajnikom in sprejemnikom.
Longley-Riceov model	Poleg osnovnih parametrov, kot so frekvenca, moč oddajnika, podatki o antenskem diagramu iz katerih se izračuna jakost polj v praznem prostoru, se pri tem modelu uporabijo še podatki o klimatskih posebnostih, radijskem horizontu, prevodnosti tal in podobno. To je najpogosteje uporabljen model v frekvenčnem območju od 30 MHz do 30 GHz.
Model Eglija	Ta model je razvilj John Egli iz izmerjenih vrednosti na skoraj ravnem terenu (povprečna višina ovir do 20 m) do razdalje 80 kilometrov. Za izračun so potrebni le frekvenca, razdalja ter višina oddajne in sprejemne antene. Model ne upošteva podatkov o razgibanosti terena. Preprost model, primeren za izračunavanje na ravnem terenu.
Bullingtonov model	Pri tem modelu se izračunane vrednosti jakosti polja za prazen prostor popravijo za slabljenje, ki je posledica difrakcije signala na ostri oviri. Ta model se uporablja za izračunavanja linkovskih in drugih mobilnih zvez, predvsem takrat, ko načrtujemo pokrivanje terena iz višinskih točk.

Izberemo tistega, ki je najprimernejši za obliko terena in frekvenco, ki jo uporabljamo. Izkustveno je za naše frekvenčno območje (146 MHz – 174 MHz) in vrsto terena (zunanje pokrivanje) najprimernejši model *Longley-Rice* [9].

Vhodni podatki so: nosilna frekvenca, efektivna izsevana moč (ang. effectife radiated power - ERP), antenski smerni diagram in višina antene nad terenom.

2.2 Izračun

Pri izračunu s programskim orodjem *Terrain Analysis Package - TAP* se upošteva klimatska karta in lastnosti geografskega območja. Pri načrtovanju pokrivanja smo za podlago uporabili digitalni model terena Slovenije, ki ima ločljivost 30 metrov. Orodje računa po izbranem modelu. Izračuni so narejeni za radialno območje 360° okoli lokacije v kotnih razmikih po 0,5° in do razdalje 30 kilometrov. Računa se z zahtevnostjo 90/90, kar statistično pomeni pokritost s signalom 90 % časa na 90 % lokacij ali boljše. To je zelo strog kriterij, ki velja za delovanje kritičnih radijskih zvez, kamor spadajo npr. službe zaščite in reševanja.

Pri načrtovanju pokrivanja terena računamo z naslednjimi osnovnimi parametri:
- frekvenco (f),
- valovno dolžino (λ)
- močjo oddajnika (P_o),
- dobitkom antene (G_o),
- močjo na sprejemniku (P_S),
- dobitkom sprejemne antene (G_S).

Po enačbi (1) se izračunava gostota pretoka moči S na razdalji d od oddajnika.

$$S = (P_o \cdot G_o)/(4\pi \cdot d^2) \quad (1)$$

Po enačbi (2) se izračunava sprejeta moč na sprejemniku mobilne postaje.

$$P_S = (P_o \cdot G_o) \cdot (\lambda^2 \cdot G_S)/(4\pi \cdot d)^2 \quad (2)$$

V enačbah (1) in (2) se uporabljajo naslednje enote: frekvenca v Hertzih, valovna dolžina v metrih, moči v vatih in dobitki anten brez enot.

Poenostavljena enačba zveze (3) med oddajnikom in sprejemnikom je

$$P_S = P_o + G_o + G_S - L_o - L_S - L_P \quad (3)$$

Vse enote v enačbi (3) so v decibelih.

Parameter L predstavlja izgube na poti med oddajnikom in sprejemnikom. Člen L_P je odvisen od prostora skozi katerega se širi signal (konfiguracije in sestave terena, vegetacije, umetnih in naravnih ovir …). Modeli, ki jih uporabljajo računska orodja, se razlikujejo po tem, kako upoštevajo vpliv prej omenjenih elementov na razširjenje signala.

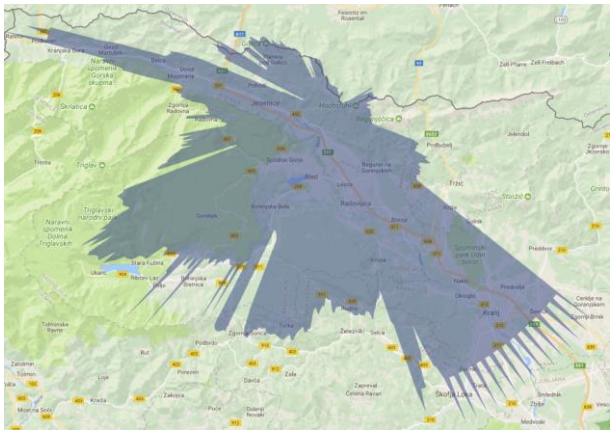
V našem primeru je izhodna moč oddajnika 10 W. Občutljivost bazne in mobilne postaje na priključni sponki sprejemnika je -117 dBm. Pri teh nivojih je pogostnost bitne napake BER (ang. Bit Error Rate) manjša ali enaka 5 %. Mejni vrednosti jakosti polja, s katerima še računamo s kakovostnim sprejemom pa sta za mobilno postajo -106 dBm in za bazno postajo -110 dBm.

Na sliki 1 je prikazano pokrivanje terena z lokacije Ravni pod Stolom. Obarvano polje prikazuje jakost signala, ki je boljša od -106 dBm.

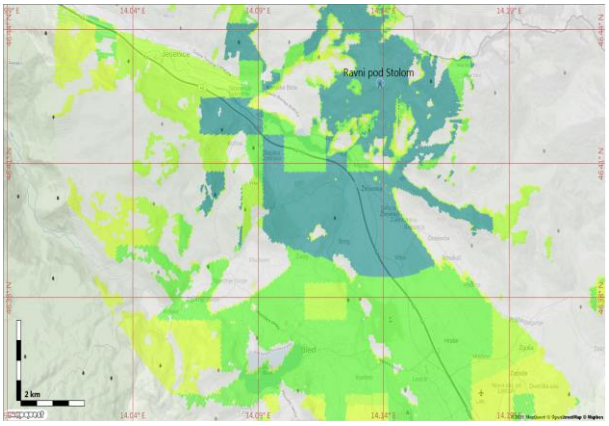
Za izračun smo izbrali model Longley-Rice z naslednjimi parametri: vertikalna polarizacija, efektivna ukrivljenost Zemlje 1.3333, klimatsko področje 5, relativna dielektričnost tal 15, prevodnost tal 0,005 S/m. Izračuni veljajo za 90 % situacij in 90 % časa.

Področje pokrivanja smo računali na radialih dolžine 10 kilometrov od lokacije, ki so si sledili po 0,5 kotne stopinje od azimuta 100 do 290 stopinj. Na vsakem radialu so se izračunavale točke električne poljske jakosti v razmikih po 50 metrov. Tudi po višinah smo nastavili resolucijo na 50 metrov.

Na sliki 2 so prikazane podrobne vrednosti polja, temnejša barva pomeni večjo jakost polja. Za bolj pregleden prikaz je bil izbran korak na grafičnem prikazu 5 dB.



Slika 1. Območje pokrivanja terena z lokacije Ravni pod Stolom s signalom boljšim od -106 dBm



Slika 2. Podrobnejši izračun pokrivanja (dBm)
(siva <-80, rumena = -75 svetlo zelena = -70, zelena = -65, temno zelena > -60)

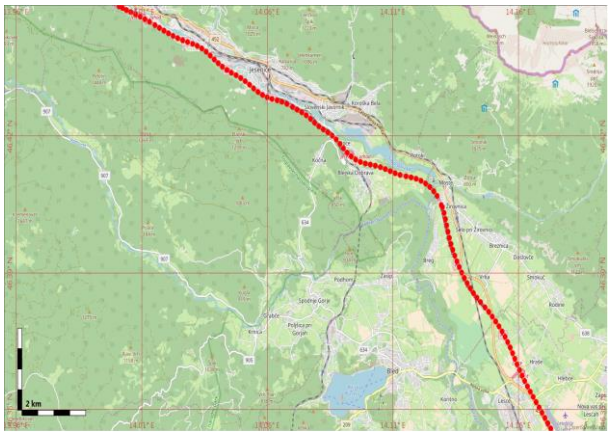
3 Testiranje/meritve

Testiranje omrežja se običajno izvaja na različnih segmentih, odvisno od namena. Če bi nas zanimala kompletna analiza signala, bi s spektralnim analizatorjem ali posebnim DMR-merilnikom merili tudi druge parametre po standardih ETSI [1 - 4], na primer: močnostni profil, frekvenčno napako, simbolno napako, napako modulacije, BER, konstelacijski diagram, kanal, barvno kodo in druge parametre.

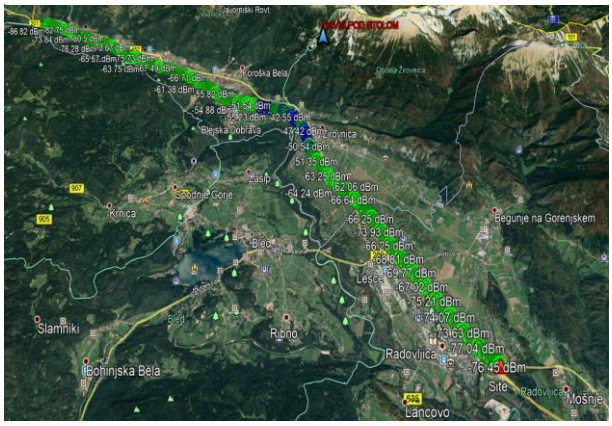
Za potrebe te analize se omejimo le na meritve jakosti sprejetega signala na izbrani trasi, hkrati pa beležimo še geografsko lokacijo s sprejemnikom globalnega sistema za pozicioniranje (ang. Global Positioning System - GPS).

Meritve smo izvajali po trasi gorenjske avtoceste, med Radovljico in Hrušico. V merilno vozilo smo namestili zunanjo sprejemno anteno, anteno GPS in mobilno radijsko postajo, ki je služila kot merilni instrument. Na postajo smo priklopili prenosnik z merilno programsko opremo. Vozilo se je premikalo s hitrostjo 60 km/h, vzorčili smo na vsakih 10 sekund (približna razdalja med posameznima točkama je 166 metrov). Višina antene nad terenom (na strehi vozila) je bila približno dva metra.

Na celotni trasi med Radovljico in Hrušico je bilo opravljenih 134 meritev. Za vsako točko je bil izmerjen podatek o geografski lokaciji in jakosti polja.



Slika 3. Trasa testnih meritev Radovljica-Hrušica (n = 134)



Slika 4. Rezultati testnih meritev (dBm)
(zelena = -105 do -50, modra > -50)

Na sliki 3 je z rdečo prikazana trasa meritev, na sliki 4 pa so podane izmerjene vrednosti jakosti polja na posameznih točkah. Iz slike 4 je razvidno, da je jakost signala na celotni trasi vedno boljša od -80 dBm, blizu oddajnika celo boljša od -50 dBm.

4 Analiza

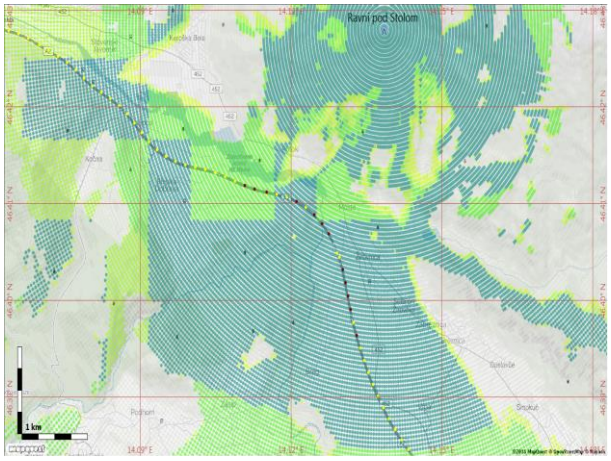
Analizo smo naredili z modulom programskega orodja *TAP6*. Vsi izračunani in izmerjeni podatki se lahko pretvorijo v format KML in so grafično berljivi v orodju *TAP Mapper* in aplikaciji *Google Zemlja*.

Izračun pokrivanja v našem primeru je bil narejen z modelom Longley-Rice, ki je imel nastavljene parametre kot je zapisano v poglavju 2.2.

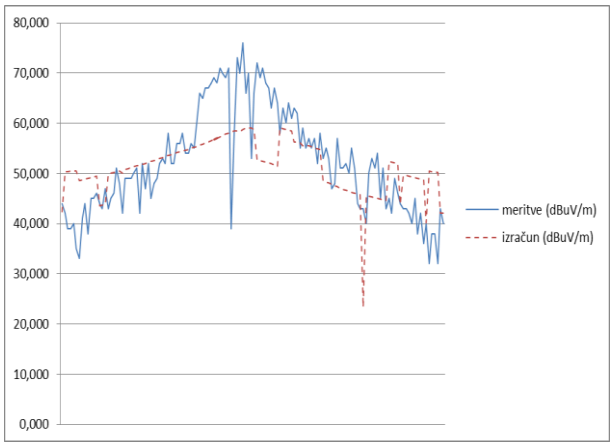
Primerjava izračunanih in izmerjenih vrednosti je prikazana na sliki 5 in v grafu 1.

Korak v grafičnem prikazu izračunane vrednosti polja je 5 dB (slika 2), prav tako je razločljivost topografskih podatkov 50 metrov. Tako grob raster je bil vzet zaradi sprejemljivo dolgega časa računanja. Bolj fina raster (npr. izračun polja za točko 10 x 10 m in razločljivost topografske karte 10 m) bi bistveno podaljšal čas računanja polj.

Primerjava izračunanih in izmerjenih vrednosti v grafu 1 kaže, da smo izbrali pravi model, vendar obstajajo manjša odstopanja, zaradi relativno velikega izbranega koraka v grafičnem prikazu.



Slika 5. Odsek meritev (črtkana krivulja) in izračunov (svetlo in temno zelene konture) za primerjavo



Graf 1. Primerjava izmerjenih (polna črta) in izračunanih (prekinjena črta) vrednosti

5 Sklep

V članku smo analizirali, kako se teoretično načrtovano omrežje obnaša v realnosti. Na izbrani trasi smo s testnim terminalom in ustrezno programsko opremo izmerili jakost sprejetega signala. Primerjava izračunane in izmerjene vrednosti (graf 1) kaže, da je za izbrani model (Longley-Rice) odstopanje med vrednostmi sprejemljivo. Za našo situacijo smo izbrali primeren model, v drugačnem okolju pa bi bil verjetno bolj primeren drug model (tabela 2). Podobno bi veljalo za drugo frekvenčno območje (tabela 1).

Ko načrtujemo npr. pokrivanje kritičnih odsekov ali področij, kjer smo zaradi geografskih ali umetnih ovir negotovi o kakovosti pokrivanja našega oddajnika, je koristno, da izračunamo poljsko jakost za večje število točk, v grafičnem prikazu pa nastavimo manjši korak (največ 3 dB).

Nadaljevanje te analize predvideva izbiro drugih modelov za izračunavanje razširjanja signala nad terenom in primerjavo med njimi. Slovenija je precej razgibana dežela, ki ima hribovito pokrajino, panonsko nižino, morje in srednje velika urbana naselja, kar pomeni, da so za različna območja primerni različni modeli. Sklepamo lahko, da bodo za druge pokrajine primernejši drugi modeli, podobno pa velja tudi za večja mesta.

Zahvala

Zahvaljujem se gospodu Jožetu Štufleku za nasvete in pomoč pri študiji programskega orodja TAP.

Literatura

[1] ETSI TS 102 361-1, Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) Systems; Part 1: DMR Air Interface (AI) protocol;

[2] Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) Systems; Part 2: DMR voice and generic services and facilities;

[3] Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) Systems; Part 3: DMR data protocol;

[4] Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) Systems; Part 4: DMR trunking protocol;

[5] ECC REPORT 102, Public protection and disaster relief spectrum requirements, Helsinki, 2007;

[6] Report ITU-R M.2377-0, Radiocommunication objectives and requirements for Public Protection and Disaster Relief (PPDR), 07/2015;

[7] TAP 6,Terrain Analysis Package, Softwirght LLC, USA, www.softwright.com

[8] MOTOTRBO Site Survey, ver. 04.03, Motorola Solutions, USA;

[9] P. L. Rice, A. G. Longley, K. A. Norton, and A. P. Barsis, "Transmission loss predictions for tropospheric communication circuits," U.S. Government Printing Office, Washington, DC, NBS Tech. Note 101, issued May 1965; revised May 1966 and Jan. 1967;

5G v ekstremnih razmerah

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 in extreme situations

Abstract. *The fifth generation of electronic communications (5G) covers much more than high data rates and low delays for modern communications. The complementarity of the 5G-compatible technologies enables the automation, control and modernization of the technologies of various industries that operate in human unfriendly environments This article covers the use of 5G technology in extreme situations, such as unfavorable working conditions (e.g. drilling of tunnels, mine work ...) and in natural and other disasters. Important techniques that enable the operation of mission-critical applications in the 5G network are also described. This article describes some research and pilot projects that gives promising results and are ready for implementation.*

1 Uvod

Peta generacija elektronskih komunikacij (5G) zajema precej več kot omogočanje visokih mobilnih hitrosti prenosa in majhnih zakasnitev za sodobne komunikacije. Komplementarnost tehnologij, ki se združujejo pod kratico 5G, omogoča tudi avtomatizacijo, krmiljenje in posodobljenje tehnologij različnih gospodarskih panog, ki delujejo v človeku neprijaznih okoljih in razmerah. Ta članek zajema uporabo tehnologije 5G v omenjenih ekstremnih razmerah ter ob naravnih in drugih nesrečah. Povzete so pomembnejše tehnike, ki v omrežju 5G omogočajo delovanje zahtevnih aplikacij. V nadaljevanju so opisane nekatere raziskave in pilotski projekti, ki že dajejo prve rezultate.

2 Ekstremne razmere

Ekstremne razmere lahko definiramo na več načinov, gotovo pa so to človeku in napravam neobičajne razmere, ki so lahko prisotne ves čas (npr. delo pod zemljo, pod vodo, v puščavi itd.) ali pa se pojavijo občasno kot posledica naravnih nesreč (kot so neurja, potresi, poplave …).

Po splošni definiciji iz [1] zahteva komunikacija v ekstremnih razmerah hitro vzpostavitev zvez, majhne zakasnitve, visoko razpoložljivost in zanesljivost delovanja, zmožnost nudenja storitev velikemu številu uporabnikov in naprav, visoko stopnjo varnosti in prednostno komunikacijo.

3 3GPP in komunikacije v ekstremnih razmerah

Združenje 3GPP [2] je že pred sprejetjem različice 13 standarda 3GPP postavila temelje za komunikacije v ekstremnih razmerah, ko je standardizirala funkciji eMBMS (Enhanced Multimedia Broadcast/Multicast) in GCSE (Group Communication System Enablers), kmalu za tem pa tudi ProSe (Proximity Based Services) za direktno komunikacijo med končnima napravama.

V različici 13 standarda 3GPP je bila standardizirana funkcija MCPTT (Mission Critical Push-to-Talk) in kasneje v različici 14 še MCData (podatkovne komunikacije v realnem času) in MCVideo (video komunikacije v realnem času).

Splošne komunikacije v ekstremnih razmerah tako zajemajo tri skupine: govorne komunikacije "pritisni-in-govori" (MCPTT, Mission Critical Push-to-Talk), video komunikacije v realnem času (MCVideo) in prenos podatkov v realnem času (MCData).

V juniju 2018 je bila zaključena prva faza standardov 5G (3GPP-različica 15), ki zajema naslednje funkcionalnosti za komunikacije v izrednih razmerah:

- povezavo med vsemi že določenimi funkcijami v 3GPP za kritične komunikacije,
- medsebojno povezljivost med LTE in obstoječimi sistemi javne varnosti, kot sta TETRA in P25, za govorne komunikacije in kratka sporočila,
- zahteve za storitve izrednih razmerah, ki se uporabljajo v železniškem prometu,
- zahteve za storitve izrednih razmerah, ki se uporabljajo v pomorskem prometu,
- vmesnike za uporabo funkcije eMBMS v izrednih razmerah.

4 Pomembni elementi komunikacijskega omrežja v ekstremnih razmerah

V ekstremnih razmerah, kjer je varnost ljudi (in infrastrukture) na prvem mestu, mora komunikacijsko omrežje delovati izjemno zanesljivo (brez izpadov) in hitro (majhne zakasnitve) [3].

Glavnina razvoja v omrežjih, ki delujejo tudi v ekstremnih razmerah, poteka na naslednjih segmentih [4]:

- zanesljivost radijskega dostopa,
- programsko nastavljiva zgradba omrežja,
- neposredna komunikacija med napravami (D2D).

Skoraj popolno zanesljivost na radijskem dostopovnem delu dosežemo z vpeljavo dodatnih poti istega signala, večkratno hkratno povezljivostjo,

omejevanjem motenj in izboljšanimi mehanizmi za ponovna oddajanja informacije. Z vpeljavo prilagodljive zgradbe radijskih okvirov pa dosežemo zakasnitve reda nekaj milisekund.

4.1 Radijski dostop

Merilo za kakovostno zvezo oz. zanesljivost zveze je zadostno razmerje med koristnim signalom in motnjo s šumom (SINR, Signal-to-Interference and Noise Ratio). To razmerje povečamo tako, da povečamo moč signala ali pa znižamo motnje z eno od metod:

- *mikroskopska raznolikost* – več anten omogoča boljše pokrivanje. Tehnika uporabe več anten pri oddaji in sprejemu (MIMO, Multiple Input Multiple Output) omogoča povečanje razmerja SINR. Pri 5G se uporabljala tehnika masovnega MIMO – uporaba matrike nekaj deset anten;
- *makroskopska raznolikost* – lahko jo izvedemo na več načinov: (a) s kombinacijo radijskih dostopovnih tehnologij (npr. 4G in 5G) ali (b) s kombinirano zvezo (sprejem/oddaja) preko več hkratnih celic in baznih postaj. Pri slednji več baznih postaj oddaja iste podatke, ki jih sprejemnik ustrezno združi.

Za uporabo v ekstremnih razmerah je še posebno pomembna uporaba načina makroskopske raznolikosti. Pri običajnih predajah zvez med celicami omrežja 4G lahko pride do zakasnitev velikosti 55 milisekund, kar je za kritične komunikacije nesprejemljivo.

Motnje se lahko odpravljajo ali na omrežju ali na terminalu. Po pravilu *čez palec* velja, da če odpravimo motnje, ki jih povzročata dva najmočnejša sosednja motilnika (bazni postaji ali terminala), smo kolikor je mogoče izboljšali razmerje SINR.

4.2 Majhne zakasnitve

Stroji (naprave) lahko sprejemajo, analizirajo in izvajajo naloge bistveno hitreje kot ljudje. To je razlog, da komunikacije med napravami zahtevajo zelo majhne zakasnitve (npr. povratne zanke v industrijski avtomatizaciji), ki so običajno manjše od ene milisekunde. Visoka zanesljivost delovanja ($BER<10^{-9}$) pa je pomembna pri sinhronizaciji. Pri komunikaciji naj bi bila režija čim manjša, da bi se pri prenosu majhnih paketov ohranila spektralna učinkovitost [4].

Upravljanje prometa in premikanje opreme v proizvodnji imata tipične časovne cikle od ene do 10 milisekund. Največje zahteve imajo nekateri regulatorji in senzorji, ki zahtevajo krajši čas od ene milisekunde z odstopanjem manj kot mikrosekunda.

4.3 Daljinsko upravljanje s haptičnim odzivom

Daljinsko upravljani roboti in druge naprave nam omogočajo, da varno delamo v okoljih ali s stvarmi, ki so lahko življenjsko nevarne ali drugače škodljive.

Na primer bolnišnice lahko organizirajo daljinsko vodene operacije brez fizične prisotnosti enega ali več kirurgov, daljinsko vodena mehanizacija lahko vrta predore ali koplje rudo globoko pod zemljo, roboti opravljajo razminiranje ali gasijo požare.

Takšen sistem mora biti izjemno zanesljiv ($BER < 10^{-9}$), zakasnitev med končnima točkama pa manjša od milisekunde, da imamo pravi občutek – haptični odziv – pri delu z oddaljeno napravo.

Danes so na trgu naprave z zasloni na dotik, ki dajejo uporabniku določen povratni odziv. Tudi ob delu na daljavo mora imeti uporabnik podoben občutek – čutiti dotik. Eden od primerov je pametni telefon, ki daje t.i. 3D-občutek pritiska.

Na primer, kombinacija haptične interakcije in 360-stopinjskih kamer, ki prenašajo dogajanje v živo preko omrežja 5G v naprave, nameščene na glavi, bo dala izjemno uporabniško izkušnjo – kot je operacija na daljavo z aktivnim sodelovanjem.

4.4 Programsko nastavljiva omrežja

Za zanesljivo omrežje z majhnimi zakasnitvami moramo imeti dovolj odporno omrežje z dinamičnim upravljanjem, ki bo pripeljalo vsebino blizu uporabnikom.

Ključne izboljšave v zgradbi omrežja je mogoče doseči z deljenjem omrežja na rezine, kar omogoča namenska virtualna mobilna omrežja z določenim naborom omrežnih virov in hkrati zagotavlja dogovorjeno kakovost storitve. Deljenje na rezine je nujno za podporo vertikalnim industrijam. Včasih smo gradili omrežja za določeno rabo, npr. 2G za govorne komunikacije ali 4G za podatkovne komunikacije. Zasnova omrežja 5G pa je takšna, da se lahko prilagodi vsakemu primeru posebej. Znotraj omrežja je lahko več hkratnih navideznih omrežij, ki se med seboj ne motijo in služijo vsako svojemu namenu. Primer soobstoja dveh navideznih omrežij sta npr. eno navidezno omrežje, ki upravlja z avtonomnimi vozili (v omrežju so pomembne majhne zakasnitve) in drugo navidezno omrežje za video prenose kakovosti 4K ali boljše, ki pa zahteva zelo velike prenosne hitrosti.

Od programsko nastavljivih omrežij pričakujemo, da se hitro prilagajajo različnim zahtevam. To pomeni, da pot paketov po omrežju ni vnaprej določena s fiksno arhitekturo omrežja in se jo sproti prilagaja, da dobimo ustrezne zakasnitve. Programsko opredeljena omrežja omogočajo v hrbteničnem delu omrežja, zbirnih točkah in zalednih povezavah optimizacijo prometa, pasovne širine in zagotavljanje potrebne procesorske moči na robu omrežja tako, da so zakasnitve najmanjše. Transportni agenti v samoorganiziranih omrežjih zbirajo podatke o zakasnitvah, izgubah, prepustnosti, dolžini čakalne vrste, aktivnih nosilnikih ali aktivnih napravah in te podatke ustrezno uporabijo za krmiljenje programirljivega omrežja.

Razpoložljivost omrežnih elementov mora biti visoka. To dosežemo z združevanjem jedrnih elementov in uravnoteženjem obremenitev. S tem zagotovimo delovanje sistema brez prekinitev, tudi če eden ali več jedrnih elementov odpove. Podobno velja, če odpovejo nekatere zaledne povezave – storitve bodo v samostojnem načinu delovanja delovale skoraj neokrnjeno.

S pomikanjem prehodov in aplikacijskih strežnikov v bližino radijskega dela (v oblak na mobilnem robu) lahko dodatno zmanjšamo zakasnitve. Tako storitve niso več odvisne od točno določenih povezav med končnima točkama, pač pa se pot za povezljivost oblikuje glede na dejanske potrebe konkretne storitve. S takšnim modelom povezovanja 'vsakega z vsakim', kjer komunicirajo naprave neposredno preko lokalnih strežnikov na ravni radijskega dostopovnega dela, se izognemo prenašanju podatkov do centraliziranih prehodov – dosežemo najkrajše in najboljše poti za usmerjanje prometa. Zakasnitve so najmanjše možne in sistem stabilno deluje.

4.5 Neposredna komunikacija med napravami

Pri neposredni komunikaciji med končnima napravama (D2D, Device-to-Device) za prenašanje podatkov (tudi govora) ne uporabljamo vmesne infrastrukture.

Ta način komunikacije v omrežjih 5G omogoča veliko zanesljivost delovanja in zelo majhne zakasnitve. Ker za obdelavo podatkov nimamo vmesnih omrežnih elementov ali transportnega dela omrežja, so zakasnitve manjše, izboljša pa se tudi zanesljivost zvez – razširjena pokritost z omrežjem.

Omrežje neposrednih povezav se lahko uporabi tudi kot rezervna ad-hoc rešitev v primeru okvar omrežne infrastrukture ali ko ta kako drugače ni na voljo.

Dodeljevanje radijskih virov za neposredne komunikacije med napravami je lahko *porazdeljeno* ali *centralizirano*. V primeru porazdeljene rešitve lahko naprave prenašajo podatke takoj in za to ni potreben poseben postopek ali sodelovanje posebnih krmilnih kanalov. Če imamo centralizirano rešitev, se sicer učinkoviteje izogibamo trkom, vendar postane rešitev bolj kompleksna in povečajo se povprečne zakasnitve.

Način upravljanja radijskih virov ponazorimo na primeru komunikacije med vozili v pametnem prometu. Vsa vozila občasno oddajajo informacije o svoji lokaciji, hitrosti, smeri gibanja itd. Ko se promet zgosti, se poveča tudi verjetnost trkov. Ena od možnosti dodeljevanja radijskih virov je izbira dostopa na osnovi geografske lokacije, ki jo pošilja vozilo. Cesta je po vsej dolžini razdeljena na odseke, tako da se vedno dodeli ustrezen radijski vir glede na informacijo o lokaciji vozila.

Neposredno komunikacijo med napravami se lahko izkoristi tudi tako, da ima naprava hkrati dve zvezi – eno neposredno z drugo napravo in drugo preko omrežja. Vedno pa izbira optimalno zvezo za prvi ali ponovljeni prenos podatkov in kombinira sprejem signala preko različnih povezav.

Neposredna zveza ima lahko korist tudi iz uporabe načina polne dvosmerne zveze (FD, Full Duplex) – sprejem in oddaja hkrati potekata na istem frekvenčnem pasu. Tako se izognemo potrebi po vnaprej nastavljenim smerem prenosa. Zato je potrebna zelo majhna kontrolna režija, kar seveda zniža zakasnitve. Način dela v polnem dvosmernem prenosu zahteva določene tehnične rešitve na anteni, da oddani signal ne moti lastnega sprejema.

5 Raziskovalni in pilotski projekti

Industrija ima močne zahteve po hitrih komunikacijah med napravami (M2M) in izjemno zanesljivih povezavah. Morebitne sistemske napake pomenijo okvaro opreme, ustavitev proizvodnje ali v skrajnem primeru hude poškodbe ali celo smrtne žrtve.

Za tovarne v prihodnosti bo ključna časovna optimizacija procesov. Tu pride do izraza potreba po izjemni zanesljivosti in skoraj ničelnih zakasnitvah – npr. stalna optimizacija proizvodnega procesa, ki temelji na senzorskem nadzoru vseh delov procesa v realnem času ali pa stalna povezanost med roboti nove generacije.

5.1 5G v rudarski industriji

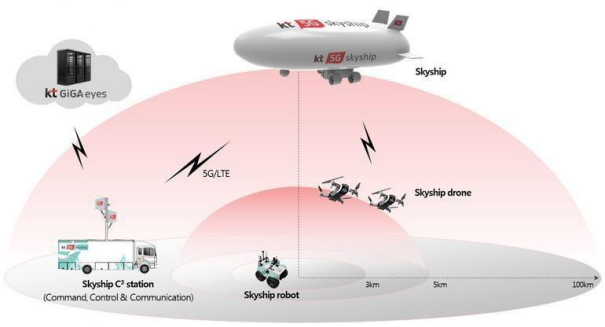
Primer združevanja tehnologij pete generacije več partnerjev in vodenja težke mehanizacije je projekt PIMM (Pilot for Industrial Mobile communications in Mining) [5], ki so ga zagnali leta 2015 na Švedskem.

Cilj tega projekta je uporaba nove mobilne tehnologije (4G, 5G) v industriji zaradi izboljšanja produktivnosti in povečanja varnosti pri delu v rudnikih. V sklopu projekta preizkušajo napredne metode avtomatizacije in daljinskega nadzora opreme, ki je montirana na mehanizaciji v rudniku Boliden na Švedskem. Z uporabo komunikacijske, senzorske in druge tehnologije je mogoče na daljavo, iz udobne lokacije na površju, krmiliti mehanizacijo v rudniških rovih globoko pod zemljo.

Preizkušajo štiri ključne segmente: industrijsko mobilno dostopovno omrežje, industrijsko mobilno jedrno omrežje, dostopovno točko za bližnje komunikacije in robustno komunikacijsko platformo za avtomatizacijo procesa.

5.2 5G pri iskanju in reševanju

Za iskanje pogrešanih oseb in reševalne akcije iz zraka je korejski operater KT razvil platformo [6], imenovano *Skyship*, ki jo sestavljajo zrakoplov, droni, roboti in center vodenja na zemlji. Slednji daljinsko nadzira drone in robote, ki izvajajo iskanje in reševalne operacije. Reševalci na terenu so opremljeni z očali za obogateno resničnost in imajo neposredno komunikacijo z zdravniki v bližnji bolnišnici, kar omogoča takojšnjo asistenco.



Slika 1. Platforma Skyship za iskalne in reševalne akcije (vir: ITU News)

Platforma je trenutno operativna v omrežju 4G LTE, prihodnje leto pa naj bi zaživela tudi na omrežju 5G. Platformo Skyship (slika 1) sestavljajo naslednje komponente:

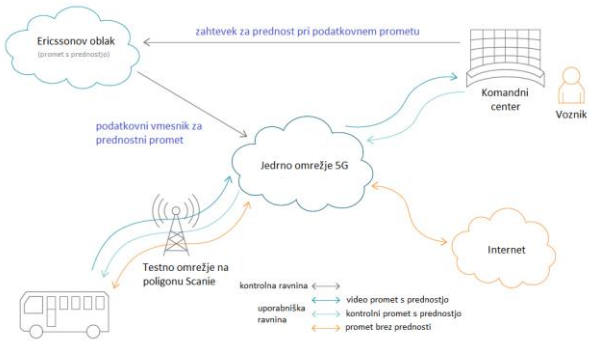
- s helijem napolnjen zrakoplov, ki pokriva večje območje do 100 kilometrov,
- komandna, kontrolna in komunikacijska mobilna zemeljska postaja (C^3),
- droni v zraku in roboti na zemlji, ki opravljajo reševalne akcije v obsegu nekaj kilometrov.

Zrakoplov je opremljen s sistemi za krmiljenje zrakoplova, kamerami, mrežnimi elementi in droni ter enoto, imenovano *skyscan*. *Skyscan* uporablja zelo majhno napravo LTE, ki zaznava signale pametnih terminalov pogrešancev v območju 50 metrov. Prejete signale primerja z bazo uporabnikov mobilnih operaterjev in tako identificira osebe. Droni nato odkrijejo točne lokacije iskanih oseb, roboti na zemlji pa jim odpeljejo osnovne stvari za prvo pomoč, dokler ne prispejo reševalci.

Zrakoplov lahko leti do šest ur s hitrostjo do 80 km/h in nosi opremo v skupni teži do 5 kg.

Zdaj je v uporabi druga različica platforme, ki uporablja omrežje 4G, naslednja različica (načrtovana za leto 2020) pa bo uporabljala omrežje 5G. Za komunikacijo bo uporabljen pas širine 100 MHz na frekvenčnem območju 3,5 GHz in 800 MHz na frekvenčnem območju 28 GHz, kar bo omogočalo prenos videa v realnem času v razločljivosti 4K UHD. Zelo majhne zakasnitve bodo reševalnim ekipam omogočale učinkovito delo.

5.3 5G v transportu



Slika 2. Platforma Skyship za iskalne in reševalne akcije (vir: Ericsson Technology Review, 2018)

Z namenom varnega vodenja samovozečih vozil preko omrežja 4G/5G podjetji Ericsson in Scania razvijata platformo za daljinsko vodenje avtobusov [7]. Za to je pomembna hitra in dovolj velika množica podatkov, ki so skoraj v trenutku na voljo operaterju. Podatki iz senzorjev, nameščenih po vsem avtobusu, se pretakajo v oddaljeni center vodenja preko radijskega dostopovnega omrežja 4G LTE in jedrnega omrežja 5G. V obratni smeri se komande za vožnjo avtobusa pošiljajo iz centra preko Scaninega komandnega sistema.

Kot je prikazano na sliki 2, imajo določeni nizi podatkov prednost, da je zagotovljena določena kakovost storitve. To so podatki, ki prihajajo iz video kamer in komande, ki na daljavo krmilijo avtobus. Vsi ostali podatki, t.i. infotainment, imajo nižjo prioriteto. Za zagotovitev prednosti uporabljajo identifikatorje QCI (QoS Class Identifier), ki izhajajo iz standardov 3GPP. Tako so identifikatorji za video prenos in podatke iz senzorjev razreda 5 in 2, infotainment (nižja prioriteta) pa ima QCI razreda 9.

Sistem deluje tako, da aplikacija v oblaku dinamično vzpostavlja povezave med vozili in paketnim jedrnim omrežjem 5G (EPC, Evolved Packet Core). Vsaka povezava ima nastavljene določene parametre kakovosti, kot sta največja dovoljena zakasnitev in najmanjša prepustnost.

6 Sklep

Ta prispevek opisuje lastnosti omrežja 5G, ki so ključne za njegovo delovanje v ekstremnih razmerah, te pa se lahko dnevno pojavljajo v nekaterih industrijskih panogah ali pa jih povzročijo nenadne vremenske spremembe.

Najpomembnejši lastnosti takšnega omrežja sta velika zanesljivost delovanja omrežja in zelo majhne zakasnitve v komunikacijah med dvema končnima napravama. Majhne zakasnitve in zanesljivost delovanja omrežja dosežemo z ustreznimi rešitvami v radijskem dostopovnem delu, izvedbi programsko nastavljivega omrežja in z vpeljavo neposredne komunikacije med napravami.

V svetu že vsaj tri leta potekajo intenzivna testiranja določenih funkcionalnosti omrežij 5G za industrijske in varnostne namene. V polnosti bodo zaživela verjetno šele čez nekaj let. Nekatere funkcije, ki so že standardizirane za omrežja 4G in so uporabne v izrednih razmerah (množične nesreče), pa že danes uspešno delujejo.

Literatura

[1] 3GPP TS 22.280, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mission Critical Services Common Requirements (MCCoRe); Stage 1, (Release 14);

[2] Mission Critical Services in 3GPP, www.3gpp.org

[3] T. Mlinar: 5G za kritične komunikacije, Zbornik referatov 32. delavnice o telekomunikacijah VITEL – Pametna omrežja informacijske družbe, 16. in 17. maja 2016, Brdo pri Kranju

[4] 5G for Mission Critical Communication, Nokia White Paper, 2016;

[5] https://www.sics.se/projects/firstpimmproject (PIMM - Pilot for Industrial Mobile communications in Mining);

[6] https://news.itu.int/kt-skyship-search-rescue-platform/ (Skyship – serch-and-rescue airborn platform);

[7] 5G Network Programmability, Ericsson Technology Review, January 26, 2018

Predlog nove tehnike sledenja izvoru signala

Peter Miklavčič[1], Tomaž Javornik[2], Boštjan Batagelj[1]

[1]*Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana*
[2]*Institut "Jožef Stefan", Jamova cesta 39, 1000 Ljubljana*
E-pošta: peter.miklavcic@fe.uni-lj.si

A new signal source tracking technique proposal

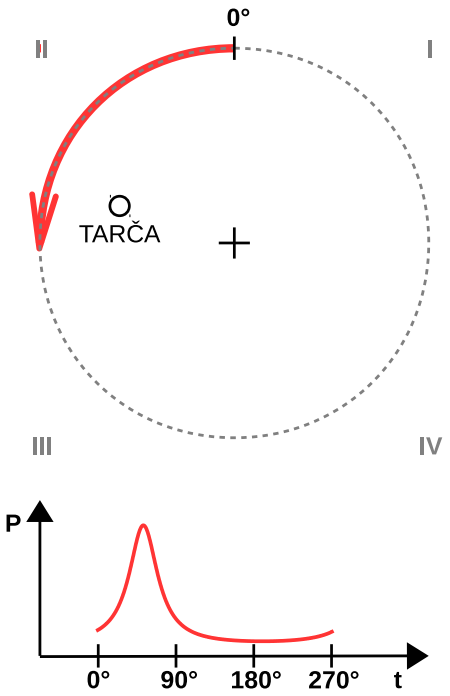
In this paper we present a novel approach in signal source tracking techniques which is theoretically applied to lobe switching and the monopulse principle. In the beginning, the main families of current signal source tracking techniques are shortly presented, namely step- and smart-track, lobe switching (sequential lobing), conical scanning and coherent monopulse operation by using an antenna group or a multimode antenna. Receiver block diagrams are also presented, for known methods as well as for the proposed three-lobe monopulse operation. Lobe switching is explained as a discretized version of conical scanning and in the case of two-dimensional operation, based on the conical scanning *nomenclature, the* pyramid scanning *term is proposed (*triangular pyramid scanning *in the case of proposed three-lobe technique).*

1 Uvod

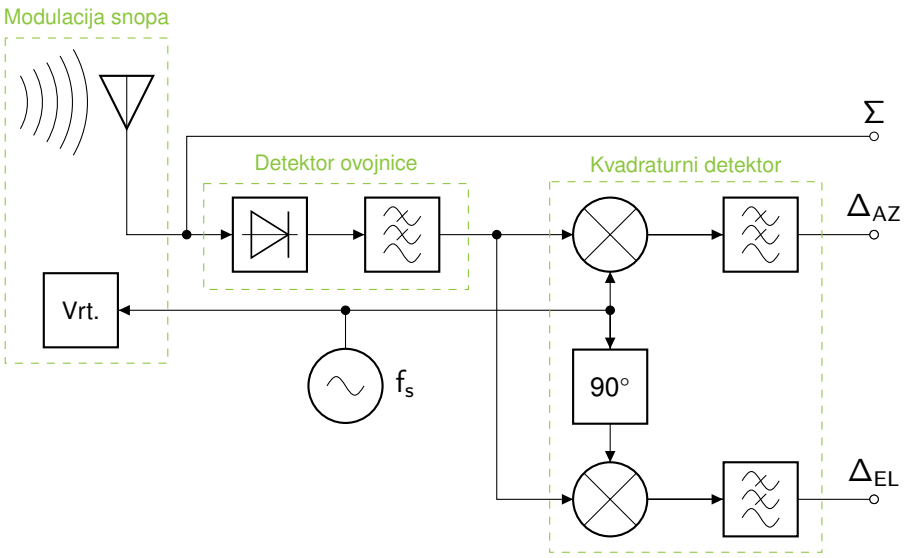
Z različnimi tehnikami sledenja izvoru signala [1, 2, 3, 4, 5, 6] je iz prostorsko odvisne fizikalne veličine smernega diagrama možno izluščiti dvodimenzijsko informacijo odklona vpadnega signala od referenčne smeri (ponavadi smeri glavnega snopa uporabljene antene oziroma antenske skupine). S pridobljeno informacijo je možno popravljati smer sprejema in s tem ohranjati zmogljivost zveze in/ali slediti smeri izvora (v radarski tehniki se uporablja izraz namerjanje tarče [1]). Sledenje izvoru signala se poleg radarske tehnike in satelitskih komunikacij uporablja tudi pri sistemih navigacije in pozicioniranja, kot npr. ILS (ang. *Instrument Landing System*), za vzdrževanje povezav med premikajočimi se vozlišči terenskih radijskih omrežij, za vzdrževanje (ali celo kot dodatno fizično zaščito) komunikacijskega kanala z letalniki, ipd.

2 Tehnike sledenja izvoru

Najpreprostejše tehnike sledenja izvoru so izvedene s fizičnim premikanjem celotne antene in tipanjem moči [1, 6, 7]. Pri antenah, namenjenih izključno sledenju, se splača tipati izrazitejši minimum, kot pri lovu na lisico. Tak princip se uporablja npr. v radarski tehniki ali pri napravah za odkrivanje (ponavadi neželenih) oddajnih naprav. Če pa je sistem namenjen prenosu informacij, se med delovanjem išče maksimum za optimalen sprejem. Pri satelitski televiziji se npr. uporabljata komercialna standarda DiSEqC in USALS, ki omogočata komunikacijo med sprejemnikom in aktivno anteno za ugotavljanje grobe lokacije na podlagi sprejemanih satelitov (za predvidevanje smeri želenega satelita) ter končno namerjanje s tipanjem moči sprejetega signala (ang. *step-track*), tudi v kombinaciji s predvidevanjem (ang. *smart-track*). Ta družina tehnik je danes večinoma izvedena v programskem delu. Sam princip sledenja ne zahteva posebnih anten (kvečjemu motorizirane podstavke), s čimer je omogočeno grobo, a poceni namerjanje velikega števila anten (npr. satelitska televizija, internet). Za hitrejše popravke in/ali natančnejše sledenje pa obstajajo tehnike sledenja, ki zahtevajo svojevrstne izvedbe anten in sprejemnih verig. Glavne so predstavljene v naslednjih poglavjih, medtem ko bolj eksotičnih, kot je npr. hibrid med stožčastim skaniranjem in monopulzno tehniko *conoscan* [8], članek ne zajema.



Slika 1: Pri stožčastem skaniranju antena z vrtečim se snopom skanira prostor okoli tarče, iz maksimuma sprejete moči signala pa sistem določi njeno smer.



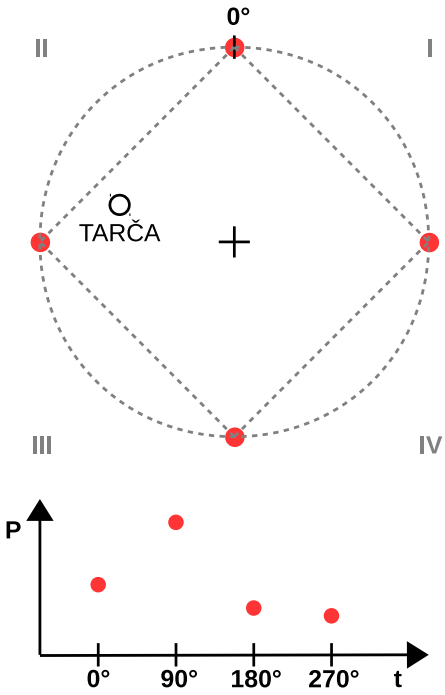
Slika 2: Blokovni načrt sprejemnika pri tehniki stožčastega skaniranja. Kadar je tarča namerjena (v osi stožca), je amplituda konstantna in signala razlike sta nič. f_S je frekvenca skaniranja oziroma vrtenja snopa.

2.1 Stožčasto skaniranje

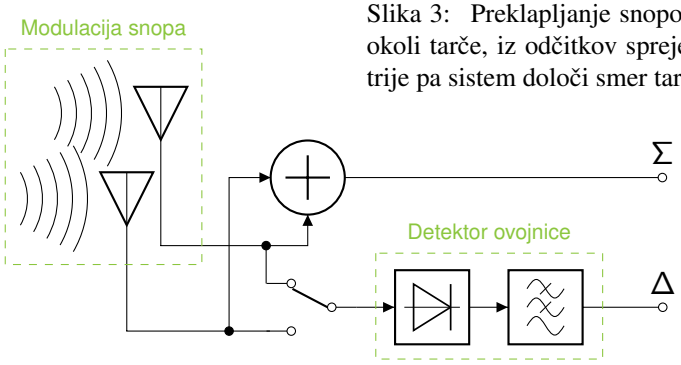
Stožčasto skaniranje (ang. *conical scanning*) vrti od referenčne smeri odklonjen snop (kot je prikazano na sliki 1), ki v prostoru okoli tarče orisuje stožec [1, 4, 8]. Kadar je tarča v osi stožca (namerjena), je odčitana moč v celotni periodi konstantna, sicer pa doseže maksimum, ki določa smer tarče. Vrtenje snopa je lahko izvedeno električno ali mehansko. Blokovni načrt sprejemnika je na sliki 2.

2.2 Preklapljanje snopov

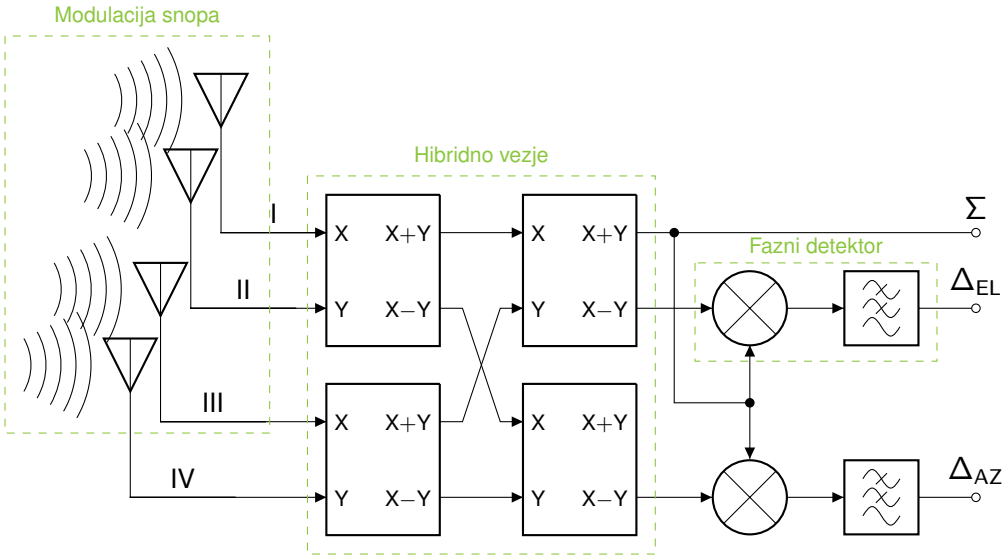
Preklapljanje snopov (ang. *lobe switching*, tudi *sequential lobing*) je tehnika, ki primerja odčitke sprejete moči z od referenčne smeri odklonjenih snopov [1, 3, 4, 8]. Za dvodimenzijsko sledenje se uporabljajo štirje (kot je prikazano na sliki 3), z uporabo štirih anten, sevalnikov ali stanj antenske skupine [1, 4]. Ko je tarča namerjena, so vsi odčitki enaki in razlik ni. Dvodimenzijsko preklapljanje snopov je diskretizirana različica zveznega stožčastega skaniranja in se lahko analogno imenuje *piramidno skaniranje*. Blokovni načrt sprejemnika za namerjanje v eni dimenziji je na sliki 4.



Slika 3: Preklapljanje snopov diskretno skanira prostor okoli tarče, iz odčitkov sprejete moči signala in geometrije pa sistem določi smer tarče in potreben popravek.



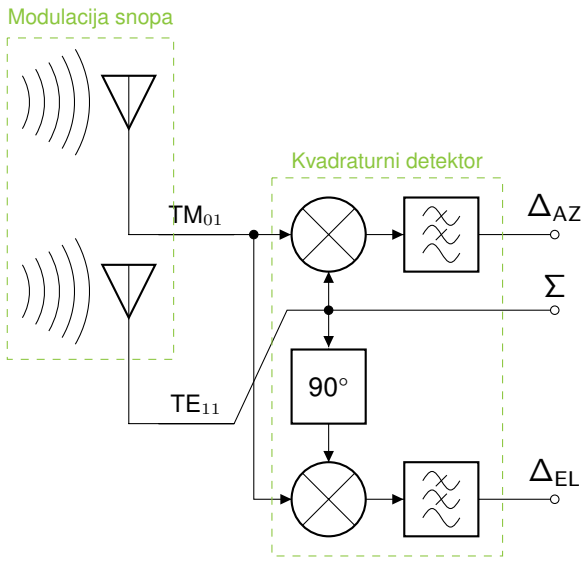
Slika 4: Blokovni načrt sprejemnika pri preklapljanju snopov v eni dimenziji. Kadar je tarča namerjena (v ravnini točno med antenama), sta odčitka moči enaka. Za dvodimenzijsko delovanje sta potrebna dva taka sprejemnika (tehnika s štirimi snopi) ali trije sevalniki z enim sprejemnikom (predlagana tehnika s tremi snopi).



Slika 5: Blokovni načrt sprejemnika pri monopulzni tehniki s skupino štirih anten.

2.3 Monopulzni princip

Vsem do sedaj opisanim tehnikam je skupno to, da za detekcijo oziroma potrebno obdelavo signalov potrebujejo določen čas, kar je omejitev pri najhitrejših tarčah in pri dosegljivi ločljivosti (npr. radarskega) sistema. Tudi hitra električna izvedba stožčastega skaniranja ali preklapljanja snopov še ni končna omejitev tehnik sledenja, pač pa obstaja še realnočasovni monopulzni princip, ki s koherentnim sprejemom in fazno interferometrijo med dvema ali več sevalniki oziroma antenami v trenutku določi smer izvora, vsa potrebna obdelava pa je analogno seštevanje in odštevanje, s čimer se oblikujejo ustrezne oblike snopov za signale vsote in razlike [1, 2, 3, 9, 8]. Izraz monopulzni princip prihaja iz področja radarske tehnike, saj omogoča detekcijo tarče v trenutku oziroma iz enega samega odboja radarskega impulza, s čimer se povečajo zmogljivosti radarskih sistemov [8].

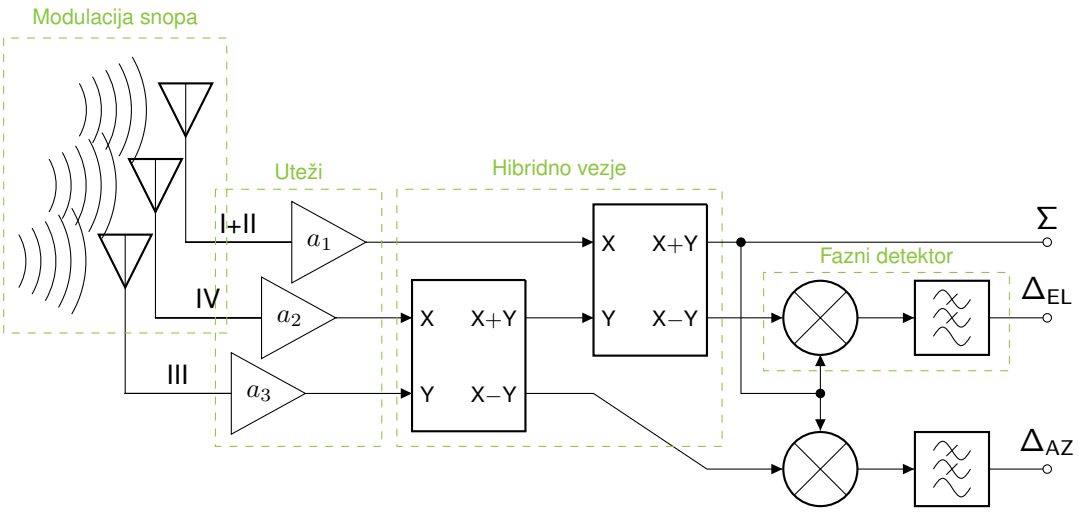


Slika 6: Blokovni načrt sprejemnika pri monopulzni tehniki z dvorodovno anteno s krožno polarizacijo.

Za doseganje raznolikosti smernih diagramov v istem časovnem trenutku imajo monopulzne antene več ločenih sevalnikov oziroma so sestavljene kot skupine (ponavadi) štirih anten [1, 3, 9]. Potrebna obdelava je prikazana na blokovnem načrtu na sliki 5. Skupaj s hibridnim vezjem za obdelavo signalov so take monopulzne antene nepraktične za poceni oziroma masovno izdelavo. Izvedene so lahko tudi kot večrodovna antena (ang. *multimode antenna*), pri kateri so neodvisno vzbujani rodovi s primernimi smernimi diagrami za monopulzno delovanje [1, 2, 8]. Take antene so že kompaktnejše, v primeru krožno polariziranih signalov pa je možna še dodatna poenostavitev, saj je informacija o smeri tarče že prostorsko-časovno modulirana v sprejetem kvadraturnem signalu razlike s signalom vsote kot referenco [10, 5]. Take antene so lahko kompaktne, cenejše in primernejše za masovno izdelavo [5], kar je razvidno tudi iz enostavnejšega blokovnega načrta na sliki 6.

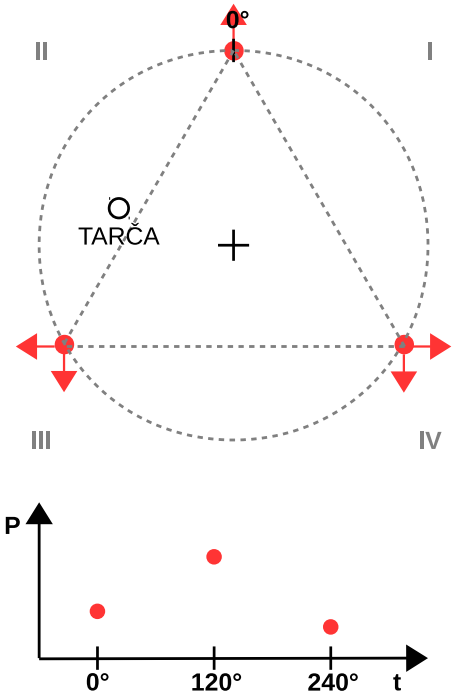
3 Predlog nove metode

Tako pri obstoječih izvedbah metode preklapljanja snopov, kakor pri monopulzni tehniki z antensko skupino so za dvodimenzijsko sledenje v obstoječi literaturi uporabljene štiri antene (snopi) ali več. Ker gre matematično gledano za iskanje točke v ravnini, so za dvodimenzijsko lociranje potrebne meritve samo v treh nekolinearnih točkah (triangulacija). To pomeni možnost poenostavitev tehnik preklapljanja snopov (sliki 8 in 4) in tudi monopulzne (slika 7) z uporabo le treh anten, sevalnikov, stanj antenske skupine oziroma v splošnem treh snopov. Ker je na koncu za motoriziran popravek ponavadi potrebna informacija v karteziečenem sistemu, je potrebno upoštevati transformacijo točk oziroma uteži pri seštevanju vektorjev tako da so simetrični oziroma v vsoti enaki nič. Na sliki 8 je skica namerjanja s tremi točkami in označeno kartezično informacijo posameznih točk. Z večanjem števila točk narašča ločljivost, a hitrost skaniranja upada.



Slika 7: Blokovni načrt sprejemnika pri predlagani monopulzni tehniki s skupino treh anten. Operacije seštevanja in odštevanja morajo biti ustrezno uteženе za transformacijo v kartezični sistem.

Klasične tehnike s štirimi snopi ostajajo zanimiv kompromis med ločljivostjo in hitrostjo, saj v primeru glede na tri kljub daljši meritvi odpade transformacija v kartezični sistem (enostavnejša obdelava).



Slika 8: Predlagana tehnika sledenja s triangulacijo točke v ravnini. Na sliki je označeno, informacijo katere kartezične smeri nosi posamezna točka trikotnika.

4 Zaključek

Na podlagi obstoječih metod sledenja izvoru signala in preproste matematike ter geometrije v koherentni in nekoherentni različici predstavljena nova, ki za sledenje uporablja samo tri snope namesto obstoječih metod s štirimi, kar omogoča kompaktnejšo izdelavo na račun dodatne obdelave za potrebno transformacijo. V naslednjem koraku predlagamo prototipiranje in praktičen preizkus predlaganih sistemov - preklapljanja treh snopov oziroma monopulznega s tremi snopi.

Zahvala

Raziskovalni program št. P2-0246 je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Zahvala gre tudi Univerzitetni ustanovi ing. Lenarčič Milana za raziskovalno štipendijo.

Literatura

[1] R. B. Dybdal: Antenna Tracking, *Antenna Engineering Handbook*, str. 42-12, 2007

[2] J. S. Cook in R. Lowell: The Autotrack System, *The Bell System Technical Journal*, št. 4, str. 1283-1307, 1963

[3] G. J. Hawkins, David J. Edwards in J. McGeehan: Tracking Systems for Satellite Communications, *IEE Proceedings F - Radar and Signal Processing*, št. 5, str. 393-407, 1988

[4] J. H. Cook Jr in W. K. Dishman: An Improved Scanning Automatic Tracking Antenna for Telemetry Applications, *International Telemetering Conference Proceedings*, 1989

[5] P. Miklavčič: Optimizacija preprostega monopulznega antenskega sevalnika za primarno gorišče globokega zrcala, 2017, http://repozitorij.uni-lj.si/IzpisGradiva.php?id=94823

[6] P. Miklavčič: Tehnike sledenja izvoru signala, *Zbornik 23. seminarja radijske komunikacije, 31. januar-2. februar 2018*

[7] W. Gawronski in E. M. Craparo: Antenna Scanning Techniques for Estimation of Spacecraft Position, *Aerospace Conference Proceedings* 2, str. 2-2, 2002

[8] S. M. Sherman in D. K. Barton: Monopulse Principles and Techniques, 2. izd., Norwood: Artech House, 2011

[9] E. Kornaros, S. Kabiri in F. De Flaviis: A Novel Model for Direction Finding and Phase Center With Practical Considerations, *IEEE Transactions on Antennas and Propagation* 65, št. 10, str. 5475-5491, 2017

[10] M. Vidmar: Sestavljanje krožne polarizacije, 1994, http://lea.hamradio.si/~s53mv/archive/a167.pdf

Determining the locking state of a Fabry–Pérot laser

Jakup Ratkoceri[1,2] and Boštjan Batagelj[1], Member, IEEE

[1] Radiation and Optics Lab., Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[2] IPKO Telecommunication, Pristina, Kosovo
e-mail: Jakup.Ratkoceri@fe.uni-lj.si

Abstract. *In this paper we experimentally examine the phenomenon of optical injection locking and its stability using a Fabry–Perot (FP) semiconductor laser with a wavelength of 1550 nm, which is widely used for the transmission in optical access networks due to its simplicity and low cost. Injection locking makes it possible to overcome the fundamental limits of the multi-longitudinal-mode behavior of the FP laser. Under injection-locking conditions, the linewidth characteristics of the master laser are transferred to the slave laser, which leads to a drastic improvement in the performance of the slave laser. However, the system dynamics of injection locking is very complex, which makes it very difficult to determine the locking state of the laser. Therefore, we propose a method whereby observing the laser's spectral coherence using self-homodyne detection we are able to determine the stable locking condition of the laser.*

Keywords. Fabry–Pérot laser diodes, injection locking, locking range, linewidth, self-homodyne

1 Introduction

In today's telecommunications networks the fiber-optic access network is wieldy accepted among academia and industry as being a future-proof solution. [1] There are several fiber-optic access solutions that enable high-data-rate capacities, but not many of them come without economic implications. One of the solutions that could be acceptable from the technological and economic points of view is the injection-locking technique. Using this technique a multi-longitudinal-mode Fabry–Pérot (FP) laser can operate in a single-longitudinal-mode regime, giving it greatly improved characteristics. [2] In a wavelength division multiplexing passive optical network (WDM-PON) the source laser in the optical network units is an optical injection-locked laser. [3]

Injection locking greatly enhances the parameters of a laser such as a multimode FP laser. This defines most of the fundamental limits of the laser, such as the mode partition noise, relaxation oscillation frequency, non-linear electron-photon coupling, relative intensity noise etc., to be improved. However, due to its rich dynamic behavior, determining the stable locking conditions is a particularly difficult task [4].

For the case of a multimode FP laser in the injection-locked state, the longitudinal modes that are not locked experience suppression. Usually, in the literature we find that for a side-mode suppression ratio (SMSR) between 30 and 35 dB the slave laser is considered to be in the injection-locking mode [4]. However, due to the complex behavior of a slave laser, the stability locking map is not only single-valued and correlated to the frequency detuning $\Delta\omega$ (between the master-laser and slave-laser frequencies) and r (ratio of the injected photon density S_i from the master laser and the photon density of the injected locked mode of slave laser S), but the stability map is a multi-valued stability function.

Under the conditions of optical injection locking, the master laser imposes its spectral distribution on the slave laser. Due to the fact that we have the complex behavior of the injection-locked slave laser, we will experimentally observe the process of transferring the linewidth of the master laser to the slave laser using self-homodyne detection, which to the best of our knowledge has not yet been done.

This paper is organized as follows. A short description of the injection-locking phenomenon is followed by different experimental measurements. First, the linewidth of a single-longitudinal-mode laser and a multi-longitudinal-mode laser (FP laser) is measured using the self-homodyne technique. Then, we present the difference between the locking and the unlocking range of the FP laser in the injection-locking map. At the end of the experimental part we present the difference in the locked laser's linewidth on the negative and positive frequency detunings. In the conclusion we propose a novel, self-restorable, injection-locking feedback loop that can contribute to a self-organized optical access network based on a WDM-PON.

2 Injection-locking phenomenon

Injection locking is a nonlinear phenomenon of synchronization that exists when coupling two or more self-sustained oscillators. The optical injection offers many dynamic regimes, such as locking, wave-mixing, relaxation and chaos. [4]

In the injection-locking process the light of the master laser is injected into the cavity of the slave laser, where the injected light coherently combines with the slave laser's internal light. If the injection power is high enough and the frequency difference between the master laser and the slave laser is not large, the slave laser is "dragged" towards the master laser's frequency, locking both the frequency and the phase [3]. The injection locking is affected by the fact that the presence of the out-of-phase component will introduce a phase shift and by the fact that to maintain the steady state of the slave laser, the gain of the injection field is changed.

The slave laser's dynamics can be analyzed through a system of rate equations, with additional terms describing the locking phenomenon [4].
Two adjustable parameters through which we control the injection locking are the injected power ratio r (the ratio of the photon density S_i injected into the slave laser's cavity and the photon density S of the injection-locked mode of the laser) and the frequency detuning ($\Delta\omega = \omega_i - \omega_0$), where ω_i is the master laser's frequency and ω_0 is the cavity resonance frequency of the slave laser. The locking bandwidth obtained from the rate equation is:

$$\Delta\omega_{min} = -(c/2n_cL)\sqrt{((S_i/S)(1+\alpha^2))} < \Delta\omega < (c/2n_cL)\sqrt{(S_i/S)} \quad (1)$$

where n_c is the refractive index, L is the laser-cavity length, c is the speed of light and α is the linewidth-enhancement factor. [5], [6]

3 Experimental measurements

3.1 Linewidth measurements

The laser's linewidth represents the width of the optical spectrum, or more precisely, the width of the power spectral density of the emitted electrical field in terms of the frequency, wavelength or wavenumber. [7] The width of the laser can be seen as a consequence of the fluctuations in the phase of the optical field. The laser's linewidth broadening is attributed to the change of the phase and the intensity change due to the spontaneous-emission event. [8]
To be able to analyze the interrelation between the linewidths of the master laser and the slave laser during the injection locking we need to measure the linewidth of both the lasers used in this experiment before engaging them in the injection locking.
For the linewidth measurement we deployed a technique known as the delayed self-homodyne (Fig. 1). With this technique the optical light of the laser is divided into two branches, where in one branch of the Mach-Zehner interferometer a 20-km-long single-mode fiber is introduced to add around 100 microseconds of time delay. With this time delay we are able to measure the linewidth down to the range of 10 kHz. For the detection of the optical signal after the recombination of the two branches of the interferometer, a PIN-FET photodiode is used. From the photodiode the signal is sent to a radio-frequency spectrum analyzer for further analysis. Below is a schematic drawing of the delayed self-homodyne technique.
We measured the linewidths of two laser sources. The single-longitudinal-mode laser is a HP 8153A tunable laser source (TLS). The measurement result for its linewidth is presented in Figure 2. When the optical output power was set to 6 dBm, the linewidth of 100 MHz is measured at -3 dB threshold.

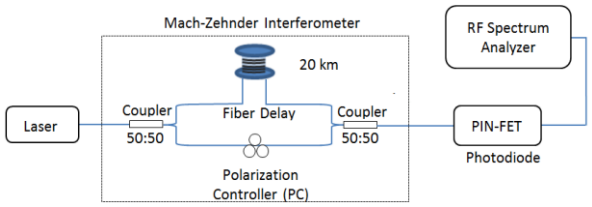


Figure 1. Delayed self-homodyne technique

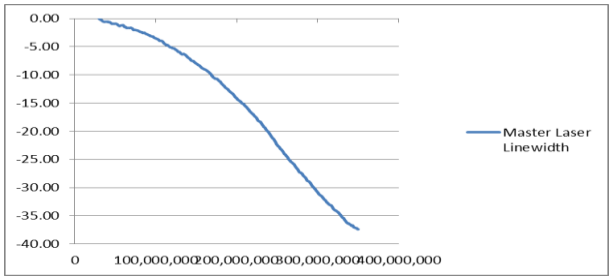


Figure 2. Linewidth measurements for a single-longitudinal-mode laser.

The second laser is a multi-longitudinal-mode laser with attached external cavity. We measured the linewidth of a 240-micrometer-long InGaAsP (indium gallium arsenide phosphide) FP laser. The external cavity is made up of 1-m long single-mode optical fiber and physical contact (FC) optical connector of 0.5% reflectivity. This FP laser has a threshold current of 10 mA. When the FP laser is stabilized at 20°C, with a bias current of $I=3I_{th}$, we find the dominant mode at 1546.244 nm. The free spectral range of external cavity is 100 MHz. For the linewidth of the FP laser with attached external cavity, due to its multi-longitudinal spectrum, in the RF spectra we have the corresponding mixing products of the optical longitudinal modes and the total linewidth is quite broad (Figure 3).

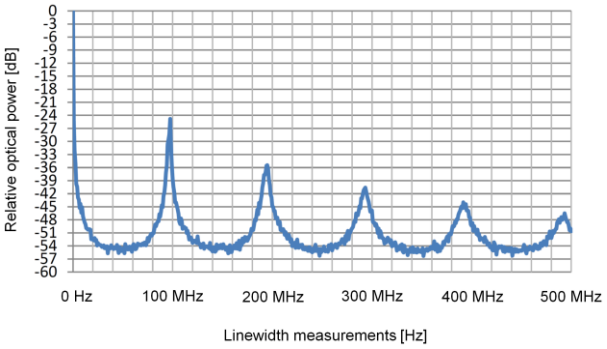


Figure 3. Linewidth measurements for the multi-longitudinal-mode laser

3.2 Injection-Locking Measurement

For the measurement of the injection-locked FP laser the reflection style set-up presented in Figure 4 is used [3]. For the master laser the single-longitudinal-mode laser is employed, while for the slave laser the multi-longitudinal-mode laser is employed. We used two optical spectrum analyzers: the OSA1 is used to control the power of the output signal from the master laser, and

the OSA2 is used to observe the injection-locking phenomenon.

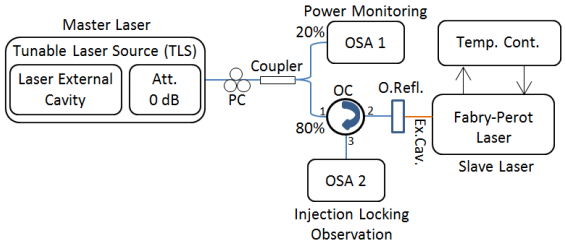


Figure 4. Experimental set-up for injection locking

The mode chosen to be locked was the first mode on the right of the central mode with a wavelength of 1457.660 nm. The FP laser was forward biased with a current of 30 mA, and was stabilized at 20°C. For this experiment we considered the FP laser to be in the injection-locked state when we measured the SMSR to be higher than 30 dB. Above this value the FP laser is operating as a single-longitudinal-mode laser. The optical spectrum of the FP laser in the injection-locked state and the free-running state is presented in Figure 5 and the locking map of the FP laser achieved from our experiment is presented Figure 6.

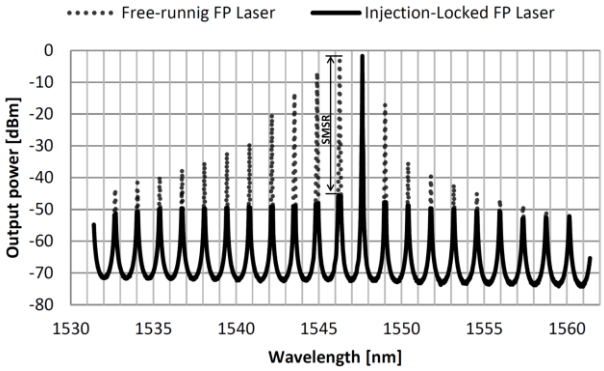


Figure 5. Optical spectrum of the FP laser

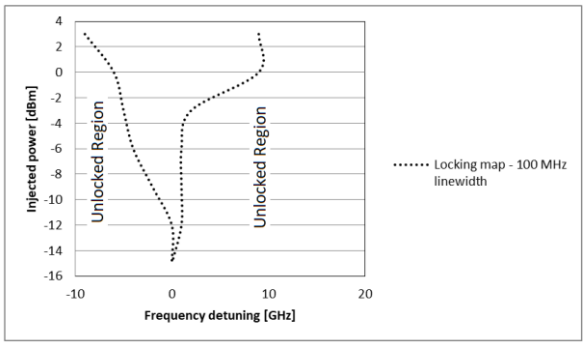


Figure 6. Injection locking map of the FP laser

As expected, with a higher injected power we could achieve a higher frequency detuning and a broader locking region.

4 Difference with respect to the negative and positive frequency detunings

It is well known that a FP laser under injection locking inherits most of the characteristics of the master laser. [4] The linewidth is one of the characteristics that is passed to the slave laser, and in our experiment we expected to observe a radical linewidth transformation, because we will transform the linewidth of a multi-longitudinal-mode laser such as a FP laser into a single-longitudinal-mode linewidth source.

The set-up for this experiment is a combination of the two previously presented set-ups. We preserved the same set-up values as in the previous section for both lasers. With the combined set-up we were able to engage the laser in the injection-locking process and also observe the process of transforming the FP laser linewidth from its free-running mode into the injection-locked state linewidth.

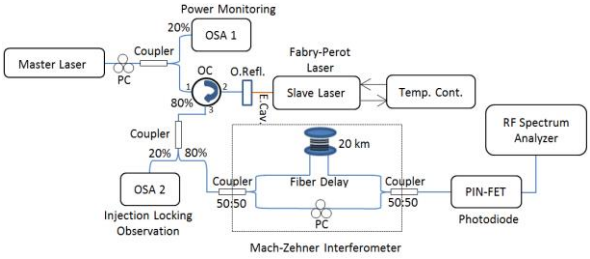


Figure 7. Experimental set-up for injection locking and linewidth observation using the delayed self-homodyne technique.

The output power of the master laser is fixed for all the conducted experiments presented in this section. In first experiment we analyzed the transformation of the linewidth of the FP laser in three different stages, related to the SMSR, as shown in Figure 8. It is worth underlining that we performed this experiment for negative $\Delta\omega$, which means that we had a lower master-laser frequency in comparison to the slave-laser frequency.

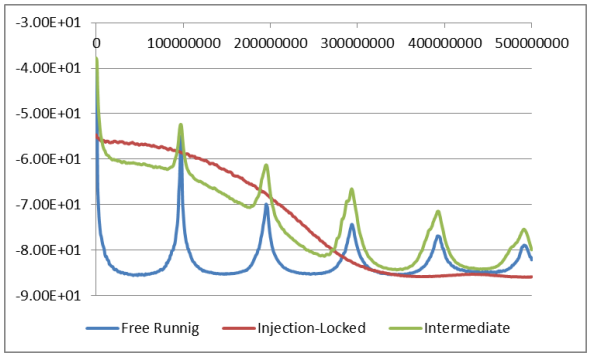


Figure 8. Experimental results for negative frequency detuning

The blue line represents the spectrum of the FP laser in the free-running state when no optical injection is applied to the FP laser. As expected, we have a similar curve as in Figure 2. For the spectrum of the FP laser when fully injection locked we have the same linewidth as for the case when we measure the linewidth of the master laser. Therefore, we have a complete transformation of the FP laser. At this stage the SMRS was higher than 35 dB. We also measured an intermediate stage, where we see that we have an in-between curve. At this point the measured SMSR was around 15 dB.

The second part of our measurement was to analyze the spectrum of the FP laser during the process of injection locking for the positive $\Delta\omega$.

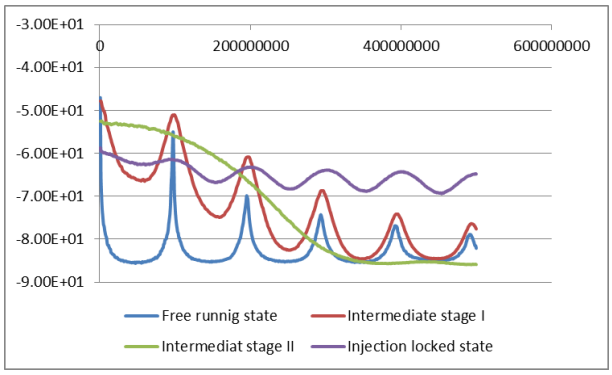


Figure 9. Experimental results for positive frequency detuning

It is worth noting that for this set of measurements, even though for the intermediate stage II we had a SMSR higher than 35 dB, we can see that the transformation of the linewidth is not complete, which could pose a problem when using an injection-locked FP laser.

For the free-running state and the injection-locked states, the optical spectra of the FP laser that correspond to the self-homodyne spectra presented in Figures 7 and 8 are clearly depicted in Figure 6, where the longitudinal free-running modes are presented with a dotted line and with a solid line we have the injection-locking spectra, where we see the injection-locked mode and the other suppressed modes. In addition to the explanation for the results in Figure 8, for the intermediate stage II in Figure 9 we present the corresponding optical spectrum, where we can clearly see that even though we have a SMSR higher than 30 dB, the linewidth transfer is not completed.

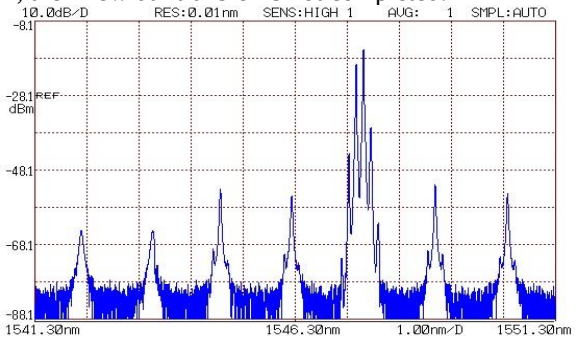


Figure 9. Optical spectrum for the intermediate stage II.

5 Conclusion

In this paper we determined the stable locking range for an injection-locked FP laser by observing the linewidth transformation of the injection-locked FP laser. By using a delayed self-homodyne technique we were able to measure the linewidths of the master laser and the FP laser in the free-running mode and then observe the transfer phenomenon of the linewidth during the injection-locking process. After performing the linewidth analysis, for the positive and negative frequency detunings, we came to the conclusion that for the stable injection locking of the FP laser the widely used condition of an SMSR higher than 30 dB is a necessary but not sufficient condition. As we have shown from the results for positive frequency detuning, even when we had a 30-dBm SMSR, the linewidth was not completely transferred for certain values of the frequency detuning.

In this paper we show that the injection-locking condition can be observed from the radio-frequency spectrum using the delayed self-homodyne technique and FP laser with attached external cavity. Based on the achieved results we propose a self-restorable architecture for the stable injection-locking behavior, as presented in Figure 9. The proposed system makes it possible to observe the linewidth change of the injection-locked FP laser, and through the temperature changes, the feedback loop keeps the injection-locked laser in a stable operational state.

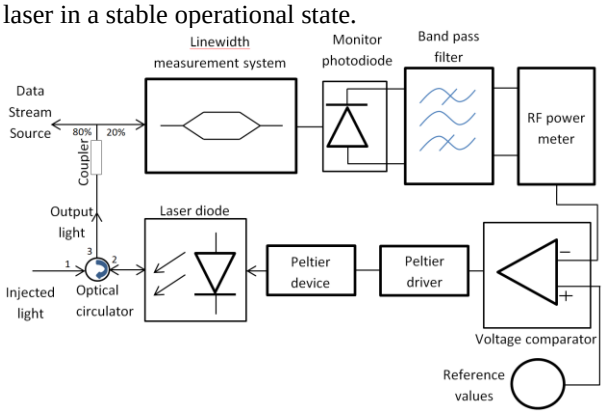


Figure 9. Self-restorable architecture for stable injection locking of a FP laser

References

[1] B. Batagelj, V. Janyani, S. Tomazic, "Research challenges in optical communications toward 2020 and beyond," Informacije Midem, Vol. 44, No. 3, 177 – 184

[2] E. K. Lau, L. J. Wong, M.C. Wu, "Enhanced modulation characteristics of optical injection-locked lasers: A tutorial," IEEE J. Sel. Top. Quant. Electron. 15(3), 618-633(2009)

[3] V. Eržen, B. Batagelj, "Rodovno vklenjen FP-LD za brezbarvni WDM-PON," ERK 2015, A:61-64

[4] M. M. Krstić, et al, "Multivalued stability map of an injection-locked semiconductor laser," IEEE J. Sel. Top. Quant. Electron. 19(4), 1501408-1501408 (2013)

[5] X. Jin, S.-L. Chuang, "Bandwidth enhancement of Fabry-Perot quantum-well lasers by injection-locking," Solid State Electronics, Vol. 50, Iss. 6, 2006, Pages 1141-1149

[6] M. Bondiou, et al, "Linewidth of an optically injected semiconductor laser," Journal of Optics B: Quantum and Semiclassical Optics, Vol. 2, No. 1, 1999

[7] https://www.rp-photonics.com/linewidth.html?s=ak

[8] C. Henry, "Theory of the linewidth of semiconductor lasers," in IEEE Journal of Quantum Electronics, vol. 18, no. 2, pp. 259-264, February 1982

A preliminary study on side mode suppression in optoelectronic oscillators using a two-mode optical fiber in the feedback loop

Antonio Astorino[1], Mehmet Alp Ilgaz[2], Karsten Rottwitt[1], Boštjan Batagelj[2]

[1]*DTU Fotonik, Ørsteds Plads, Building 343, 2800 Kongens Lyngby, Denmark*
[2]*University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, 1000 Ljubljana, Slovenia*
email: antonio.astorino.ing@gmail.com

Abstract

We discuss the use of a two-mode step-index fiber in the optical path of an optoelectronic oscillators as an alternative to a dual loop. The idea is to exploit the different group velocities of two LP fiber modes instead of using two single-mode fibers with different lengths. The basic theory of dual-loop oscillators can therefore be reused in this context. Nevertheless, due to a generally small group delay between two fiber modes, compared to the loop delay, the proposed idea needs a preliminary analysis.

1 Introduction

An optoelectronic oscillator (OEO) is able to produce high spectral purity carriers in the millimeter-wave range [1–3]. The limit in the spectral purity is mainly due to the relative intensity noise of the laser, the excess noise added by the optical amplifier, and shot noise related to the photodetection. As demonstrated in [2], the phase noise can be reduced by increasing the loop-delay time. In addition, using two feedback loops with different delay times, i.e. two single-mode fibers with different lengths, allows to suppress unwanted resonant frequencies, otherwise difficult to suppress with standard radio-frequency band-pass filters.

The idea proposed in this work is to replace the dual loop with a single one, maintaining the same working principle; instead of using two different fiber lengths, one fiber and two modes with different group velocities are considered. Multiple modes can be simultaneously and easily excited by means of radial offset launching [4, 5] or mechanically induced long-period gratings [5, 6], for example. Besides the use of one fiber only in the feedback loop, the main advantage of the proposed configuration, compared dual-loop OEOs, is the absence of the optical splitter and the optical or electrical combiner. The reader can refer to [7, 8] for a comparison with the proposed schematic, depicted in Fig. 1. A laser source is modulated by a Mach-Zehnder modulator (MZM) and amplified by an erbium-doped fiber amplifier (EDFA). The output fiber of the EDFA is spliced to a two-mode fiber (TMF) using the offset splicing technique, in order to equally excite both propagating modes. Using the same technique, the other end of the TMF is then spliced to the single-mode input fiber of the photodetector (PD), whose radio-frequency (RF) output is then amplified and filtered. The resulting signal is a high purity carrier, used as both system output (indicated with "RF out") and feedback signal, the latter modulating the laser source. Notice that the optical path does not involve optical filters. In fact, it is our intention to address the problem of filtering out spurious harmonics without involving such a filter.

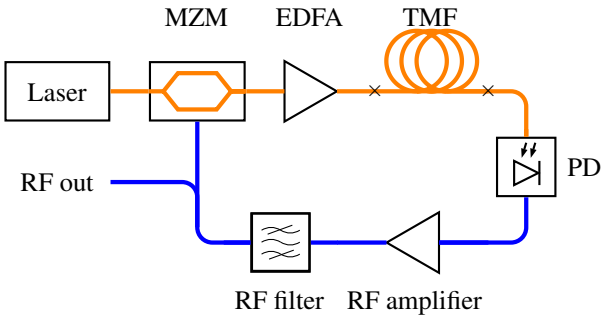


Figure 1: Simplified schematic of an optoelectronic oscillator adopting a two-mode fiber in the feedback loop

2 Background

The working principle of a dual-loop OEO is thoroughly described in [3]. Considering the case in which the modulus of the open-loop gain is larger than unity, the oscillation condition at frequency f_{osc} is achieved if

$$f_{\text{osc}} = \frac{m_1}{\tau_1} = \frac{m_2}{\tau_2} \tag{1}$$

where m_1 and m_2 are integers, and τ_1 and τ_2 denote, respectively, the loop delays of the two modes under consideration.

Let $n_{\text{g},01}$ and $n_{\text{g},11} > n_{\text{g},01}$ be the group indices of the LP_{01} and LP_{11} modes, respectively, propagating in a TMF of length L. By definition of group index, the delays associated to the two modes are $\tau_1 = n_{\text{g},01}L/c_0$ and $\tau_2 = n_{\text{g},11}L/c_0$, c_0 being the speed of light in vacuum. Equation (1) can therefore be written as

$$f_{\text{osc}} = \frac{m_1 c_0}{n_{\text{g},01}L} = \frac{m_2 c_0}{n_{\text{g},11}L}. \tag{2}$$

From (2), it immediately follows that

$$\frac{m_2}{m_1} = \frac{n_{\text{g},11}}{n_{\text{g},01}}. \tag{3}$$

Table 1: Test parameters

Δf_{FWHM} (kHz)	δ (Hz^{-1})
32	$4.707\,344 \times 10^{-8}$
16	$2.353\,672 \times 10^{-8}$
8	$1.176\,836 \times 10^{-8}$
4	$5.884\,180 \times 10^{-9}$
2	$2.942\,090 \times 10^{-9}$
1	$1.471\,045 \times 10^{-9}$

For a two-mode step-index fiber with refractive indices 1.450 and 1.444 and core diameter of 12 µm, at the wavelength of 1550 nm, the ratio $n_{\text{g},11}/n_{\text{g},01} \approx 5219/5216$. Therefore, $m_1 = 5216$ and $m_2 = 5219$ are the smallest integers satisfying (3). Consequently, according to (2), in order for f_{osc} to be 10 GHz, a length $L \approx 107.8$ m is to be chosen. However, it has been demonstrated that full width at half maximum (FWHM) of the oscillator decreases quadratically with the loop-delay time, i.e. the fiber length [2]. Hence, a longer fiber may be selected in order to reduce the line width of the resulting carrier signal. This is always possible, provided that m_1, m_2, and L are multiplied by the same positive integer, so that the oscillation frequency remains constant [see (2)]. Moreover, as numerically demonstrated in the next section, a longer fiber results in a larger delay difference between the two propagating modes, hence achieving a better filtering of the undesired resonance frequencies in the oscillator. On the other hand, in a long fiber, undesired effects such as chirp and signal degradation due to intrinsic losses become significant. For this reason, an intuitively effective choice is to keep the value of L as low as possible. This work aims at analyzing how the value of L and the purity in the carrier signal are related to the filtering property of the proposed system.

3 Simulations

For the sole purpose of calculating the noise-to-signal ratio δ, defined below, we first consider a single-loop OEO with the desired output frequency $f_{\text{osc}} = 10$ GHz, loop length $L = 100$ m, and the group index $n_{\text{g}} = 1.450\,587$. For sufficiently small values of loop-delay time, it is possible to express the proportionality constant δ that relates the FWHM and the fiber length as [2]:

$$\delta = 2\pi \Delta f_{\text{FWHM}} \left(\frac{n_{\text{g}} L}{c_0} \right)^2 \tag{4}$$

We then consider, as an example, the set of values of Δf_{FWHM} reported in Table 1, along with the corresponding values of δ, calculated by using (4).

We now focus on the OEO depicted in Fig. 1, assuming $f_{\text{osc}} = 10$ GHz, $n_{\text{g},01} \approx 1.450\,713$ and $n_{\text{g},11} = 1.451\,548$. The value of L is varied from $L_{\text{min}} = 107.8$ m to 10 780.0 m in steps of L_{min}. This ensures that (1) can be always verified, for any length, by conveniently multiplying m_1 and m_2 by the same integer number $p > 0$. Here, p runs from 1 to 100. In other words, (1) can be

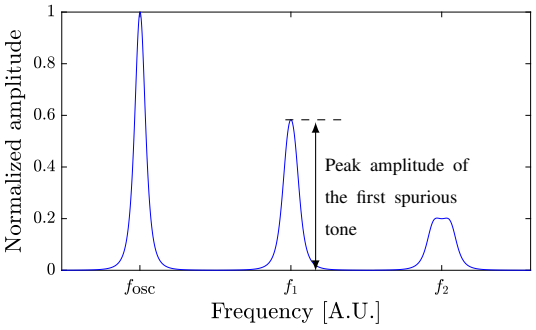


Figure 2: Spectrum of a dual-loop OEO in a region close to the central frequency f_{osc}. The frequencies of the first two higher harmonics are indicated with f_1 and f_2.

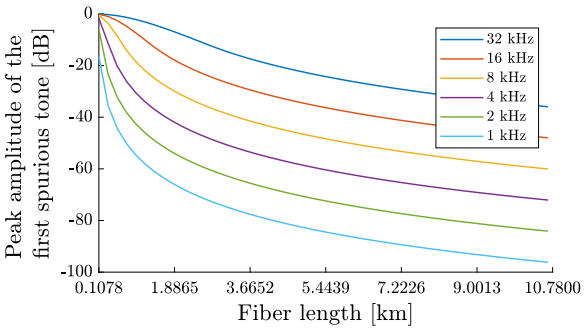


Figure 3: Peak amplitude of the spurious tone closest to the desired oscillation frequency as a function of the TMF length for selected values of the carrier FWHM.

written as

$$f_{\text{osc}} = \frac{p m_1 c_0}{n_{\text{g},01} L_p} = \frac{p m_2 c_0}{n_{\text{g},11} L_p}, \quad p = 1, 2, \ldots, 100. \tag{5}$$

where $L_p = p L_{\text{min}}$.

The spectrum of the dual-loop OEO is calculated as the product between the two spectra produced by two independent single-loop OEOs, one in which only the LP_{01} mode propagates in the TMF and one where only the LP_{11} mode is present. The FWHM of all tones is calculated by solving (4) for Δf_{FWHM}, namely,

$$\Delta f_{\text{FWHM}} = \frac{\delta}{2\pi} \left(\frac{c_0}{n_{\text{g}} L_p} \right)^2, \quad n_{\text{g}} = n_{\text{g},01}, n_{\text{g},11}. \tag{6}$$

The spectra of both tones are assumed to have a Lorentzian amplitude distribution. [2]

The effectiveness of the proposed approach as a filter for spurious tones is evaluated by measuring the simulated peak amplitude of the undesired harmonic closest to f_{osc} (see Fig. 2) as a function of L_p. This is, in fact, assumed to be the predominant unwanted harmonic.

The simulation result is plotted in Fig. 3. The figure shows that, as expected, spurious modes are better filtered when Δf_{FWHM} is low. Also, when the fiber length becomes larger than about 5 km, the improvement achieved by halving Δf_{FWHM} is constant and approximately equal to 12 dB, whereas and the curve slope is

around $3\,\mathrm{dB\,km^{-1}}$. Moreover, the impact on the system performance is much higher when the fiber length increases from its minimum up to around $5\,\mathrm{km}$. This means that increasing the fiber length may not be a good strategy to attenuate undesired harmonics when L_p becomes larger than $5\,\mathrm{km}$. In fact, as already mentioned, excessively long fibers may worsen the system performance due to attenuation and dispersion.

4 Conclusion

In this work, a preliminary study on an optoelectronic oscillator involving a two-mode fiber in the feedback loop has been conducted. The system has been analyzed as a dual-loop oscillator in which the short loop is associated with the LP_{01} mode propagating in the fiber and the long loop is associated with the LP_{11} mode.

The main goal has been analyzing the impact of the fiber length on the suppression of spurious modes. The simulated results show that, when the fiber length is close to its minimum, any small variation of its length largely affects the system filtering performance. When instead the fiber is sufficiently long, an increment in its length produces a reduction in the amplitude of the first harmonic of about $3\,\mathrm{dB\,km^{-1}}$.

The obtained results are very promising; an attenuation of the first undesired harmonic of around $100\,\mathrm{dB}$ has been achieved with a fiber spool roughly $10\,\mathrm{km}$ long, despite the assumption of a large noise in the setup.

Acknowledgment

This work has received funding from the European Union's Horizon 2020 Research and Innovation Program 20142018 under the Marie Sklodowska-Curie grant agreement No. 642355.

References

[1] X. S. Yao and L. Maleki. High frequency optical subcarrier generator. *Electronics Letters*, 30(18):1525–1526, Sep 1994.

[2] X. S. Yao and L. Maleki. Optoelectronic oscillator for photonic systems. *IEEE Journal of Quantum Electronics*, 32(7):1141–1149, Jul 1996.

[3] X. S. Yao, L. Maleki, Yu Ji, G. Lutes, and Meirong Tu. Dual-loop opto-electronic oscillator. In *Proceedings of the 1998 IEEE International Frequency Control Symposium (Cat. No.98CH36165)*, pages 545–549, May 1998.

[4] L. Raddatz, I. H. White, D. G. Cunningham, and M. C. Nowell. An experimental and theoretical study of the offset launch technique for the enhancement of the bandwidth of multimode fiber links. *Journal of Lightwave Technology*, 16(3):324–331, Mar 1998.

[5] Keigo Iizuka. *Elements of Photonics, Volume II For Fiber and Integrated Optics*, chapter 11. John Wiley & Sons, Inc., 2002.

[6] S. Savin, M. J. F. Digonnet, G. S. Kino, and H. J. Shaw. Tunable mechanically induced long-period fiber gratings. *Opt. Lett.*, 25(10):710–712, May 2000.

[7] J. H. Cho, H. Kim, and H. K. Sung. Reduction of spurious tones and phase noise in dual-loop oeo by loop-gain control. *IEEE Photonics Technology Letters*, 27(13):1391–1393, July 2015.

[8] D. Eliyahu and L. Maleki. Low phase noise and spurious level in multi-loop opto-electronic oscillators. In *IEEE International Frequency Control Symposium and PDA Exhibition Jointly with the 17th European Frequency and Time Forum, 2003. Proceedings of the 2003*, pages 405–410, May 2003.

Examination of Power Penalty due to Chromatic Dispersion Effect on mm-Wave Optical-Communications Fiber-Link with Mach-Zehnder Modulator

Mehmet Alp Ilgaz[1], Antonio Astorino[2], Guillermo J. Martínez T.[1,3], Boštjan Batagelj[1]

[1]*University of Ljubljana, Faculty of Electrical Engineering, Radiation and Optics Laboratory*
Tržaška cesta 25, 1000 Ljubljana, Slovenia
[2]*Technical University of Denmark,*
Ørsteds Plads, Building 343, 2800 Kongens Lyngby, Denmark
[3]*Universitat Politècnica de Valencia, Escola Tècnica Superior de Enginyers de Telecomunicació*
Camí Vera s/n, 46022, València, España
tel: +386 1 4768 400, e-mail: mehmet.ilgaz@fe.uni-lj.si

Abstract*. Analog optical fiber link is one of the common configuration to transmit high frequency signals in the microwave (μ-W) and millimeter wave (mm-W) ranges. In the mm-W signal range, only external modulation can be used, with the intention to overcome the limited bandwidth of a directly modulated laser. Such an external modulation can be implemented by using a Mach-Zehnder modulator (MZM). In mm-W optical fiber links, one of the main issues is the power penalty due to the chromatic dispersion of the optical fiber. This problem is thoroughly analyzed in this paper.*

1 Introduction

In today's communication technology, optical transmission is widely adopted due to the low losses and the high bandwidth characterizing optical fibers. To transfer mm-W signals over the optical domain, one of the possibilities is using analog optical fiber links (AFOL). This technique, makes use of several components such as laser diodes, opto-electronic modulators, optical fibers and photodetectors (Fig. 1).

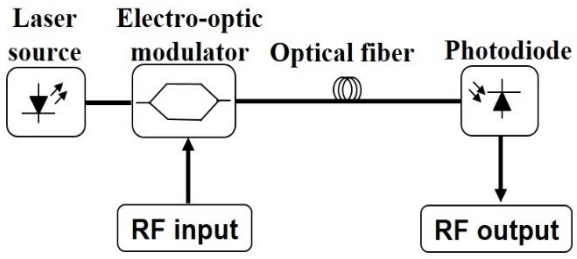


Figure 1. AFOL configuration with external modulation.

The optical loss of the fiber depends on the wavelength of the light source. For example, the G.652D optical fiber has a loss of only 0.2 dB/km at the wavelength of 1550 nm. Hence, the power penalty due to attenuation is low but additional power penalty occurs due to the fiber chromatic dispersion [1-3]. In addition, the chromatic dispersion is more dominant in the mm-W range compared to the μ-W range [4]. In the mm-W range, the light source needs to be modulated with an external modulator such as the MZM [5-6] or an electro-absorption modulator [7-8], as an alternative to the direct laser modulation which is affected by frequency chirp. For the purpose of effectively transmitting the mm-W signal through analog optical links, the external modulation and power penalty need to be analyzed.

2 Optical signal transmission from Laser diode to MZM

In this section, the transmission of optical signals from a distributed feedback laser (DFB) to a MZM is investigated. We have shown how the input signal changes with a mm-W modulated signal, and we observe the optical signal behavior at the output of the MZM. These calculations are based on [3] and [9-14].

The optical field at the input of the MZM [$E_{in}(t)$] can be described as:

$$E_{in}(t) = E_0 \cos(\omega_0 t) \quad (1)$$

where E_0 is the electric field amplitude and ω_0 is the angular frequency of the light produced by the laser. The output of the MZM can be written as

$$E_{out}(t) = E_0 \left\{ \left\{ \cos\left(\beta \frac{\pi}{2}\right) \cos\left(\alpha \frac{\pi}{2} \cos(\omega_{RF} t)\right) - \sin\left(\beta \frac{\pi}{2}\right) \sin\left(\alpha \frac{\pi}{2} \cos(\omega_{RF} t)\right) \right\} \cos(\omega_0 t) \quad (2)$$

where ω_{RF} is the modulation angular frequency, β is the normalized bias, and α is the normalized radio frequency (RF) amplitude driving the MZM. The normalized drive voltage can be defined [13] as:

$$\frac{U}{U_\pi} = u_0 + u\cos(\omega_{RF} t + \varphi_m) \quad (3)$$

where U_π is the half-wave voltage and U is the drive voltage, u_0 is the normalized bias voltage, u is the normalized voltage amplitude, φ_m is the initial phase. Fig. 2 shows the transfer function of MZM versus applied bias voltage u.

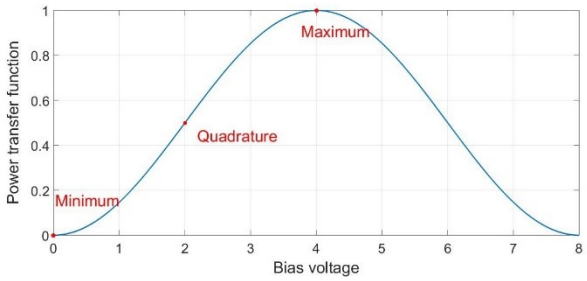


Figure 2. Transfer Function of MZM versus applied bias voltage where the half-wave voltage is 4 V.

There are three major points of interest from the transfer function curve of the MZM when a bias is applied (Fig. 3). These are maximum (when transfer function is 1), quadrature (when transfer function is 0.5) and minimum (when transfer function is 0) points. The quadrature point is preferred for mm-W signal applications [14]. The Fast Fourier Transform (FFT) of the transfer function versus modulation frequency is shown in Fig. 3.

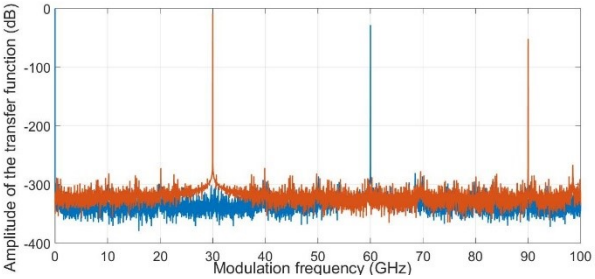


Figure 3. Power of the transfer function versus the modulation frequency.

In Fig. 3, blue lines show the amplitude of the maximum and minimum points and orange lines show the quadrature points.

3 Optical Signal transfer from MZM to Photodiode via optical fiber

The optical field at the output of SMF can be written [14] as:

$$E_{out}(\omega) \approx E_{in}(\omega) e^{-j(\beta_0+\beta_0'\Delta\omega+\beta_0''\frac{\Delta\omega^2}{2})L} \quad (5)$$

where β_0 represents the phase delay and $\beta_0'\Delta\omega$ represents the group delay. The $\beta_0''\frac{\Delta\omega^2}{2}$ is the phase change related to the chromatic dispersion where

$$\beta_0'' = \frac{-\lambda^2 D}{2\pi c} \quad (6)$$

and the photocurrent after photodetection can be written [14] as:

$$i_{PD}(t) = R_{PD}\left\{E_0^2 + 2E_{RF}^2 + 4E_0E_{RF}\cos\left(\frac{LD}{4\pi c}\lambda^2\omega_{RF}^2\right).\cos[\omega_{RF}(t-\tau)] + 2E_{RF}^2\cos[2\omega_{RF}(t-\tau)]\right\} \quad (7)$$

where E_0 is the field amplitude, E_{RF} is the radio frequency amplitude, R_{PD} is the photodetector responsivity, L is the fiber length, c is the speed of light in vacuum, D is the fiber-dispersion coefficient at the wavelength of the optical carrier, τ is the fix delay λ is the optical carrier wavelength. With some approximations, (7) can be represented as (8) in electrical domain [1], [14] as:

$$P \propto 20log\left(cos^2\left(\frac{\pi L D_0}{c_0}\left(\frac{f_{RF}}{f_0}\right)^2\right)\right) \quad (8)$$

where f_0 is the optical carrier frequency and f_{RF} is the modulation frequency.

Fig. 4 describes how the signal behavior changes during the transmission via optical fiber from laser diode to photodetector.

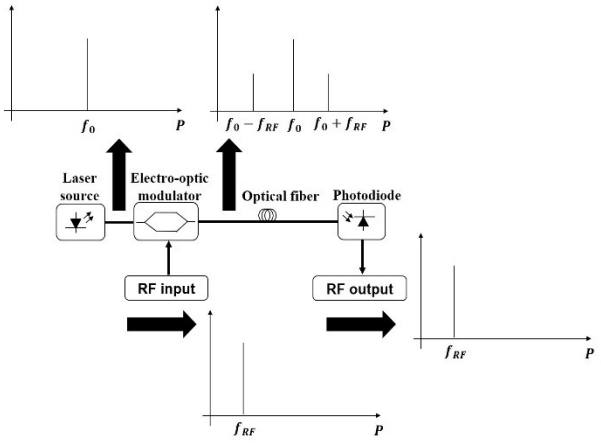


Figure 4. Electrical and optical signal behavior during the transmission via AFOL.

When 1550 nm wavelength is used for optical signal transfer from laser diode to photodetector, the chromatic dispersion is very dominant and has an approximate value of 17 ps/nm/km. Eqn. (8) describes the dispersion penalty [2], [15]. Recently, the 1550 nm wavelength window has become quite popular for optical application because of its low optical loss characteristics to other regions. In addition, below 20 GHz, chromatic dispersion poses a minimal problem to the transmission of optical signal. Coherent beating can help to generate μ-W and mm-W signal optically. In Fig. 5, we demonstrate how the chromatic dispersion affects the signal transfer on the first, third and fifth harmonics. The even harmonics cannot be obtained in the MZM if it is biased on quadrature point (linear region). Also shown in Fig. 5 is the dispersion effect with a 10 GHz modulating signal.

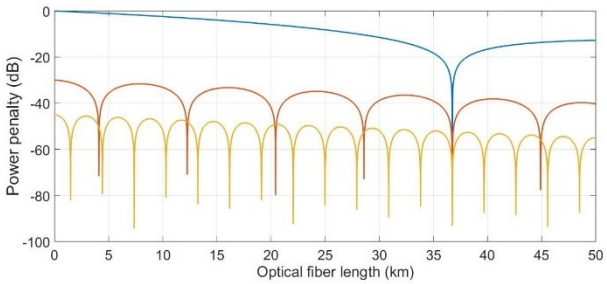


Figure 5. Power penalty of the 1st harmonics (blue) and 3rd (red) and 5th harmonics (orange).

Fig. 5 shows how the chromatic dispersion affects the main mode and harmonics. We have already generated µ-W and mm-W signal in this example. If we compare the µ-W with mm-W signals, it is clearly seen that the chromatic dispesion is less dominant in the µ-W range (especially below 10 GHz). This challenge is very crucial for signal distribution from central station to base stations in the 5G mobile and wireless networks [16]. The reason is that the mm-W signals are preferred to use for next generation mobile systems. In the next example, the optical fiber length is changed from 0 to 10 km and two frequencies are selected; 3-GHz and 39-GHz. The results are shown in Fig. 6.

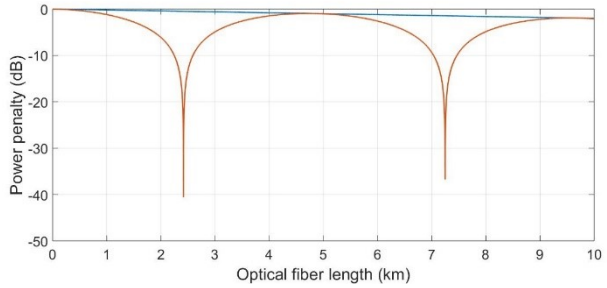


Figure 6. Power penalty of the 3-GHz (blue) and 39 GHz (red).

As can be seen from the graph, there is no power penalty in 3 GHz, and there are two deep points in 39 GHz due to the power penalty. That simulation supports the theory on the power penalty due to the chromatic dispersion.

4 Conclusion

In this paper, we have explained the basics of AFOL. We described the operational regions of MZM and presented simulation studies, highlighting the quadrature bias region as the most useful area for mm-W signal applications. In addition, we have shown the equations for photocurrent after quadratic photodetection and described the electrical power penalty. Finally, we have performed simulations to demonstrate how the µ-W signal and its harmonics in mm-W range are affected by chromatic dispersion and how power penalty affects µ-W and mm-W signals.

Acknowledgment

The authors would like to express their gratitude to the company InLambda BDT d.o.o. for the research equipment and devices. The work presented in this article was created within ARRS Algorithms and optimization procedures in telecommunications program and FIWIN5G Innovative Training Network, which has received funding from the European Union's Horizon 2020 Research and Innovation Program 2014–2018 under the Marie Sklodowska-Curie grant agreement No.642355.

References

[1] M.A.Ilgaz, E. Udvary, and B.Batagelj: Influence of Fibre Chromatic Dispersion on the Performance of Analogue Optical Links for an Opto-electronic Oscillator within a 5G Network Structure in ERK 2016, 19.-21. September 2016, Portoroz, Slovenia.

[2] M.A.Ilgaz, and B.Batagelj: Opto-Electronic Oscillator in the mm-W range for 5G Wireless and Mobile Networks: Design Challenges and Possible Solutions in ONDM 2017, 15. -17. May 2017, Budapest, Hungary.

[3] H. Schmuck: Comparison of optical millimeter-wave system concepts with regard to chromatic dispersion in Electron. Lett., vol. 31, no. 21, pp. 1848–1849, October 1995.

[4] U. Gliese, S. Norskov, and T. N. Nielsen: Chromatic dispersion in fiber-optic microwave and millimeter-wave links in IEEE Trans. Microwave Theory Tech., vol. 44, pp. 1716-1724, October 1996.

[5] A. Garcia, S. Fedderwitz, and A. Stöhr: Optical generation of low-phase noise microwave signals using nonlinear MZM and ultra-long SOA in IMOC 2011, 29. October-1. November 2011, Natal, Brazil.

[6] G. E. Betts, and F. J. O'Donnell: Optical Analog Link Using a Linearized Modulator, Lasers and Electro-Optics Society Annual Meeting in LEOS '94, 31. October-3 November 1994, Boston, USA.

[7] M. Shin, P. S. Devgan, V. S. Grigoryan, and P. Kumar: Low Phase-Noise 40GHz Optical Pulses from a Self-Starting Electroabsorption-Modulator-Based Optoelectronic Oscillator in OFC, 5.-10. March 2006 Anaheim, USA.

[8] G. W. Lee, and S.K. Han: Linearization of a narrowband analog optical link using integrated dual electroabsorption modulator in Microwave Photonics, vol. 1, pp. 21-24, November 1999.

[9] H. Salgado: Dispersion in Optical Fibers, Dec. 2007, https://paginas.fe.up.pt/~hsalgado/co/docs/phase_group_vel_dispersion.pdf

[10] B.Batagelj: Generiranje dveh enako polariziranih optičnih frekvenčnih komponent z Mach-Zehnderjevim modulatorjem in *ERK 2003,* 25.-26. September 2003 Portoroz, Slovenia.

[11] A.F. Elrefaie, R.E. Wagner, D. A. Atlas and D.G. Daut: Chromatic Dispersion Limitations in Coherent Lightwave Transmission Systems in Journal of Lightwave Optics, vol. 6, no. 5, pp. 704-709 May 1988.

[12] G. H. Smith, D. Novak and Z. Ahmed: Overcoming Chromatic Dispersion Effects in Fiber-Wireless Systems Incorporating External Modulator in IEEE Trans. on Mic. Theory and Techn., vol. 45, no. 8, pp.1410-1415, August 1997.

[13] Shcherbakov Vladislav V., Solodkov Anatoly F., and Zadernovsky Anatoly A.: Transmission of Light Intensity Modulation signals in Analog Fiber-Optic Links in RENSIT, vol.8, no.1, pp. 9-24, June 2016.

[14] A. Hilt: Microwave harmonic generation in fiber-optical links in Journal of Telecommunications and Information Technology, pp.22-28, January 2002.

[15] Bostjan Batagelj, et al., “Convergence of fixed and mobile networks by radio over fibre technology in Inform. MIDEM, vol. 41, no. 2, pp. 144-149, June 2011.
[16] M. A. Ilgaz, and B. Batagelj: Preliminary idea for a converged fixed and mobile network infrastructure with 5G using Radio-over-Fiber technology and an Opto-Electronic Oscillator in the millimeter-wave range in Proc. ICTON 2016, 10.-14. July 2016, Trento, Italy, paper Tu.P.5.

Second order performance of relay communication system with three sections in the presence of Nakagami-m and Rician fading

Dragana Krstic[1], Petar Nikolic[2], Danijela Aleksic[3], Dragan Vuckovic[4], Mihajlo Stefanovic[1]

[1] *Faculty of Electronic Engineering, University of Niš, Niš, Serbia*
[2] *TigarTyres, Pirot, Serbia*
[3] *College of Applied Technical Sciences, Nis, Serbia*
[4] *University Business academy, Faculty of Economics and Engineering Managment, Novi Sad, Serbia*
E-pošta: dragana.krstic@elfak.ni.ac.rs

Abstract. *In this paper, the product of arbitrary and independent random variables (RVs) is analyzed. Level crossing rate (LCR) of product of two Nakagami-m random variables and Rician random variable is calculated. This formula for LCR can be used for evaluation the average fade duration (AFD) of wireless relay communication radio system with three sections working over small scale fading channel with presence of Nakagami-m fading and Rician fading. The influence of Nakagami-m parameter, Rician factor, Nakagami-m fading power and Rician multipath fading power on the LCR is considered.*

1 Introduction

There are much discussions of statistical characteristics of the ratios and products of random variables (RVs) lately. The statistical characteristics of random processes can be of the first order and the second order. The first order statistical characteristics are: probability density function (PDF), cumulative distribution function (CDF), characteristic function, outage probability (OP), the amount of fading (AoF), bit error probability (BEP) and moments. Important moments are: the first moment or average value, the second moment or average square value, and the third mentent or cube value. are important characteristics of wireless systems. Statistical characteristics of the second order are: joint probability density function of signal envelope and the first derivative of the signal envelope, joint probability density function of signal envelope in two time instants, level crossing rate (LCR) and average fade duration (AFD). LCR can be calculated as average value of the first derivative of random process [1], AFD can be calculated as the ratio of outage probability (OP) and LCR, and OP can be calculated as probability that signal envelope falls below the determined threshold [2].

The formula for LCR obtained here can be used for calculation LCR of product of three Rayleigh random variables, or product of Nakagami-m random variable and two Rayleigh random variables, and of product of Nakagami-m random variable, Rayleigh random variable and Rician random variable. LCR of product of three Rayleigh random variables can be used for calculation of AFD of wireless relay communication system with three sections working over multipath fading channel in the presence of Rayleigh fading.

There are many works considering performance of the second order of wireless communication radio system in the presence of short term fading. In [3], LCR of product of N Rayleigh random processes is derived and this result is used for calculation AFD of wireless relay communication systems operating over Rayleigh multipath fading channels. In [4], LCR of product of two Nakagami-m random processes is obtained and used for evaluation the AFD of wireless relay communication radio system with two sections operating over Nakagami-m short term fading channel. The PDF, CDF, and moments for distribution of N∗Nakagami fading are derived by Meijer's G-function in [5], and the first order performance are derived for a few modulation schemes which work in channels under presence of N∗Nakagami fading. This derivation is useful for cascaded channels in the presence of Nakagami-m fading.

In the literature, the ratio of product of two random processes and random process and statistics of the ratio of random process and product of two random processes are considered. LCR of the ratio of product of two random processes and product of other two random processes is also analyzed in the literature. Later, newer, general distributions like α-μ are taken into considerations. So, the product of three arbitrary and independent α-μ random variates is processed in [6]. The closed-form expressions for PDF, CDF, and moments are derived. It is shown the calculation process of AoF and OP in cascaded channels.

In this paper, LCR of product of two Nakagami-m random processes and Rician random process is calculated. Obtained expression can be then used for calculation of the AFD of wireless relay communication system with three sections where in the first and second sections Nakagami-m fading is present, and in the third one, Rician fading is present. To the best authors' knowledge, LCR of product of two Nakagami-m random variables and Rician random variable is not observed in available technical literature. The results from this paper could be useful in performance analysis of wireless relay communication systems.

The paper is organized through four sections. The first, introduction, describes previous works in the field of research. In the second one section, the statistics of defined product of three RVs is presented and LCR is derived. Then, obtained formula in closed form is confirmed by some graphs and parameters' influence is analyzed. The conclusion is final part of the paper.

2 Derivation of level crossing rate of product of two Nakagami-m random variables and Rician random variable

2.1 Distribution of random variables

Random variables x_1 and x_2 have Nakagami-m distribution [7]:

$$p_{x_1}(x_1)=\frac{2}{\Gamma(m_1)}\left(\frac{m_1}{\Omega_1}\right)^{m_1} x_1^{2m_1-1}e^{-\frac{m_1}{\Omega_1}x_1^2},\ x_1\geq 0 \qquad (1)$$

$$p_{x_2}(x_2)=\frac{2}{\Gamma(m_2)}\left(\frac{m_2}{\Omega_2}\right)^{m_2} x_2^{2m_2-1}e^{-\frac{m_2}{\Omega_2}x_2^2},\ x_2\geq 0 \qquad (2)$$

Random variable x_3 has Rician distribution [8]:

$$p_{x_3}(x_3)=\frac{2(\kappa+1)}{\Omega_3}\sum_{i=0}^{\infty}\left(\frac{\kappa(\kappa+1)}{\Omega_3}\right)^{i}\frac{1}{(i!)^2}x_3^{2i+1}e^{-\frac{\kappa+1}{\Omega_3}x_3^2},\quad x_3\geq 0. \qquad (3)$$

Here Ω_1, Ω_2 and Ω_3 are powers of random variables x_1, x_2 and x_3 respectively, m_1 and m_2 are fading severity parameters of Nakagami-m RVs x_1 and x_2, and κ is Rician factor of RV x_3.

2.2 Product of three RVs

Random variable x is the product of x_1, x_2 and x_3 and is given by:

$$x=x_1\cdot x_2\cdot x_3. \qquad (4)$$

The first derivative of x is:

$$\dot{x}=\dot{x}_1x_2x_3+x_1\dot{x}_2x_3+x_1x_2\dot{x}_3. \qquad (5)$$

The mean value of x is:

$$\overline{\dot{x}}=\overline{\dot{x}}_1x_2x_3+x_1\overline{\dot{x}}_2x_3+x_1x_2\overline{\dot{x}}_3=0 \qquad (6)$$

since [9]

$$\overline{\dot{x}}_1=\overline{\dot{x}}_2=\overline{\dot{x}}_3=0. \qquad (7)$$

The variance of $\dot{x}$ is:

$$\sigma_{\dot{x}}^2=x_2^2x_3^2\sigma_{\dot{x}_1}^2+x_1^2x_3^2\sigma_{\dot{x}_2}^2+x_1^2x_2^2\sigma_{\dot{x}_3}^2 \qquad (8)$$

where:

$$\sigma_{\dot{x}_1}^2=\pi^2 f_m^2\frac{\Omega_1}{m_1} \qquad (9)$$

$$\sigma_{\dot{x}_2}^2=\pi^2 f_m^2\frac{\Omega_2}{m_2} \qquad (10)$$

$$\sigma_{\dot{x}_3}^2=\pi^2 f_m^2\frac{\Omega_3}{\kappa+1}, \qquad (11)$$

and f_m is maximal Doppler frequency.

After substituting, the expression for variance becomes:

$$\sigma_{\dot{x}}^2=\pi^2 f_m^2\left(x_2^2x_3^2\frac{\Omega_1}{m_1}+\frac{x^2}{x_2^2}\frac{\Omega_2}{m_2}+\frac{x^2}{x_3^2}\frac{\Omega_3}{\kappa+1}\right)=$$

$$=\pi^2 f_m^2x_2^2x_3^2\frac{\Omega_1}{m_1}\left(1+\frac{x^2}{x_2^2}\frac{1}{x_2^2x_3^2}\frac{\Omega_2}{m_2}\frac{m_1}{\Omega_1}+\frac{x^2}{x_3^2}\frac{1}{x_2^2x_3^2}\frac{\Omega_3}{\kappa+1}\frac{m_1}{\Omega_1}\right)=$$

$$=\pi^2 f_m^2x_2^2x_3^2\frac{\Omega_1}{m_1}\left(1+\frac{x^2}{x_2^4x_3^2}\frac{\Omega_2}{m_2}\frac{m_1}{\Omega_1}+\frac{x^2}{x_2^2x_3^4}\frac{\Omega_3}{\kappa+1}\frac{m_1}{\Omega_1}\right) \qquad (12)$$

The joint probability density function of x, $\dot{x}$, x_2 and x_3 is:

$$p_{x\dot{x}x_2x_3}(x\dot{x}x_2x_3)=p_{\dot{x}}(\dot{x}/xx_2x_3)\cdot$$

$$\cdot p_x(x/x_2x_3)\,p_{x_2}(x_2)\,p_{x_3}(x_3) \qquad (13)$$

The joint PDF of x and $\dot{x}$ is:

$$p_{x\dot{x}}(x\dot{x})=\int_0^{\infty}dx_2\int_0^{\infty}dx_3\,p_{x\dot{x}x_2x_3}(x\dot{x}x_2x_3)=$$

$$=\int_0^{\infty}dx_2\int_0^{\infty}dx_3\,p_{\dot{x}/xx_2x_3}(\dot{x}/xx_2x_3)\,p_x(x/x_2x_3)\,p_{x_2}(x_2)\,p_{x_3}(x_3)$$

where is:

$$p_x(x/x_2x_3)=\left|\frac{dx_1}{dx}\right|p_{x_1}\left(\frac{x}{x_2x_3}\right) \qquad (14)$$

$$\frac{dx_1}{dx}=\frac{1}{x_2x_3}. \qquad (15)$$

After substituting, the expression for $p_{x\dot{x}}(x\dot{x})$ becomes:

$$p_{x\dot{x}}(x\dot{x})=\int_0^{\infty}dx_2\int_0^{\infty}dx_3\,p_{\dot{x}}(\dot{x}/xx_2x_3)\cdot$$

$$\cdot\frac{1}{x_2x_3}p_{x_1}\left(\frac{x}{x_2x_3}\right)p_{x_2}(x_2)\,p_{x_3}(x_3). \qquad (16)$$

2.3 LCR of product of three RVs

Level crossing rate of x is [10]:

$$N_x=\int_0^{\infty}d\dot{x}\,\dot{x}\,p_{x\dot{x}}(x\dot{x})=$$

$$=\int_0^{\infty}dx_2\int_0^{\infty}dx_3\left(\int_0^{\infty}d\dot{x}\,\dot{x}\,p_{\dot{x}}(\dot{x}/xx_2x_3)\right.$$

$$\left.\cdot\frac{1}{x_2x_3}p_{x_1}\left(\frac{x}{x_2x_3}\right)p_{x_2}(x_2)\,p_{x_3}(x_3)\right)=$$

$$=\int_0^{\infty}dx_2\int_0^{\infty}dx_3\,\frac{1}{\sqrt{2\pi}}\sigma_{\dot{x}}\frac{1}{x_2x_3}p_{x_1}\left(\frac{x}{x_2x_3}\right)p_{x_2}(x_2)\,p_{x_3}(x_3)=$$

$$= \frac{1}{\sqrt{2\pi}} \pi f_m \frac{\Omega_1^{1/2}}{m_1^{1/2}} x^{2m_1-1} \cdot \frac{2}{\Gamma(m_1)} \left(\frac{m_1}{\Omega_1}\right)^{m_1} \frac{2}{\Gamma(m_2)} \left(\frac{m_2}{\Omega_2}\right)^{m_2} \cdot$$

$$\cdot \frac{2(\kappa+1)}{\Omega_3} \sum_{i=0}^{\infty} \left(\frac{\kappa(\kappa+1)}{\Omega_3}\right)^i \frac{1}{(i!)^2}$$

$$\cdot \int_0^{\infty} dx_2 \int_0^{\infty} dx_3 x_2^{-2m_1+1-1+2m_2-1} x_3^{-2m_1+1+2i+1-1} \cdot$$

$$\cdot e^{-\frac{m_1}{\Omega_1}\left(\frac{x}{x_2 x_3}\right)^2} e^{-\frac{m_2}{\Omega_2}x_2^2} e^{-\frac{\kappa+1}{\Omega_3}x_3^2} \cdot$$

$$\cdot \left(1 + \frac{x^2}{x_2^4 x_3^2} \frac{\Omega_2}{m_2} \frac{m_1}{\Omega_1} + \frac{x^2}{x_2^2 x_3^4} \frac{\Omega_3}{\kappa+1} \frac{m_1}{\Omega_1}\right)^{1/2} =$$

$$= \frac{1}{\sqrt{2\pi}} \pi f_m \frac{\Omega_1^{1/2}}{m_1^{1/2}} x^{2m_1-1} \frac{2}{\Gamma(m_1)} \left(\frac{m_1}{\Omega_1}\right)^{m_1} \frac{2}{\Gamma(m_2)} \left(\frac{m_2}{\Omega_2}\right)^{m_2} \cdot$$

$$\cdot \frac{2(\kappa+1)}{\Omega_3} \sum_{i=0}^{\infty} \left(\frac{\kappa(\kappa+1)}{\Omega_3}\right)^i \frac{1}{(i!)^2}$$

$$\cdot \int_0^{\infty} dx_2 \int_0^{\infty} dx_3 x_2^{-2m_1+2m_2-1} x_3^{-2m_1+2i+1} e^{-\frac{m_1}{\Omega_1}\frac{x^2}{x_2^2 x_3^2} - \frac{m_2}{\Omega_2}x_2^2 - \frac{\kappa+1}{\Omega_3}x_3^2}$$

$$\cdot \left(1 + \frac{x^2}{x_2^4 x_3^2} \frac{\Omega_2}{m_2} \frac{m_1}{\Omega_1} + \frac{x^2}{x_2^2 x_3^4} \frac{\Omega_3}{\kappa+1} \frac{m_1}{\Omega_1}\right)^{1/2} \tag{17}$$

Last integral can be solved using Laplace approximation theorem for solution two-fold integral [11]:

$$\int_0^{\infty} dx_2 \int_0^{\infty} dx_3 g(x_2, x_3) e^{\lambda f(x_2, x_3)} = \frac{\pi}{\lambda} \frac{g(x_{20}, x_{30})}{\left(B(x_{20}, x_{30})\right)^{1/2}} e^{\lambda f(x_{20}, x_{30})}, \tag{18}$$

where x_{20} and x_{30} are solutions of:

$$\frac{\partial f(x_{20}, x_{30})}{\partial x_{20}} = 0 \;\; ; \;\; \frac{\partial f(x_{20}, x_{30})}{\partial x_{30}} = 0 \; . \tag{19}$$

and:

$$B(x_{20}, x_{30}) = \begin{vmatrix} \frac{\partial^2 f(x_{20}, x_{30})}{\partial x_{20}^2} & \frac{\partial^2 f(x_{20}, x_{30})}{\partial x_{20} \partial x_{30}} \\ \frac{\partial^2 f(x_{20}, x_{30})}{\partial x_{20} \partial x_{30}} & \frac{\partial^2 f(x_{20}, x_{30})}{\partial x_{30}^2} \end{vmatrix} . \tag{20}$$

For our case, it is:

$$g(x_2, x_3) = x_2^{-2m_1+2m_2-1} x_3^{-2m_1+2i+1} \cdot$$

$$\cdot \left(1 + \frac{x^2}{x_2^4 x_3^2} \frac{\Omega_2}{\Omega_1} \frac{m_1}{m_2} + \frac{x^2}{x_2^2 x_3^4} \frac{\Omega_3}{\Omega_1} \frac{m_1}{m_3}\right)^{1/2} \tag{21}$$

$$f(x_2, x_3) = -\frac{m_1}{\Omega_1} \frac{x^2}{x_2^2 x_3^2} - \frac{m_2}{\Omega_2} x_2^2 - \frac{\kappa+1}{\Omega_3} x_3^2 \tag{22}$$

$$\frac{\partial f(x_2, x_3)}{\partial x_2} = \frac{2m_1}{\Omega_1} \frac{x^2}{x_2^3 x_3^2} - \frac{2m_2}{\Omega_2} x_2 \tag{23}$$

$$\frac{\partial f(x_2, x_3)}{\partial x_3} = \frac{2m_1}{\Omega_1} \frac{x^2}{x_2^2 x_3^3} - \frac{2(\kappa+1)}{\Omega_3} x_3 \; . \tag{24}$$

The x_{20} and x_{30} are the solutions for the last two equations to be equal to zero.

3 Presentation of numerical results

In this paper, LCR of product of two Nakagami-m random variables and Rician random variable is evaluated and presented in the next three figures. The impact of Nakagami-*m* parameters m_1 and m_2, Rician factor κ, Nakagami- *m* fading powers Ω_1 and Ω_2, and Rician multipath fading power Ω_3, on the LCR is analyzed.

In Fig. 1, LCR normalized by f_m is shown versus signal envelope x for some values of parameters m_1, Ω_1 and Ω_3. It is evident from this figure that LCR reaches maximum for very small x, and start to decrease for higher values of signal envelope. Therefore, the influence of signal envelope on the LCR is bigger for low values of signal envelope. For bigger Nakagami-*m* fading parameter m_1, LCR is smaller. It is also visible that LCR grows up with an increase of Nakagami-*m* fading power Ω_1 and Rician multipath fading power Ω_3.

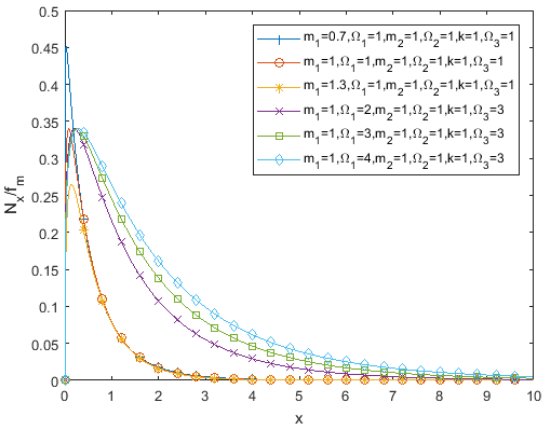


Figure 1. LCR normalized by f_m versus signal envelope for various values of parameters m_1, Ω_1 and Ω_3

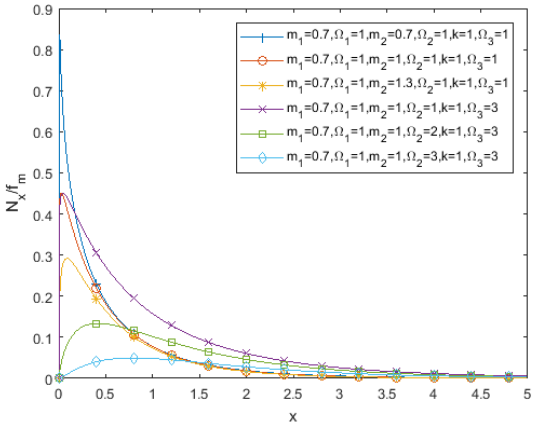


Figure 2. Normalized LCR for various m_2, Ω_2 and Ω_3

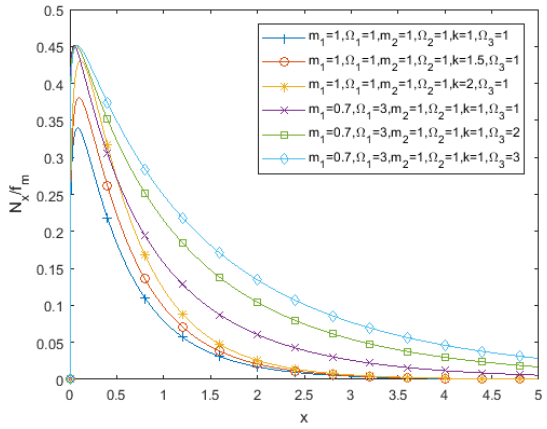


Figure 3. LCR normalized by f_m depending on signal envelope x for different values of parameters κ, Ω_1 and Ω_3

LCR versus signal envelope x for several values of Nakagami-m parameter m_2, and fading powers Ω_2 and Ω_3 is presented in Fig. 2. From Fig. 2, one can see that LCR decreases as Nakagami-m parameter m_2 increases. The influence of signal envelope on the LCR is greater for smaller values of Nakagami-m parameter. When power Ω_1 increases, the LCR wanes, and when Ω_3 enlarges, LCR enlarges too.

Normalized LCR depending on the signal envelope x, for different values of Rician factor κ, Nakagami- m fading power Ω_1, and Rician fading power Ω_3, is given in Fig. 3.

Can be noticed from the last figure that LCR decreases as Rician factor increases. The impact of signal envelope on the LCR is smaller for higher values of Nakagami-m parameter and higher values of Rician factor. The influence of Rician factor on the LCR is biger for greater values of Nakagami-m fading parameter. LCR decreases as Rician factor increases.

4 Conclusion

The product of two RVs has application in double and composite fading channels for characterizing shadowed fading in device-to-device channels. The product of N RVs is suitable for cascaded fading channel with N sections. In this work, we decide to examine triple fading channel and derived LCR for such scenario. As an application, this formula for LCR in closed form is used as the multihop wireless relaying system performance. The obtained closed form expression is verified for some values of fading powers and parameters.

The outcomes of the paper are expected to be useful for researchers studying fading models for multiple scattering radio propagation scenarios which, among other things, appear in forests and urban microcells via diffracting street corners.

It would be interesting to measure exact channel parameters which can verify the suitability of the proposed distribution of the product of double Nakagami and Rician processes and verify modeling of realistic wireless fading channels, which are theoretically analzyed here.

Acknowledgment

This paper is partially supported by the Ministry of Education, Science and Technological Development of Republic of Serbia within the projects TR-33035 and III-44006.

References

[1] G. L. Stuber: Mobile communication, 2nd ed. Dordrecht: Kluwer Academic Publisher, 2003.

[2] P. Mohana Shankar: Fading and shadowing in wireless systems, Springer, New York Dordrecht Heidelberg London, 2012, DOI 10.1007/978-1-4614-0367-8

[3] Z. Hadzi-Velkov, N. Zlatanov, and G. K. Karagiannidis: Level crossing rate and average fade duration of the multihop Rayleigh fading channel, IEEE International Conference on Communications, 19-23. May 2008, Beijing, China, DOI: 10.1109/ICC.2008.835

[4] N. Zlatanov, Z. Hadzi-Velkov, and G. K. Karagiannidis, "Level crossing rate and average fade duration of the double Nakagami-m random process and application in MIMO keyhole fading channels", IEEE Communications Letters, vol. 12, no. 11, pp. 822 – 824, Nov. 2008, DOI: 10.1109/LCOMM.2008.081058

[5] G. Karagiannidis, S. Nikos, and M. Takis: N*Nakagami: a novel stochastic model for cascaded fading channels, IEEE Transactions on Communications, Vol. 55, No. 8, 2007, pp. 1453–1458.

[6] E. J. Leonardo and M. D. Yacoub:, Statistics of the product of arbitrary α-μ variates with applications, 2014 IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications, 2-5 Sept. 2014, Washington, DC, USA, pp. 73-76, DOI: 10.1109/PIMRC.2014.7136135

[7] M. Nakagami: The m-distribution: A general formula of intensity distribution of rapid fading, in W. C. Hoffman (ed.), Statistical Methods in Radio Wave Propagation: Proceedings of a Symposium held June 18–20, 1958. Pergamon Press, New York, pp. 3–36, 1960.

[8] Rice, S. O., Mathematical Analysis of Random Noise. Bell System Technical Journal 24 (1945), pp. 46–156.

[9] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd. ed., New York, McGraw-Hill, 1991.

[10] S. Panic, M. Stefanovic, J. Anastasov, P. Spalevic: Fading and interference mitigation in wireless communications, USA: CRC Press, 2013.

[11] Jose L. Lopez and Pedro J. Pagola: A simplification of the laplace method for double integrals. application to the second appell function, Electronic Transactions on Numerical Analysis, ISSN 1068-9613, Vol. 30, 2008.

Multimedija

Multimedia

Pilotna rešitev za prikazovanje komplementarnih vsebin s pomočjo nadgrajene resničnosti

Matej Koplan, Klemen Pečnik, Matevž Pogačnik

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: matkoplan @gmail.com

A pilot augmented reality solution for displaying complementary content

Abstract. *This paper presents an implementation of a pilot augmented reality solution which allows users to view complementary content as holograms using a Microsoft HoloLens headset. The main advantage of the solution is being able to display complementary content without modifying the original, thus preserving the experience for the users without the headset.*

In this way we have also solved the ergonomic problems that come hand-in-hand with second-screen solutions such as the gorilla arm syndrome.

There are many potential use cases for this type of a solution, such as complementary information about/for: tourist attractions, museums, guided tours, education, cinema, and television channels

Although the application of this solution was developed for television channels, it can be applied to any of the listed options.

As a pilot project we have made an additional video stream for the deaf and hard of hearing. If applied by internet service providers, this solution would remove the need for duplicate video channels for the deaf, since the only difference is the sign language interpreter, effectively halving the bandwidth needed for each such channel.

The solution consists of a web application, containing a video player for the main content, a Microsoft HoloLens application, displaying complementary content and a NodeJS server, synchronising the two.

1 Uvod

1.1 Cilj projekta

Cilj tega projekta je bil izdelava aplikacije za prikaz dodatnih vsebin, ki nima slabosti trenutnih rešitev, in sicer s pomočjo nove tehnologije – z Microsoft HoloLens očali za mešano resničnost, ki prikazujejo holograme, s katerimi obogatimo prostor, v katerem se nahajamo.

Želeli smo rešiti več težav, ki se pojavljajo pri uporabi trenutnih rešitev za dodatno vsebino. Tovrstni pristopi k prikazovanju dodatnih vsebin so uporabni za številne scenarije, kot so uporaba v muzejih, galerijah in v turizmu, prikazovanje dodatnih vsebin v športu ali za izobraževalne namene, ne glede na področje uporabe pa tudi za nadomeščanje zvočnih informacij za gluhe in naglušne uporabnike.

Prva težava teh rešitev je 'ločenost' komplementarnih vsebin od glavne vsebine zaradi fizične razdalje med njimi.

Druga težava so ergonomske težave, ki pridejo z roko v roki skupaj z rešitvami tipa "second-screen", saj je velikokrat treba napravo držati v roki dalj časa neprekinjeno, kar predstavlja ergonomski problem. Pod to točko spadajo tudi vse ostale ergonomske težave, do katerih pride pri uporabi fizičnih rešitev.

Tretja s tem povezana težava je žrtvovanje glavne izkušnje za namene prikaza dodatnih informacij.

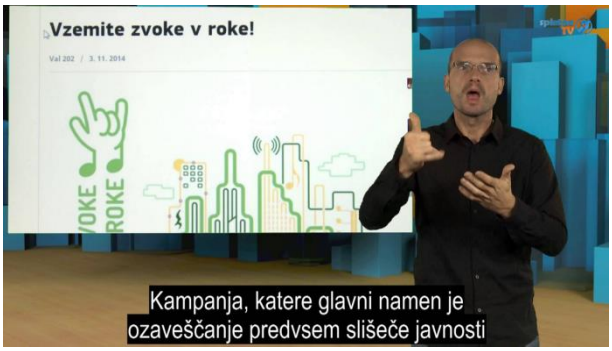
Vse znane rešitve imajo vsaj eno izmed naštetih težav, naš cilj pa je bil razvoj aplikacije, s katero bi se jim ognili.

1.2 Obstoječe rešitve

Nekaj rešitev z dodatnimi vsebinami že obstaja, a menimo, da ima vsaka vsaj kakšen problem, ki ga ne moremo prezreti, če hočemo uporabniku ponuditi kvalitetno izkušnjo. Te težave so večinoma ergonomske narave ali pa žrtvujejo kvaliteto izkušnje za uporabnike, ki dodatne vsebine ne potrebujejo.

Identificirali smo tri glavne obstoječe rešitve:

Prva rešitev so kanali [1], kjer je video v načinu slika v sliki (ang. picture in picture video). To je način, ki se trenutno uporablja v televizijskih programih. Ta rešitev prekrije del zaslona z osnovno vsebino, hkrati pa ga še zmanjša. Tako lahko gledalec izgubi pomembne informacije, ki bi jih na navadnem kanalu videl. Tako ta rešitev poslabša izkušnjo za vse gledalce, tudi za tiste, ki dodatno vsebino potrebujejo. Tovrstna rešitev velikokrat zahteva prenos dveh televizijskih kanalov, enega za navadne gledalce in enega za slušno prizadete. S tem operaterjem efektivno podvoji zahtevano pasovno širino za prenos televizijskega kanala.



Slika 1. Rešitev slika v sliki. [1]

Druga rešitev je rešitev tipa drugi zaslon (ang. second-screen) [2]. Takšna rešitev večinoma uporablja telefonske in tablične aplikacije na drugi fizični napravi. Le-te prikazujejo dodatne informacije, kot je statistika o

trenutni igri športa, dodatne informacije o artiklu, ki ga reklama predstavlja ipd.

Največja težava tega tipa rešitve je razmik med televizijo in dodatno napravo, kar oteži hkraten ogled obeh vsebin. To lahko rešimo s tem, da napravo držimo pred seboj, vendar je to za uporabnika precej utrujajoč položaj, ki ne more biti dolgo vzdrževan in lahko povzroči sindrom goriline roke (ang. gorilla arm syndrome [8]).



Slika 2. Rešitev tipa drugi zaslon podjetja 2Immerse. [3]

Tretja rešitev [3], ki jo je naredil laboratorij IRT, pa temelji na Microsoft Hololens očalih. Pojavila se je šele v končnih fazah izdelave naše rešitve in smo jo vzeli kot potrditev, da gremo v pravo smer.

Ta rešitev prikazuje hologram kretalca znakovnega jezika poleg televizijskega zaslona.

Osebi, ki še ni uporabljala HoloLens očal, lahko izgleda ta rešitev ustrezna, čeprav ima skrit problem. To je omejen kot vida. Takšna rešitev zahteva, da mora uporabnik gledati televizijo od daleč ali pa mora imeti glavo usmerjeno v hologram kretalca znakovnega jezika. Če uporabnik tega ne naredi, bo hologram ob robu odrezan ali pa sploh ne bo viden. Zaradi implementacije rešitve je okno kretalca tudi nemogoče zožiti.

Iz teh razlogov je ta rešitev precej neudobna za dolgotrajno uporabo, hkrati pa se je težko osredotočati na obe vsebini hkrati, saj je med njima veliko prostora.



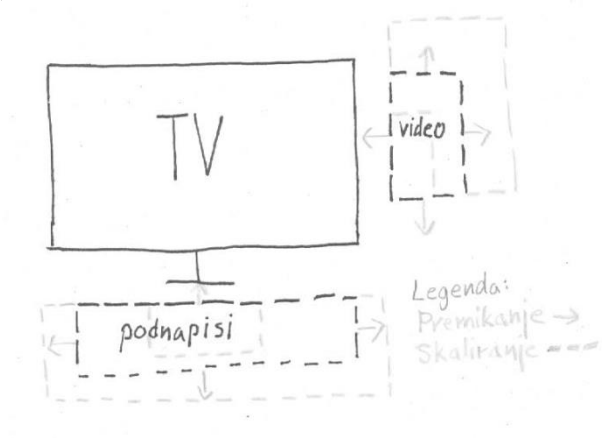
Slika 3. Rešitev laboratorija IRT. [3]

1.3 Zasnova naše rešitve

V projektu smo se osredotočili na nudenje dodatne vsebine za video gluhim in naglušnim.

Ugotovili smo, da potrebujemo dva hologrami. Enega za video kretalca, drugega pa za podnapise. Vsakemu smo določili kontrole za povečavo, premikanje in izbris.

Odločili smo se, da bomo gledalcu omogočili prilagoditev velikosti in lokacije teh dveh hologramov, saj niso vse televizije enako velike. Poleg tega smo tudi omogočili, da uporabnik posamezen hologram izklopi, če ga ne potrebuje.



Slika 4. Prva zasnova naše rešitve.

Slika 4 predstavlja naš začetni koncept izgleda aplikacije. Na sredini je televizor z glavno vsebino, zraven pa hologram z dodatno vsebino. Razporeditev je podobna sedanjim izkušnjam, ki jih imajo uporabniki – podnapisi na dnu slike so sedaj spodaj, dodatna vsebina, ki je bila prej v desnem predelu slike, pa je sedaj na desni strani zaslona.

Med zasnovo smo pazili na omejeno območje prikaza hologramov in na lahkost interakcije, saj lahko natančno klikanje majhnih elementov zelo hitro postane utrujajoče. Pri kliku je namreč potrebno element centrirati z obratom glave, ki pa se vedno nekoliko premika.

2 Izdelava rešitve

Izdelava je potekala v več fazah. Pri vsaki smo se soočili z različnimi izzivi.

2.1 Uporabljene tehnologije

Pri razvoju smo uporabili naslednjo opremo in tehnologije:

- Unity3D razvojno orodje za izdelavo video igrer, ki je hkrati uradno razvojno orodje za Microsoft Hololens aplikacije.
- Microsoft HoloLens SDK (ang. software development kit), ki v kombinaciji z Unity3D omogoči izgradnjo (ang. build) aplikacije in prenos le-te na očala.
- Visual Studio 2017 in Windows 10, saj sta zahtevana za razvoj Hololens aplikacij [6].
- NodeJS strežnik s procesnim upravljalnikom PM2.
- Spletne tehnologije za implementacijo video predvajalnika (HTML, CSS, JavaScript).
- HoloLens emulator in fizično napravo.

2.2 Interakcija z napravo in omejitve naprave

Med razvojem smo se spopadali z dvema glavnima omejitvama.

Prva omejitev je bila novost Microsoft HoloLens. To je pomenilo, da še ni bilo na razpolago veliko programskih knjižnic, ki bi pomagale pri razvoju takšnih aplikacij, tiste, ki pa so že bile, pa so bile še vedno v povojih. Poleg tega je bilo pri razvoju treba upoštevati, da HoloLens SDK ne podpira vseh funkcionalnosti razvojnega orodja Unity3D. Posledično smo morali veliko stvari implementirati sami.

Druga težava je bila omejen vidni kot, ki ga očala omogočajo. To pomeni, da elementi niso smeli biti preveliki, saj ne želimo prekriti glavne vsebine.

2.3 Implementacija uporabniškega vmesnika in HoloLens različica

Najprej smo naredili osnovo uporabniškega vmesnika v Unity3D brez HoloLens SDK in implementirali interakcijo, potem pa smo jo prenesli na očala.



Slika 5. Hologram podnapisov in video predvajalnika.

Med prenosom aplikacije na očala HoloLens smo opazili, da navadne ‘onClick’ (koda, ki se izvede ob kliku elementa) funkcije, ki jih Unity3D ponuja, ne delujejo na njej, čeprav se je v Unity3D emulatorju vse obnašalo po pričakovanjih.

Za rešitev te težave smo implementirali “poslušalca” (ang. “listener”, programerski izraz za element v kodi, ki čaka na določen dogodek in ko pride do njega, aktivira drugo programsko kodo). S tem smo dosegli, da so se gumbi obnašali po pričakovanjih tudi na napravi.
V tej fazi smo še vedno reševali problem ozkega vidnega kota, ki ga naprava omogoča.

Holograme smo postavili fiksno v prostor z možnostjo, da jih uporabnik po želji prilagodi. Pri takšni implementaciji se efektiven prostor za holograme še dodatno zmanjša, saj je treba preprečiti, da se holograms odrežejo, ko uporabnik nekoliko premakne glavo. To pomeni, da morajo biti holograms postavljeni z nekaj rezervnega prostora, saj uporabniki vedno vsaj malo premikajo glavo. Primerna razdalja od roba vidnega polja je odvisna od uporabnika in jo je možno tudi prilagoditi s povečavo ali premikanjem elementov.

2.4 Predvajanje video posnetka in podnapisov ter sinhronizacija

V tej fazi smo implementirali prenos podnapisov in videa.

Prenos podnapisov smo naredili s preprostim HTTP zahtevkom. Prikazujemo jih z uporabo interpreterja, ki smo ga implementirali sami. Zaradi razširjenosti smo uporabili format WebVTT [7].

Vsebino polja za podnapise nadzira sinhronizacijska koda, ki prejema informacije od kontrolnega strežnika.

Video zaradi performačnih razlogov, ki izvirajo iz prenosa videa, ne sme biti prevelik.. Ker Unity3D v kombinaciji z Microsoft Hololens nativno ne podpira pretočnega videa (ang. video streaming), ga je potrebno prenesti vnaprej. Za tovrsten prenos potrebujemo nekaj sekund, preden je video pripravljen, kar je še v okviru sprejemljivih mej. Upamo, da bo Unity kmalu v svoji programski knjižnici dodal podporo za pretočni video.

2.5 Glavni predvajalnik, strežnik in sinhronizacija

V končni fazi smo implementirali še sinhronizacijo med video predvajalnikom in očali.

V spletni aplikaciji smo uporabili HTML5 video predvajalnik. Napisali smo preprosto JavaScript kodo, s katero je možno ugotoviti status videa in predvajani delež ter posredovati te informacije strežniku.
S to kodo informacije pošljemo strežniku do desetkrat na sekundo, a le v primeru sprememb (preskok v videu ali njegova zaustavitev). Strežnik si zapomni zadnje informacije o statusu in jih ob zahtevi Hololens aplikacije posreduje naprej.

Hololens aplikacija s pomočjo sinhronizacijske kode ugotovi delež preteklega videa in status videa (to je pavza ali predvajanje). To naredi tako, da upošteva čas, ko je bila informacija poslana s spletne aplikacije, delež pregledanega posnetka ob času pošiljanja in status predvajanja. Da dosežemo želeno natančnost, je pomembno, da je podan čas naveden v milisekundah.

Čas, ko je bila informacija poslana, se primerja s časom HoloLens očal in s časom videa. S pomočjo tega lahko ugotovimo, kakšen mora biti čas videa na naši strani, in sicer s preprosto formulo (čas HoloLens – poslan čas + čas videa = trenutni čas videa).

Izračunan čas se uporabi za sinhronizacijo podnapisov in videa.

S tem pristopom trajanje zakasnitve prenosa informacij iz predvajalnika do strežnika in naprej do HoloLens naprave nima vpliva na natančnost sinhronizacije. Zanašamo se le na natančnost ur v predvajalniku glavne vsebine (TV-program) in HoloLens napravi.

Ugotovili smo, da je dovolj, če se sinhronizacija izvede le nekajkrat na sekundo.

2.6 Stopnja sinhronizacije

Pri razvoju smo testirali več različnih stopenj sinhronizacije. Pri visokih stopnjah (60 Hz ali višje) smo opazili performančne probleme, pri nizkih (1 Hz ali manjše), pa je bilo skakanje pri zaustavljanju videa preveč opazno. Na osnovi poskusov smo izbrali 10 Hz, kot dovolj dobro stopnjo brez performančnih problemov.

Ker naša sinhronizacija temelji na urah naprav, je možno, da avtomatski popravki ure, ki se na današnjih napravah dogajajo precej pogosto, povzročijo napake. Za rešitev tega problema smo predpostavili, da je popravek ure skoraj vedno majhen (manjši od 100 ms) in razlike v času, ki so manjše od tega, preprosto ignorirali. To je

bistveno izboljšalo gladkost dodatnega videa in podnapisov.

Na specifični rešitvi, ki smo jo razvili, smo ugotovili, da je potrebno upoštevati gibanje kretalca. Za razliko od podnapisov, ki jih lahko beremo tudi med pavzo, ali pa govora, kjer lahko tudi z delom besede ugotovimo, kaj je bilo izrečeno, se pri kretanju ob pavziranju izgubi veliko več pomena. To smo rešili tako, da Hololens, kljub temu da že prejme informacijo o pavzi, dodatno video vsebino nadaljuje še 1.5 sekunde, kar je dovolj, da kretalec konča začeto gesto. Pri nadaljnjem predvajanju se ta del predvaja še enkrat.

2.7 Območje vida

Kot projeciranja hologramov na očala Hololens je zelo omejen in znaša 45 stopinj [5].

Približek vidnega polja po širini si lahko najlažje predstavljamo, če iztegnemo roki predse, razširimo prste in se z enim palcem dotaknemo drugega. Če upoštevamo, da hoče uporabnik imeti v centru vida še kaj drugega, se zavemo, da pravzaprav ni veliko prostora in se moramo omejiti na manjše število elementov, ki jih porazdelimo po vidnem polju tako, da ne zavzamejo preveč prostora.

3 Možne izboljšave

V sklopu tega projekta smo naleteli na problem prenosa videa, saj Unity3D ne podpira prenosa videa v živo. To bi se dalo potencialno rešiti na več načinov in bi bilo tudi potrebno v primeru končne rešitve, saj bi lahko čakanje na prenos dolgega videa trajalo predolgo za dobro uporabniško izkušnjo.

Identificirali smo tri možne rešitve, ki bi najbolje in najbolj izvedljivo rešile omenjeno težavo. Vsaka ima svoje prednosti in slabosti.

Prva rešitev je prireditev videa, kot da bi prihajal iz spletne kamere. Jasna slabost je to, da tovrstne kamere niso temu namenjene in bi implementacija zahtevala veliko prilagoditev, saj bi bilo prejete informacije med tekom programa potrebno pretvoriti v teksture, ki bi jih potem ‘prilepili’ na element v Unity3D, ki predstavlja video predvajalnik, kar pa je zelo zahtevno.

Druga rešitev je uporaba Microsoft Azure pretočnega videa, edinega načina, ki ga HoloLens nativno podpira. Ta je od vseh najbolj preprosta, poleg tega pa se je v zadnjem času tudi bistveno pocenila. V času pisanja tega članka je mesečna cena storitve, ki premore pasovno širino 600 Mb/s, kar zadostuje za do 200 uporabnikov hkrati, 54 € [4], kar predstavlja za veliko ustanov še sprejemljive stroške.

V primeru, da plačevanje za storitev ni možno, pa obstaja tudi tretja možnost, ki je implementacija lastnega protokola za psevdo-pretočni video.

Druge izboljšave zahtevajo izboljšavo vidnega polja naprave.

4 Sklep

Menimo, da je projekt uspešen, saj smo uspeli narediti rešitev, ki:

- je vsebinsko primerljiva obstoječimi rešitvami, saj podobno kot rešitev “slika v sliki” uporabi podnapise in kretalca znakovnega jezika. S tem zajame vse prednosti omenjene rešitve in lahko postane njena nadgradnja,
- odpravi probleme dvojnega prenosa vsebin in s tem povezanega sočasnega oddajanja dveh programov. Oddaja se le še originalen program, dodatno vsebino pa se lahko prenese tudi drugače,
- ne zniža kvalitete izkušnje gledalcu brez očal, saj osnovna vsebina ostane enaka,
- omogoča, da gledalec zlahka sledi hkrati glavni in dodatni vsebini v udobnem položaju, ker se ne zanaša na naprave, ki jih je treba držati v rokah,
- dobro izkoristi območje za izris hologramov in je uporabna za številne scenarije, ne glede na domeno.

Rešitev bi bilo v prihodnosti, ko se bo vidni kot izboljšal, možno uporabiti tudi v kinu, kjer do sedaj ni bilo cenovno smiselno narediti posebnega predvajanja filma za gluhe in naglušne uporabnike. Z našo rešitvijo, pa to ni več potrebno.

5 Zahvale

Zahvaljujemo se podjetju SportRadar, ki je laboratoriju za multimedijo (LMMFE) na Fakulteti za elektrotehniko Univerze v Ljubljani posodilo očala za razvoj aplikacij, in laboratoriju, ki je omogočil razvoj te rešitve.

Poleg tega se zahvaljujemo vsem, ki so pomagali pri razvoju te rešitve z nasveti in mnenji.

Literatura

[1] Spletna TV, S kampanjo Zvoke v roke o pomembnosti znakovne govorice, https://www.youtube.com/watch?v=FsSuWsxmzSU
[2] Rešitev 2Immerse, https://2immerse.eu/
[3] Rešitev laboratorija IRT, https://lab.irt.de/synchronised-holograms/
[4]Cenik Microsoft Azure storitev za pretočni video, https://azure.microsoft.com/en-us/pricing/details/media-services/
[5] Specifikacije Microsoft HoloLens očal, https://docs.microsoft.com/en-us/windows/mixed-reality/Hololens-hardware-details
[6] Zahteve za razvoj Microsoft HoloLens aplikacij https://docs.microsoft.com/en-us/windows/mixed-reality/install-the-tools
[7] WebVTT format podnapisov, https://developer.mozilla.org/en-US/docs/Web/API/WebVTT_API
[8] Gorilla arm syndrome, https://www.deccanchronicle.com/lifestyle/health-and-wellbeing/230617/beware-of-gorilla-arm-syndrome.html

Tehnologije navidezne resničnosti za prikaz športnih dogodkov

Nejc Povšič, Matevž Pogačnik

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: np2758@student.uni-lj.si

Use of virtual reality technology to display sports statistics

Abstract. *This paper addresses a mobile application which uses virtual reality technology to display sports matches and statistics. It was designed for the Android mobile operating system with the use of the Unity 3D tool, a graphics engine that allows for very fast and efficient development of graphically intensive games and simulations. The application is created for the ever-present sports market using technology that has taken the field by storm. The application displays a football game in a virtual world in conjunction with some statistics information about both teams and a complete timeline of the event. Interaction is performed with the user's gaze, which controls a cursor that can trigger various actions, such as playing or pausing a video, setting the video player to full screen, selecting a timeline event, etc. This paper explains how the application was created and which methods and technologies were used in its development. It also touches upon the possible future development in the case of a successful productivisation.*

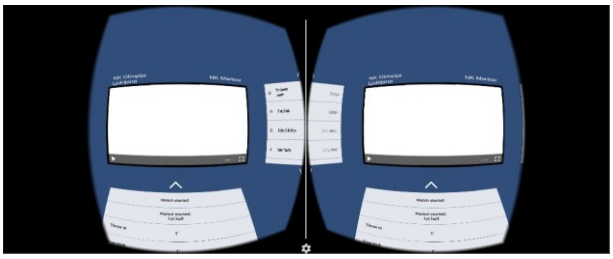
1 Uvod

Z uporabo navidezne resničnosti smo ustvarili mobilno aplikacijo za gledanje nogometa v virtualnem svetu. Uporabnik je postavljen v umetno ustvarjeno okolje, kjer ima pred seboj zaslon, ki predvaja nogometno tekmo, poleg zaslona pa so postavljeni seznami s podatki o sestavi ekipe in seznami s časovnico tekme, ki vsebuje vse omembe vredne dogodke. Športne statistike za aplikacijo so bile pridobljene od podjetja Sportradar, ki se ukvarja z agregacijo in analizo športnih podatkov.
Aplikacija je narejena za operacijski sistem Android z uporabo orodja Unity, igralnega pogona, ki doda vmesni nivo abstrakcije, saj sam poskrbi za računanje v prostoru, izrisovanje in grajenje za določeno platformo, tako da se lahko razvijalec posveti stvarem, ki se vsebinsko tičejo njegovega projekta. Prav tako ima zelo dobro spisano dokumentacijo in podporo uporabniške skupnosti, zanj pa je ustvarjenih ogromno vtičnikov, ki razvijalcem olajšujejo delo.

2 Uporabljene tehnologije za izvedbo rešitve

2.1 Navidezna resničnost

Virtualna ali navidezna resničnost (ang. »Virtual Reality - VR«) je definirana kot računalniška simulacija, pri kateri ima uporabnik občutek, da se nahaja v umetnem okolju [1]. Kot okno v to okolje uporabimo VR vizir, ki ma vgrajeni dve leči, vsako posebej usmerjeno proti določenemu očesu, naši možgani pa nato iz obeh pogledov sestavijo enotno sliko, ki ustvarja videz, da smo v drugem svetu. Vizir ima lahko zaslona za vsako oko že vgrajena ali pa je narejen tako, da vlogo zaslona prevzamejo zunanje naprave kot so mobilni telefoni, ki jih moramo vstaviti v sam vizir.
Prikaz navidezne resničnosti se od klasičnega prikaza, ki smo ga vajeni, precej razlikuje, kot vidimo na spodnjih slikah 1 in 2. Za razliko od slednjega imamo pri navidezni resničnosti namesto ene kamere, skozi katero opazujemo sceno, sedaj dve. To pomeni, da mora grafična procesna enota celotno sceno izrisati dvakrat in to z večjim vidnim kotom, dvakrat izločiti elemente, ki niso vidni, preslikati tridimenzionalno sceno na dvodimenzionalni zaslon za vsako kamero, na vsaki kameri dodati distorzijo, ki pomaga pri prikazu skozi leče, vse to pa pomeni veliko več dela, ki ga mora računalnik opraviti [2].



Slika 1: Prikaz scene v navidezni resničnosti



Slika 2: Prikaz scene na klasičen način

Da človek zares dobi občutek, da je v navideznem svetu, mora prikaz navidezne resničnosti slediti premikom glave in, če je potrebno, tudi premikom celotnega telesa. Pri tem je zelo pomembna latenca ali zakasnitev, torej čas med sprožitvijo akcije in prejemom odgovora nanjo. Poskrbeti moramo, da je ta čim manjša, da uporabnika ne zmoti in mu ne uniči predstave, saj lahko v skrajnem primeru vodi tudi do simulacijske bolezni (slabost, izguba orientacije) [3],
Da ohranjamo občutek navidezne resničnosti, je zelo pomembna tudi ločljivost zaslona, ki ga opazujemo. Obraz imamo namreč zelo blizu ekrana, zato se lahko

zgodi, da vidimo posamezne slikovne elemente (piksle), kar je pri virtualni resničnosti skrajno nezaželeno.
Računalnik opravi vse potrebne izračune in izriše sceno na zaslon, a za prikaz potrebujemo neko napravo, da dve različni sliki, ki sta na zaslonu, prikažemo tako, da ju naši možgani lahko zlijejo. Tu pride v poštev že omenjeni VR vizir.
Zelo pomembna komponenta VR vizirjev so leče. Te imajo dve glavni nalogi: dovolj povečati prikaz zaslona, tako da ne vidimo robov, in ustvariti gorišče v neskončnosti [4]. Objekti na zaslonu, ki je za lečo, tako izgledajo oddaljeni dlje, kot dejansko so. Leče tudi popačijo sliko, ki jo dobijo od zaslona, zato ji je treba na zaslonu dodati inverzno popačitev, da jo naše oči vidijo čim bolj resnično [4].

2.2 Unity

Unity je platforma za razvoj iger, v katerem lahko na preprost način ustvarjamo tri- ali dvodimenzionalne računalniške igre s pomočjo najsodobnejših tehnik in tehnologij, ki so na voljo. Ideja Unity-ja je bila, po besedah ustvarjalcev, »demokratizacija razvijanja iger«, kjer bi lahko vsakdo, ne glede na poklic in izobrazbo, ustvarjal računalniške igre [5]. Uporablja ga ogromno število razvijalcev, saj je enostaven za uporabo, ima zelo dobro dokumentacijo, veliko skupnost, je zelo zmogljiv, dobro podprt in razširljiv. Zanj je ustvarjenih mnogo vtičnikov, ki uporabnikom še dodatno olajšajo delo, čeprav jih je večina plačljivih. Z Unity-jem lahko zgradimo igro za 27 različnih platform, med katerimi je tudi Android.
Unity podpira dva programska jezika, v katerih lahko programiramo, C#, ki je uraden programski jezik orodja in jezik, v katerem je logika aplikacije spisana, in Javascript, ki pa ga počasi umikajo. Vse skripte za delovanje v Unity-ju so sestavljene iz dveh glavnih delov, iz funkcije *Start*, ki se zažene takoj, ko je skripta dodana v sceno, in iz funkcije *Update*, ki se požene pri vsakem izrisu nove sličice.
Velika prednost Unity-ja je razširljivost, saj omogoča, da v projekt dodajamo razne vtičnike. Za želeno delovanje aplikacije smo jih morali nekaj dodati. Med njimi so bili GoogleVR, ki ga je ustvaril Google za potrebe platform Daydream in Cardboard ter je dodal nek nivo abstrakcije, saj je v ozadju poskrbel za prikaz virtualne resničnosti in pomagal pri gradnji za Android operacijski sistem, vtičnik za predvajanje posnetkov Easy Movie Texture, ki omogoča predvajanje videov v prostoru na poljubnem objektu, in vtičnik za interpretiranje oblike JSON v programskem jeziku C#.

2.3 Android

Android je mobilni operacijski sistem za pametne telefone, ki je zgrajen na Linuxovem jedru. Prva različica je bila izdana leta 2008 in je bila na voljo le za peščico mobilnih naprav, ker pa je Android odprtokodni sistem, ga je kmalu začelo uporabljati vse več proizvajalcev mobilnih naprav. Tako je počasi pridobival na popularnosti in je do danes postal najbolj razširjen mobilni operacijski sistem.

2.4 Strojna oprema

Za prikaz aplikacije smo potrebovali Android telefon in VR vizir, v katerega smo lahko mobilno napravo vstavili. Uporabljen je bil telefon LG G6, ki vsebuje vse potrebne senzorje in je dovolj zmogljiv, da aplikacijo poganja učinkovito in brez zatikanja. Prikazan je na sliki 3.



Slika 3: Mobilni telefon LG G6

Da virtualna resničnost lahko deluje, mora telefon nekako zaznati svoj položaj in orientacijo v prostoru. Moderne naprave to dosegajo s pospeškomerom in žiroskopom.
Pospeškomer (ang. »accelerometer«) je, kot ime nakazuje, zmožen meriti veličino in smer pospeška ter s tem tudi sile, ki vplivajo nanj. Te so lahko dinamične, kot so recimo sile pri premikanju naprave, ali statične, kot je Zemljina gravitacija. Zmožen je zaznati tudi blage spremembe rotacije naprave, ampak je pri tem zelo nenatančen, potrebuje pa tudi pomoč vgrajenega kompasa [6]. Android zna iz pospeškomera pridobiti veličino pospeška v vseh treh oseh (prišteto skupaj z gravitacijo ali samo linearni pospešek brez gravitacije).
Žiroskop (ang. »gyroscope«) je naprava za merjenje kotne hitrosti naprave, ki je predstavljena v stopinjah ali radianih na neko časovno enoto. S tem omogoča zaznavanje hitrosti rotacije naprave v vseh treh oseh, s čimer lahko zelo točno določimo orientacijo telefona v prostoru.

Za VR vizir smo izbrali preprost VR BOX, ki ima prostor v katerega vstavimo telefon in dve leči, ki jih lahko premikamo naprej, nazaj, levo in desno, da si pravilno nastavimo razdaljo do očes in med očesoma. Če imamo razdaljo pravilno nastavljene, lahko naši možgani brez problemov sliki združijo v prostorsko celoto. Vizir ne vsebuje nobenih senzorjev, saj vse delo opravi telefon s svojimi senzorji. Viden je na sliki 4.



Slika 4: VR BOX vizir

3 Športni podatki

Športne podatke, uporabljene v aplikaciji, smo pridobili od podjetja Sportradar, ki se ukvarja z urejanjem in analizo športnih podatkov za večino obstoječih športov v okviru katerih potekajo tekmovanja. Za najpopularnejše, kot je npr. nogomet, imajo na voljo več vrst podatkov, kot so celotni urnik turnirjev, statistika ekip, statistika posameznih igralcev, primerjava dveh ekip, statistika tekem, itd.
Največji del aplikacije in njen namen je seveda prenos nogometne tekme. Za potrebe demonstracije naše rešitve je bil izbran posnetek tekme med Mariborom in Olimpijo, ki se je odvijala 8. avgusta 2016.

3.1 Športni trg

Šport je gonilo za ogromno novih tehnologij, predvsem na področju multimedije. V prenosih športnih dogodkov se že nekaj časa uporablja obogatena resničnost (npr. namišljena pot žoge pri nogometu), vedno bolj pa se veča zanimanje tudi za navidezno resničnost.
Pojavili so se že 360 stopinjski posnetki iz tekem za navidezno resničnost, ki gledalca postavijo na sam štadion (tudi za svetovno prvenstvo v nogometu 2018 so za določene tekme na voljo prenosi, namenjeni naglavnim zaslonom). Nekateri so virtualno resničnost uporabili kot orodje za analizo preteklih tekem ali za učenje novih spretnosti [7]. Možnosti za implementacijo navidezne resničnosti je na športnem trgu ogromno in naša rešitev je le ena izmed mnogih, ki so na voljo.

4 Rešitev virtualnega prikaza tekem in statistik

4.1 Funkcionalnosti aplikacije

Aplikacija vsebuje samo dve sceni, med katerima lahko menjamo. Po končani inicializaciji in začetnem zaslonu se odpre prva scena, v kateri je v prostoru dodan samo gumb za odpiranje druge scene. V osnovi aplikacija omogoča ogled športne tekme s prikazom statistik tekme, ekip igralcev in seveda rezultat.
Za uporabo aplikacije potrebujemo VR vizir, čeprav deluje tudi brez njega. Trenutno je omogočeno samo rotacijsko sledenje glavi, kar pomeni, da se ne moremo premikati po navideznem prostoru, ampak sceno opazujemo le iz ene točke.

Ker je aplikacija narejena za VR naglavni sistem brez dodatnih gumbov, je krmiljenje zelo omejeno, saj imamo na voljo le pogled. Edini način interakcije z okoljem je usmerjanje pogleda na interaktivne objekte, ki akcijo sprožijo instantno ali z nekim časovnim zamikom. Ko s kazalcem pristanemo na predmetu, ki ima nase vezano neko akcijo, se, če je bilo določeno, pokaže odštevalnik časa. Če takrat, ko se odštevanje konča, še vedno opazujemo isti objekt, se akcija izvede. Ta sistem je za navidezno resničnost brez dodatnih krmilnikov najbolj ustrezen.
Ko aktiviramo gumb v prvi sceni, se odpre druga, ki vsebuje prikaz tekme in podatkov. Postavljeni smo v nek prostor, kjer imamo postavljena 4 platna. Na sredini je zgoraj predvajalnik video posnetka, spodaj časovnica, levo in desno pa je prikaz postavitve obeh ekip, kot je prikazano na sliki 5.



Slika 5: Druga scena aplikacije

Tudi tukaj ima zelo veliko vlogo kazalec, saj zoper ga zopet potrebujemo za interakcijo z objekti v okolju. Vsak interaktivni objekt, ki ga pogledamo, povzroči, da se kazalec poveča, kar nakazuje, da je z njim povezana neka akcija. Na sezname so dodane puščice, ki ob pogledu premikajo seznam navzgor ali navzdol, saj je podatkov preveč, da bi jih prikazali naenkrat.

4.2 Arhitektura

Za potrebe aplikacije smo postavili centralen strežnik, kjer so športni podatki zapisani v mongoDB podatkovni bazi. Da smo podatke lahko posredovali do aplikacije, smo ustvarili REST API končne točke, na katere smo potem izvajali klice s protokolom http (ang. »Hypertext Transfer Protocol«). Aplikacija posnetek pridobiva iz strežnika z uporabo protokola HLS (ang. »Http Live Streaming«)
Kot omenjeno, sta v aplikaciji na voljo dve sceni. Prva ima vlogo začetnega menija, kjer lahko trenutno samo zaženemo drugo sceno, a se gumbi lahko prosto dodajajo. Njena hierarhija je zato zelo preprosta, saj vsebuje le pred nastavljen objekt za kamero, globalno luč, prazen objekt, na katerega vežemo skripte in pred nastavljen objekt, ki vsebuje interaktivne gumbe. Pred nastavljeni objekti oz. *prefabi*, so v Unity-ju zelo koristni, saj nam dovolijo, da vstavimo identičen objekt v več scen, ali pa ga vstavimo večkrat v sklopu posamezne funkcije.
Na objekt kamere je seveda vezan tudi kazalec, ki se skupaj s njo transformira, tako da vedno stoji točno na sredini pogleda. Deluje na principu sledenja žarku (ang.

»Ray tracing«), ki je način preverjanja kolizij med žarkom in objekti. Predstavljajmo si ga, kot da iz našega pogleda v smeri kazalca sije žarek, ki seka vse predmete na njegovi poti. Da vpliva na nek predmet, moramo temu dodati komponento za detekcijo trkov in poslušalec dogodkov, ki zajame dogodek, ko se ta sproži.
Zraven kazalca smo dodali tudi objekt za časovno zamaknjeno akcijo. Pri nekaterih objektih nismo želeli, da se akcija izvede takoj, ker lahko uporabnik nenamerno pogled usmeri v objekt. Da bi si olajšali delo, smo dodajanje števca hoteli narediti čim bolj enostavno, saj ga je bilo potrebno dodati na kar nekaj objektov.
Ko se druga scena zažene, se takoj na začetku izvede klic na strežnik za podatke, v skripto, ki skrbi za nalaganje posnetka, pa se nastavi URL naslov video posnetka. Klic na strežnik se izvede v posebni funkciji, ki jo omogoča Unity, imenovani *Coroutine*. Ta poskrbi, da izvajanje ni vezano na posodobitev okvirjev v *Update* funkciji, ampak teče vzporedno in je pri vsaki posodobitvi zmožna nadaljevati od tam, kjer je končala izvajanje. Ker je pridobivanje podatkov iz strežnika asinhrona operacija in nočemo, da nam prekine delovanje aplikacije dokler traja, je uporaba tega tipa funkcije nujna.
Ko športni podatki pridejo do odjemalca, se seznami napolnijo z njimi. Za vsak element se v sceno doda nov pred nastavljen objekt, ki vsebuje skripto za nadzor vseh komponent za prikaz besedila, ki se v Unity-ju imenuje *Text*. Glede na tip podatka (časovnica ali postavitev) smo ustvarili dva pred-nastavljena objekta, ki imata seveda drugačno postavitev.
Video posnetek se med tem prav tako naloži preko spleta na izbrano platno z uporabo Easy Movie Texture vtičnika. Uporaba je zelo preprosta; sam vtičnik nam da dostop do poslušalcev, ki opazujejo ali se je posnetek naložil, ali je prišlo do napake, itd. Javno dostopne so tudi funkcije, ki skrbijo za ustavljanje ali predvajanje posnetka, za premikanje po posnetku in ugotavljanje časa trajanja ter koliko časa je že preteklo pri predvajanju.

4.3 Problemi pri razvoju

Pri razvoju rešitve smo naleteli na nešteto težav, ki smo jih morali rešiti. Ena izmed največjih, vsaj v sklopu virtualne resničnosti, je bila učinkovitost aplikacije. Narejena je za mobilne naprave, ki so zelo omejene v procesorski in grafični moči. S tem, ko smo zmanjšali podrobnosti, izključili sence, zmanjšali ločljivost tekstur in uporabili materiale, ki niso odbijali svetlobe, smo občutljivo izboljšali delovanje aplikacije in zmanjšali porabo baterije.
Glavni del aplikacije, prikazovanje tekme na nekem samostojnem zaslonu v okolju, bi bilo na Android napravi zelo težko implementirati, saj Unity privzeto podpira samo prikazovanje video posnetka čez celoten zaslon. Rešitev smo našli v vtičniku Easy Movie Texture. Z njim je bil prikaz tekme zelo nezahteven proces, saj podpira tudi strujanje preko http protokola.

5 Zaključek

V članku je opisana izdelava delujočega prototipa mobilne aplikacije, ki je osnovana na virtualni resničnosti. Namenjena je športnim navdušencem, ki bi radi tekme spremljali na nov način z uporabo najnovejših tehnologij računalniške grafike.
Seveda aplikacija še ni končna. Trenutno je rešitev implementirana samo za točno določeno tekmo, dalo pa bi se jo razširiti, da bi podpirala več tekem in turnirjev. Dodati bi se morale nove končne točke na strežniku in omrežni klici iz naprave, ki bi do teh točk dostopale. Podatki bi se tudi morali sami posodabljati.
Sam grafični vmesnik je dokaj preprost in vsebuje samo komponente, ki so nujne za delovanje, ker je aplikacija omejena s strojno opremo telefonov, zato je tudi izgled sveta zelo generičen in dokaj pust. Z razvojem mobilnih naprav se bo izboljševala tudi strojna oprema, kar pomeni več procesorske in grafične moči ter posledično več prostih rok pri ustvarjanju virtualnega sveta.

Literatura

[1] Navidezna resničnost. Dosegljivo: https://sl.wikipedia.org/wiki/Navidezna_resni%C4%8Dnost [Dostopano: 3.4.2018]

[2] Stereoscopic Rendering in WebVR. Dosegljivo: https://hacks.mozilla.org/2015/09/stereoscopic-rendering-in-webvr/ [Dostopano 6.5.2018]

[3] The Technical Challenges of Virtual Reality. Dosegljivo: https://iq.intel.com/the-technical-challenges-of-virtual-reality/ [Dostopano 6.5.2018]

[4] VR Lens Basics: Present And Future. Dosegljivo: https://www.tomshardware.com/news/virtual-reality-lens-basics-vr,36182.html [Dostopano: 18. 6. 2018]

[5] Unity (game engine). Dosegljivo: https://en.wikipedia.org/wiki/Unity_(game_engine) [Dostopano: 23.5.2018]

[6] D. Avguštin, Analiza tresljajev s senzorji v mobilnih napravah. Dosegljivo: http://eprints.fri.uni-lj.si/2269/1/Avgu%C5%A1tin_D%2D1.pdf [Dostopano: 28.5.2018]

[7] How virtual reality is transforming the sports industry. Dosegljivo: https://techcrunch.com/2016/09/15/how-virtual-reality-is-transforming-the-sports-industry/ [Dostopano: 18. 6. 2018]

Urban gamification with Augmented Reality and Geo-Referencing: An Innovative and Engaging Cultural Heritage App

Klemen Babuder
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
klemen.babuder@ltfe.org

Vid Stropnik
Faculty of Electrical Engineering
University of Ljubljana
Velenje, Slovenia
vid.stropnik@ltfe.org

Jure Idzig
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
jure.idzig@gmail.com

Emilija Stojemnova Duh
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
emilija.stojmenova@fe.uni-lj.si

***Abstract*—Creating an application, focused on the promotion and demonstration of local cultural heritage can prove to be a task of ample difficulty. It is paramount to recognize that the aforementioned solution is designed to tackle a problem with which the target audience doesn't concern themselves habitually. Thus, the approach towards publicizing a smartphone solution for such a problem should be comprehensive, continual, inclusive and resourceful. The fact that we now live both in the age of mobile-first software development and oversaturation of application markets, only further emphasizes the importance of studying the state of the art. Furthermore, this should not only be done in the field of mobile solutions promoting cultural heritage, but also in those similar to it - be it in the way that they both strive to engage their audience with urban exploration or presentation and discovery of specific content. In addition to that, studying the advances in modern mobile technology may yield other use cases that can be used in such an application, perhaps in ways similar solutions are yet to implement.**

The authors of this article have aimed to include the overview of all above-mentioned fields herein, with the goal of making kulTura, their application for the promotion of cultural heritage in the city of Jastrebarsko and the municipality of Črnomelj, as alluring and modern as possible.

Keywords—Cultural heritage, AR, urban gamification, UX, content discovery, geo referencing

I. INTRODUCTION

The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines *Cultural heritage* as the *legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations*.[1] The institution does not specify in which way the artifacts of the past should be entrusted to the descendants of their creators, but to benefit from their cultural significance, the latter generation need to, above all else - comprehend their value.

It is no mystery that the current generation's lives heavily depend on the usage of mobile devices. In 2018, 52.2% of all worldwide online traffic was generated through smart phones,[2] with the 79% of internet users in the European Union using their smart phones to surf the web in 2016.[3] Furthermore, in 2017 the world's largest video streaming service – YouTube, announced that over 1 billion hours of content is being broadcasted every single day[4], with 70% of it being done on mobile platforms.[5]

These numbers alone speak volumes about the fact that the generation of today can be approached through the mobile world in an impactful way. It is here that we receive information and expect to be communicated with, with 75% of all US digital ad spend going towards native mobile advertisements.[6] Hence, mobile applications are a suitable means for our comprehension of - and consequential benefit from cultural heritage, now more than ever.

The authors of this article aim to create such an application in the following years. As part of the kulTura project funded by The European Union through the Interreg Slovenia-Croatia programme and the Slovenian Research Agency, it will seek to inform its users about the tradition and legacy of over 50 time-honored smaller towns along the Slovenian-Croatian border, starting with the cross-border pair of the Municipality of Črnomelj in Slovenia and the city of Jastrebarsko in Croatia.

However, said mobile solution is to be fashioned in such a way, that the possibility of it realizing its intent – bestowing the cultural heritage of these locations upon younger generations, will be as high as possible. Therfore, we decided to preface our work with a thorough review, resulting in this article. As the explosive growth of the mobile application market has made it a significant challenge for the users to find interesting applications in crowded App Stores[7], we concluded that a study of the state of the art is one of the methods required to accomplish this task. Furthermore, this should not only be done in the field of mobile solutions promoting cultural heritage, but also in those like it - be it in the way that they both strive to engage their audience with urban exploration or the presentation and discovery of specific content. In addition to that, studying the advances in modern mobile technology may yield other use cases that can be used in such an application, perhaps in ways similar solutions are yet to implement. The

overview of our study of these fields can be found within the following sections.

II. State of the art

When conducting the study of the newest ideas and features, as well as the highest level of general development in a specific field, it is important to adequately define it. In our case, this field does not only envelop all other solutions, oriented towards the propagation of cultural heritage, but also all of those that strive to either inform their users of points of interest in their surroundings, or educate them about a certain topic. Furthermore, any successful app features that attempt to diversify a user's – be it a tourist's, urban explorer's or local's interaction with the world around them, either through gamification or enhancing the world through the inclusion of multimedia content or AR technology, are worth considering as well.

Many of this section's subsidiaries will refer to an application aspect's effect on User experience (UX). UX refers to a person's emotions and attitudes about using a particular product, system or service. Factors influencing UX have been classified into three main categories: user's state and previous experience, system properties and the usage context.[8] While aesthetic and functional design of a service or product, respectively make it visually appealing and optimally operative, user experience design makes sure these aspects of the product work in the context of the rest of it.[9]

During the process of our study, we examined 16 applications [10] - [25] from a broad scope of categories: from those very similar to our cause, to those that only incorporate certain features of interest to us. In addition to those, we also researched several city-focused or regional applications and those that focus on pathfinding. All applications were reviewed in July of 2018 on the Android operating system running Android version 8.0.0.

The remainder of this section will describe, explain and highlight good practices of the most cutting-edge aspects found in the analyzed applications, as well as emphasize any other related conclusions about said features we made while researching them.

A. Inclusion of and redirection to third-party services

Several local or heritage focused applications can benefit a lot from third-party services. This may include, but is not limited to weather and traffic information, ticket reservations and purchases, pathfinding, etc. By integrating already polished, more focused solutions into their app, developers can dedicate more time and resources towards disruptive innovation, rather than work towards something that already exists. Some companies, like Google with its Maps Platform[26], even build their business model around third-party applications integrating their services in the form of application programming interfaces (API).

With the fact that each of a services' functions needs to be evaluated in the context of its entirety in mind, developers should be wary of including any features that redirect to external sources in their applications without proper optimization. This is largely on account of these third-party services having different user interfaces – with the desired contexts of components vastly different from those of the original app.

The aforementioned importance of caution is emphasized when competing with solutions from companies spearheading modern UX design – for example the American travel and restaurant website company TripAdvisor, that doesn't only include third-party services, but makes them a cornerstone of its functionality. While certain reviewed applications only use hardcoded information and simple hyperlinks that lead to sites offering lodging or navigation these features are all integrated into the TripAdvisor application in the form of a browser extension – ticket purchases and reservations can all be done entirely inside the app and in most if not all cases require no redirection, while still offering great detail in communicating all relevant information for the specific use-case. Furthermore, the design complements the content in such a way, that nothing feels out of place.

While redirection towards services that have their native smartphone applications already pre-installed on the operating system may work just as well, we recognized the described approach as the best practice for including external services.

B. Gamification

Gamification refers to the use of game design elements within non-game contexts.[27] This method employs the use of 'building blocks' of games and implements them into real world-settings, often with the goal of motivating specific behaviors within the gamified situation.[28] As the practice has proven to be a success in achieving several different goals, notably the improvement of user engagement[29][30] and learning[31], it is a concept worth exploring in any field, not just application design. Furthermore, it is expected that gamification can also foster the initiation or continuation of goal-directed behavior – i.e. motivation.[32]

Some of the applications we reviewed during the process of our study focus on gamifying the acts of walking, sightseeing and exploration. This is the case with Nexto and Six to Start's The Walk: Fitness tracker game. On the other hand, some others, like Pokemon GO and Ingress, are branded as games, but heavily incorporate the aforementioned acts into their core mechanics.

We found several good executions of sightseeing gamification in *Nexto*. Throughout a guided tour Legends of Piran, the user is met with a plethora of simple game-like mechanics, ranging from quizzes to AR-powered puzzles. Two of the aspects of these minigames stood out as good practices: Their simplicity and diversity. The term 'sustained attention' is often used in reference to a putatively unitary capacity to remain engaged in tasks that are lengthy or characterized by long intervals between relevant events.[33] The gamification aspects of the application not being repetitive is a worthwhile aim of trying to retain both sustained attention and audience engagement, while them (as a service's component) being

The research presented in this paper was funded by The European Union through the Interreg Slovenia-Croatia programme and the Slovenian Research Agency.

simple corresponds with the Nielsen Norman Group's definition of an exemplary user experience.[34]

The application also utilizes fully voiced storytelling, which is a popular game component on its own, with Time magazine's 2014 Game of the Year creative director Jon Ingold calling video games *'the most fascinating medium for storytelling available right now.'[35]* In addition of interactive storytelling being a trending game design element,[36] building parts of a business around a story is also a global trend in the experience industry, revealing new insights into conceptualizing tourism via story-telling tourist guides.[37]

Another popular form of gamification is the implementation of a visible representations of a player's achievements in the form of collectibles or badges,[38] items such as virtual souvenirs, star tokens or achievement trophies, amassed by performing application related tasks like completing tours or correctly answering quizzes. They confirm the players' achievements, symbolize their merits[39] and serve to numerically represent a player's progress.[40] Additionally, badge-owners can also exert social influence on players and co-players, particularly if they are rare or hard to earn.[41]

Niantic's *Pokemon GO* itself is focused around the amassing of collectible virtual creatures, while also highlighting another interesting form of gamification – a mechanic in which a reward is presented to the user upon completing a certain number of steps.

C. Augmented Reality (AR)

Augmented Reality (AR) is an interactive experience of a real-world environment whereby the objects that reside in the real-world are "augmented" by computer-generated perceptual information. This is sometimes done across multiple sensory modalities, including visual, auditory, haptic, somatosensory, and olfactory.[42]

The usability of this technology resides in the ability to manifest experiences that would otherwise be unreasonably expensive or simply impossible to produce. As our analysis of the reviewed applications will show, when implemented correctly, AR can be used to simulate the inclusion of virtual 3D objects into a real environment in real time – cutting costs and expended time in comparison with manufacturing such objects, while also making them accessible to a wider audience through the inclusion of AR on a user's mobile platform.

The technology's appeal can also be credited to the fact that the emulation of reality and the act of overlaying computer-generated imagery or other synthetic content over physical objects have been some of the more common science-fiction tropes since the late 20th century, prominently featuring in works and franchises like Star Trek[43], Ender's Game[44] and Iron man[45]. However, the capabilities of AR have not yet caught up with the imagination of sci-fi creators and currently still face difficulty displaying objects on non-flat surfaces whilst taking in account camera movements effectively making sure that the projected object abides by the laws of gravity. In spite of its technical limitations, AR offers a very useful tool for creating an engaging user experience, as proven by the reviewed applications below.

The content used most widely in cultural heritage apps relating to AR regularly involves projecting images onto a specific surface. Google Arts & Culture, for example, enables the user to project famous paintings on empty walls, simulating the experience of visiting it in an actual museum. The Nexto app goes even a dimension further by allowing the user to view the 3D Model of a famous composer's violin in the real world. In other instances, the application augments elements of one's surroundings in the pursuit to showcase the state of an object or room in times past – a very useful use-case for an application, showcasing cultural heritage.

However, the UX of applications that include this feature is heavily reliant on the capabilities of the user's terminal. For reference, roughly 77% of mobile devices are running android[46], out of which only around 43% support ARCore - Google's AR software development kit (SDK)[47][48]. On the other hand, Apple's AR SDK - ARKit is supported on 81% of all mobile devices running iOS.[49][50] This represents almost 19% of the mobile OS market share.[51] Consequentially, only close to 50 % of all smartphones can run either ARKit or ARCore. This is mostly due to holographic processing being a computationally-demanding task, as illustrated by leading AR-dedicated devices, like the Microsoft HoloLens headset, including separate processing units, solely devoted to hologram-related tasks.[52] However, several other mobile AR SDKs are available for both Android and iOS. One of them – Vuforia is the kit used in Nexto. These third-party solutions enable the use of AR content for the vast majority of smartphones.[53] In order to support such a large range of less-capable devices, Vuforia uses image recognition technology to target specific real-world objects (targets)[54] in order to correctly project holograms onto actual surfaces. Hence only by using targets can an AR experience reach the majority of end users. The problem with target, as mentioned before, is they have to be present at a given location for the AR content to work. More importantly, the target must remain in sight of the camera or just barely outside of it.

D. Personalization

Personalization is defined as the act of designing or tailoring to meet an individual's specifications, needs, or preferences.[56] With personalization, apps can sift through large arrays of their content and present only that, which is relevant to the preferences of the user. Perhaps the biggest indicator of the value of personalization is the fact that some of the largest companies in the world massively invest in perfecting personalization algorithms.[57] Google, for example, use these algorithms for both delivering relevant ads or further improving their services,[58] some of which were also part of the described study. They are not alone in this – with other examined applications also including features for curating the content they display to an individual user.

Thorough studies have already been conducted in the field of recommender systems and collaborative filtering. The former can be separated into content-based, collaborative and hybrid systems. Content-based approaches recommend content

similar to that a user has liked in the past.[59] As such, the user has to rate a certain number of items for the system's understanding of their preferences – hindering the experience for new users of the platform. In contrast, collaborative filtering systems try to predict the utility of items for a particular user based on the items previously rated by other users, often taking into account the user's similarity to specific user groups by gender, age, area code, education, employment information etc. Additionally, modern iterations of these systems use advanced technologies such as machine learning, clustering and artificial neural networks to determine additional similarities between users. While a pure collaborative recommender improves upon the content-based system's shortcomings, it too has shortcomings – notably with recommending newly added or less exposed items, with a sparse amount of ratings.[60] Due to the drawbacks of both kinds of systems, hybrid approaches have emerged, combining both methods.[61]

Considering the level of complexity of recommender system studies, further expanded upon in the sources referenced in the last paragraph, it would not be very cost and time efficient for developers of a dedicated heritage focused application to work towards a bleeding edge solution in this field. However, some of the reviewed applications have incorporated more basic means of recommending content. Culture trip's approach for instance, gives its users the ability to "like" certain content and later modifies the section displayed on the default Explore page – a content-based approach, augmented with some collaborative elements, mostly utilizing the user's location. Google trips on the other hand, applies Google's own technology by reserving a dedicated 'For you' section, specifically for content that the app deemed the user might enjoy, taking in account the content that the user expressed a liking for, as well as he user's location and search history.

E. Geo-location specific features

Culture and heritage-oriented apps can be separated into two categories. Some, for example Google arts and Culture, aim to bring attractions to the user, no matter where the user is currently located. These types of solutions often offer several features for discovery of new artworks, artists and mediums, as well as include content like written articles, to be enjoyed at the user's leisure. In contrast, other applications (Slovenia's top 100, Nexto, Google trips, Visit London) focus on orienting the user towards the attraction and augmenting the experience of enjoying them in the real world. Admittedly, many of the reviewed applications strive to achieve both, but do prioritize one aspect of use over the other, upon closer examination.

Most of the applications in the latter of the two categories provide several features, relating to the geographic location of the user. Among others, features like visualizing the real world in map form, routing the user from their current location to different points of interest and the discovery of relevant nearby locations are of note here. While some of these features can be done via redirection to third-party services (as is done in Culture Trip and Visit London), advanced use like map customization requires a dedicated in-app solution.

With 99% coverage of the world and 1 billion monthly active users, the Google maps platform is the biggest actor in the geo-location related industry.[62] By using its API or an API of one of its competitors,[63] developers can showcase relevant content on interactive maps, with several added features at their disposal. Nexto, for example, highlights already visited locations in a guided tour differently than those, which the user is yet to visit, also showcasing the virtual souvenirs acquired at a certain location, right on the map. Use cases like this can improve user experience, and with it – engagement, as they allow for further design in the context of the remainder of the application.

Similarly, a lot of applications include a dedicated 'near me' feature, highlighting the points of interest, that are closest to the user, sometimes even combining geo-specific features with their personalization and recommendation systems, as seen in Culture trip.

F. Offline usage

Internet access is seeing better coverage and lower access costs, with 4G covering an average of 91% of EU's population.[64] However, offline features still have a place in mobile applications. In our use case, we can highlight Croatia having the second-lowest DESI connectivity score in the EU[65] as one of the reasons for their importance, but others might include personal preference, privacy, or battery-saving, as the technology behind connecting the 4G LTE network and end-devices is very battery consuming.[66]

While some apps with a thinner range of multimedia content opt to provide most of their services offline by default (Slovenia's Top 100), this isn't a valid solution in applications that strive to continuously upkeep and update their content. In addition, an offline-only application may also consume an amount of storage users aren't willing to dedicate to it.

Most state of the art apps therefore include options to download the most basic type of content e.g. an article (Visit London, Culture Trip, Beyondarts), allowing the users to handpick the content they would like to consume offline. Nexto and Google Trips on the other hand, allow the user to download and locally store large bundles of multimedia content (several times the size of the app itself), relevant to a larger area, containing numerous interest points. By doing this, these applications avoid the negative user experience of downloading data to consume offline, only to find out they want to consume more shortly after. Furthermore, this approach also enables the presentation of the same, wholly experience they offer to connected users, offline.

G. User generated content

Several of the examined applications welcomed some sort of personal input by their users. In most cases, for example in Foursquare or Tripadvisor, this user-generated content (UGC) appeared in the form of user reviews and scores. We recognized that these elements mostly benefited the user in situations, where they would be choosing between competing services: e.g. restaurants or lodging. One study finds that the valence of traveler reviews has a significant impact on the online sales of hotel rooms, as these reviews may serve to

reduce the cognitive load of potential travelers, and thus increase their awareness.[67] This same reduction of the cognitive load may also influence the simplicity of the user experience of a potential application, which may in turn, as mentioned before, improve engagement.

In addition to collecting valuable feedback, a comprehensive rating and review system may also aid the users lacking the time or desire to experience all of the landmarks a given location has to offer. It is important to note that this may actually hinder the experience, however, if the goal of our application is to present the cultural heritage of a certain area in its entirety – as the better reviewed portion of points of interest would receive more exposure than its lesser counterpart.

Only one of the reviewed applications – Nexto, featured a different type of UGC. However, we still recognized it carries particular importance and usefulness considering the developmental circumstances of our solution.

Most of the reviewed cultural heritage applications incorporate content created either by in-house experts or by an outside team of content creators. In this way, the solution's content is guaranteed to be tailored to its desired standards and – at the same time, is exclusive to it. Google Arts & Culture's Home tab, for example, boasts a plethora of expertly written editorial features, written by experts in their field, hand picked by the application team to ensure diversity and quality. This sort of content creation can, however, prove costly and time consuming when contemplating the scaling of an application – thus an alternative might be considered.

This is where Nexto tries to incorporate UGC. The application provides users with a web-based tool, enabling the production of specific multimedia experiences for certain cultural hotspots. While the tool utilizes a very well designed and intuitive UI, with no coding required, Nexto Inc.'s team of heritage interpreters, designers and 3D artists is also available to design said experiences by commission.

While relying solely on UGC isn't a sensible move in designing such an application due to the variability in the content's quality, we recognized several positives in incorporating it – such as the utilization of user engagement to target long-term interactions, loyalty and advocacy through word-of-mouth.

III. Conclusion and Future Work

Throughout the process of our study and the writing of this article, we attempted to feature aspects of applications that stood out as good practices in their specific use cases. It is worth noting, however, that our intent was not to imply that all of the described features should be incorporated in a successful heritage-focused application.

When designing an application, one of the key steps is to define its target audience. The application user's demographic, interests and goals should be heavily considered when deciding on which of the above features to incorporate, and to what degree. Audience groups vary heavily between analyzed applications: from tourists to locals, from those who seek to experience the cultural heritage digitally, from afar or those who seek to engage with it in person.

Furthermore, the authors of this article would argue that including all of the described functionalities could even result in a hindered UX. Instead of implementing all of them, features inspired by those described in sections A through G should only be implemented upon deciding that they will indeed benefit the application's userbase. Otherwise, the oversaturation of features can contradict the application's simplicity – already established as one of the core principles of UX beforehand. Furthermore, an application is never only a sum of its features. Creating a successful and engaging solution may be aided by several techniques not described herein, such as creating a comprehensive marketing plan, following platform-specific UI design guidelines, or the employment of design thinking during the course of development.

References

[1] Tangible Cultural Heritage (n. d.). Retrieved from http://www.unesco.org/new/en/cairo/culture/tangible-cultural-heritage/

[2] Percentage of all global web pages served to mobile phones from 2009 to 2018 (2018, January). Retrieved from https://www.statista.com/statistics/241462/global-mobile-phone-website-traffic-share/

[3] Statistical Office of the European Communities (2016, December 20). *Almost 8 out of 10 internet users in the EU surfed via a mobile or smart phone in 2016…* Retrieved from http://ec.europa.eu/eurostat/documents/2995521/7771139/9-20122016-BP-EN.pdf/f023d81a-dce2-4959-93e3-8cc7082b6edd

[4] Goodrow, C. (2017, February 27). *You know what's cool? A billion hours*. Retrieved from https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html

[5] Lella, A. (2017, February 23) *Unlocking Mobile Measurement for YouTube in the U.S.* Retrieved from https://www.comscore.com/Insights/Data-Mine/Unlocking-Mobile-Measurement-for-YouTube-in-the-US/

[6] Koetsier, J. (2018, February 23) *Mobile Advertising Will Drive 75% Of All Digital Ad Spend In 2018: Here's What's Changing.* Retrieved from: https://www.forbes.com/sites/johnkoetsier/2018/02/23/mobile-advertising-will-drive-75-of-all-digital-ad-spend-in-2018-heres-whats-changing/#5e516c39758b

[7] Yan, B., & Chen, G. (2011, June). AppJoy: personalized mobile application discovery. In *Proceedings of the 9th international conference on Mobile systems, applications, and services* (pp. 113-126). ACM.

[8] Hassenzahl, M. & Tractinsky, N. (2006). User Experience – a Research Agenda. In *Behaviour and Information Technology, Vol. 25, No. 2, March-April 2006* (pp. 91-97). Archived 2014, October 20 at the Wayback Machine.

[9] Garrett, J. J. (2010). *Elements of user experience, the: user-centered design for the web and beyond*. Pearson Education.

[10] Dspot (2018). *iKranj* (1.3.10) [Mobile application or software]. Retrieved from: https://play.google.com/store/apps/details?id=si.dspot.ikranj

[11] London & Partners Ltd (2017). *Visit London Official City Guide* (2.20.8.190) [Mobile application or software]. Retrieved from:

https://play.google.com/store/apps/details?id=com.londonandpartners.londonguide

[12] Slovenian Tourist Board (2015). *Slovenia's Top 100* (1.0) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=info.slovenia.top100.app

[13] TripBucket (2015). *World Heritage Sites* (4.4) [Mobile application or software]. Retrieved from: https://play.google.com/store/apps/details?id=com.tripbucket.unesco

[14] Culture Trip (2015) *Culture Trip: Explore & Travel* (2.1.2) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=culturetrip.com

[15] TripAdvisor (2010) *TripAdvisor Hotels Flights Restaurants Attractions* (27.2) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.tripadvisor.tripadvisor

[16] Google LLC (2015). *Google Arts & Culture* (6.3.12) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.google.android.apps.cultural

[17] Nexto inc. (2016). *Nexto* (1.7.6) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=io.nexto

[18] Niantic, Inc. (2012). *Ingress* (1.132.2) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.nianticproject.ingress

[19] Niantic, Inc. (2016) *Pokemon GO* (0.111.3) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.nianticlabs.pokemongo

[20] beyondarts (2014). *Beyondarts Art & Culture Guide* (2.0) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.beyondarts.app

[21] Touch Screen Travels (2016). *Bologna + Modena Art & Culture Travel Guide* [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.facarospauls.bologna

[22] Google LLC (2016). *Google Trips – Travel Planner* (1.10.0.205086730) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.google.android.apps.travel.onthego

[23] Foursquare (2018). *Foursquare City Guide* (11.9.1) [Mobile application or software]. Retrieved from https://play.google.com/store/apps/details?id=com.joelapenna.foursquared

[24] Six to Start (2017). *The Walk: Fitness Tracker Game (FREE)* (2.1.1) [Mobile application or software]. Retrieved from: https://play.google.com/store/apps/details?id=com.sixtostart.thewalk2

[25] Niantic, Inc. (2012) *Field Trip* (2.0.9) [Mobile application or software]. Retrieved from: https://play.google.com/store/apps/details?id=com.nianticproject.scout

[26] Welcome to Google Maps Platform (n.d.). Retrieved from https://cloud.google.com/maps-platform/

[27] Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). *From Game Design Elements to Gamefulness: Defining "Gamification"*, presented at the 15th International Academic MindTrek Conference, Tampere, 2011. New York, NY: ACM.

[28] Sailer, M., Hense, J. U., Mayr, S. K. & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. In *Computers in Human Behavior, Volume 69, April 2017* (pp. 371-380).

[29] Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. In *Computers in Human Behavior, Volume 71, June 2017* (pp. 469-478)

[30] Ruhi, U. (2015). Level Up Your Strategy: Towards a Descriptive Framework for Meaningful Enterprise Gamification. In *Technology Innovation Management Review, Volume 5, Issue 8, August 2015.*

[31] Hamari, J., Shernoff, D. J., Rowe, E., Coller. B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. In *Computers in Human Behavior, Volume 54, January* 2016 (pp. 133–134)

[32] Schunk, D. H., Pintrich, P. R. & Meece, J. L. (2010). *Motivation in education: Theory, research, and applications* (3 rd ed.), Pearson, Upper Saddle River

[33] Grahn, J. A., Manly, T. (2012). *Common Neural Recruitment across Diverse Sustained Attention Tasks*. Retrieved from https://doi.org/10.1371/journal.pone.0049556

[34] Norman, D., Nielsen, J. (n.d.). *The Definition of User Experience (UX)*. Retrieved from https://www.nngroup.com/articles/definition-user-experience/

[35] Burks, R. (2015, September 14). *How Video Games Have Become The Perfect Storytelling Medium*. Retrieved from https://www.techtimes.com/articles/84016/20150914/how-video-games-have-become-the-perfect-storytelling-medium.htm

[36] Blacker, A. (2018, March 05). *Interactive Story Games Are Trending*. Retrieved from http://blog.apptopia.com/interactive-story-games-tap-into-our-love-for-storytelling

[37] Mossberg, L. (2008). Extraordinary Experiences through Storytelling. In *Scandinavion Journal of Hospitality and Tourism, Volume 8, Issue 3: Strategy in Hospitality Manegement, 2008*. (pp. 195 – 210)

[38] Werbach, K., Hunter, D. (2012*). For the Win: How game thinking can revolutionize your business*. Philadelphia: Wharton Digital Press.

[39] Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013). *Steering user behavior with badges*, presented at the 22nd international conference on World Wide Web, Rio de Janeiro.

[40] Werbach, K., Hunter D. (2015). *The gamification toolkit - dynamics, mechanics, and components for the win.* Philadelphia: Wharton Digital Press.

[41] Hamari, J. (2013). Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. In *Electronic commerce research and applications, Volume 12, Issue 4, July-August 2013(*pp. 236-245).

[42] Schueffel, P. (2017). The Concise Fintech Compendium. *School of Management Fribourg/Switzerland*.

[43] Roddenberry, G. (Producer), & Butler, R. (Director), (1988). The Cage (Star Trel: The Original Series). United States: Culver Studio.

[44] Card, O. S. (1991). Ender's Game. 1985. *Author's Definitive Edition." NY: Tor*.

[45] Favreau, J., Arad, A., & Feige, K. (2008). Iron Man.[Motion picture]. *USA: Paramount Pictures*.

[46] Percentage of all mobile devices running android (2018, June). Retrieved from: http://gs.statcounter.com/os-market-share/mobile/worldwide

[47] Percentage of android devices supporting ARCore SDK (2018, July) Retrieved from: https://developer.android.com/about/dashboards/

[48] Devices supporting ARCore SDK (2018, July). Retrieved from: https://developers.google.com/ar/discover/supported-devices

[49] Precentage of devices running iOS supporting ARKit (2018, May). Retrieved from: https://developer.apple.com/support/app-store/

[50] Documentation of ARKit SDK (n.d.). Retrieved from: https://developer.apple.com/documentation/arkit

[51] Percentage of all mobile devices running iOS (2018, June). Retrieved from: http://gs.statcounter.com/os-market-share/mobile/worldwide

[52] Hardware details of the Microsoft Hololens headset (2018, March). Retrieved from: https://docs.microsoft.com/en-us/windows/mixed-reality/hololens-hardware-details

[53] Mobile devices supporting Vuforia SDK (n.d.). Retrieved from: https://vuforia.com/devices.html

[54] How image targets work (n.d.). Retrieved from: https://library.vuforia.com/content/vuforia-library/en/articles/Training/Image-Target-Guide.html

[55] Documentation on extended tracking (n.d.). Retrieved from: https://library.vuforia.com/articles/Training/Extended-Tracking

[56] Defenition of personalization (n.d.). Retrieved from: https://www.dictionary.com/browse/personalization?s=t

[57] Amount of money spent on Netflix search engine per year (2018, June). Retrieved from: https://www.businessinsider.com/netflix-recommendation-engine-worth-1-billion-per-year-2016-6

[58] How google uses gathered data for improving their services. (n.d.) Retrieved from: https://policies.google.com/technologies/partner-sites

[59] Mooney, R. J., & Roy, L. (2000, June). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries* (pp. 195-204). ACM.

[60] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6), 734-749.

[61] Lewis, D. D., & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In Machine Learning Proceedings 1994 (pp. 148-156).

[62] Data on Google Maps Platform (n.d.). Retrieved from: https://cloud.google.com/maps-platform/

[63] Website of google maps competitor skobbler (n.d.). Retrieved from: https://www.skobbler.com/apps

[64] Broadband connectivity in Europe (2018). Retrieved from: https://ec.europa.eu/digital-single-market/en/connectivity

[65] DESI report on connectivity in the EU (2018). Retrieved from: http://ec.europa.eu/information_society/newsroom/image/document/2018-20/1_desi_report_connectivity_DFB52691-EF07-642E-28344441CE0FCBD1_52245.pdf

[66] Effect of LTE on battery life (2015, April). Retrieved from: https://www.extremetech.com/mobile/110711-what-is-lte/2

[67] Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human behavior, 27(2), 634-639.

Mobilna aplikacija za predstavitev produkta s pomočjo obogatene resničnosti

Luka Bezovšek, Matevž Pogačnik, Jože Guna

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: luka.bezovsek@gmail.com

Mobile Application for Product Presentation with Augmented Reality

Abstract. *Presentation of any physical product is most effective when the product is presented directly at the location in a tangible way. Since it is difficult or impossible for some products to be carried around due to large dimensions, some other alternative solutions are necessary. Augmented reality enhanced applications can be used to tackle these issues. In this paper, we introduce the augmented reality technologies and apply them to enhance a presentation of an autonomous robot. Based on Vuforia and Unity software development kits, interactive mobile application is developed and evaluated in a basic user study.*

1 Uvod

Kadar predstavljamo fizični produkt, želimo karseda natančno prikazati vse poglavitne lastnosti le-tega. Tovrstna predstavitev največji učinek doseže takrat, ko poslušalcem pokažemo pravi, oprijemljivi izdelek. Na tem mestu se zlasti pri objektih velikih dimenzij pojavi problem njihove prenosljivosti. Skupaj z razvojem sodobnih tehnologij v zadnjem času, se je pojavila tudi tehnologija obogatene resničnosti, ki klasične predstavitve lahko dvigne na povsem nov nivo.

Namen tega članka je predstavitev mobilne aplikacije za predstavitev produkta z uporabo tehnologije obogatene resničnosti, ki smo jo razvili s pomočjo razvojnih okolij Vuforia in Unity.

Glavno vprašanje na katerega želimo odgovoriti v članku je, ali je tehnologija obogatene resničnosti primerna za predstavitev fizičnih objektov v navidezni obliki oz. ali si uporabnik na podlagi interaktivnih virtualnih objektov ustvari dovolj dobro predstavo o dejanskem fizičnem produktu.

Na to in še nekaj drugih vprašanj bomo poskušali odgovoriti s pomočjo testiranja aplikacije na dejanskih uporabnikih.

V prvem delu članka predstavimo nekaj osnovne teorije o tehnologiji obogatene resničnosti. Teorijo v nadaljevanju dopolnimo s predstavitvijo okolja za implementacijo aplikacij navidezne resničnosti Vuforia. Članek nadaljujemo z opisom implementacije in delovanja aplikacije ter ga zaključimo s predstavitvijo rezultatov testiranja aplikacije, podajanjem ocene opravljenega dela in predstavitvijo možnosti za nadaljevanje.

2 Obogatena resničnost

Obogatena resničnost (ang. Augmented Reality, AR) je tehnologija prikaza informacij na zaslonu, pri čemer so informacije iz resničnega sveta dopolnjene z navideznimi objekti, ki so ustvarjeni s pomočjo računalniške grafike [3,7].

Obogatena resničnost, označujemo jo tudi z izrazom dopolnjena resničnost, ima štiri poglavitne lastnosti [1,2]:

- Združuje resnični in navidezni svet
- Prikazuje informacije v realnem času
- Omogoča interakcijo z objekti
- Prikazuje informacije v treh dimenzijah

Tridimenzionalni navidezni objekti, ki dopolnjujejo resničnost, so ustvarjeni na podlagi senzorskih podatkov iz okolja, na zaslonu pa so izrisani kot del slike, ki predstavlja resničnost. Tehnologija obogatene resničnosti se uporablja na številnih področjih, na primer v [1]: zdravstvu, arhitekturi, izobraževanju, inženiringu, filmski in zabavni industriji. Zaradi velikega napredka računalniških tehnologij v zadnjem času, na tem mestu moramo omeniti predvsem področji računalniškega vida in interaktivne računalniške grafike [5], se je zelo hitro lahko razvila tudi tehnologija obogatene resničnosti [3]. Z naraščanjem zmogljivosti in razširjenosti uporabe mobilnih naprav, se je navidezna resničnost najbolj uveljavila na področju prenosnih terminalov [4]. Ti namreč združujejo vse potrebne komponente za delovanje aplikacije navidezne resničnosti. Vhodne senzorske naprave (kamera, GPS, žiroskop) omogočajo pridobivanje informacij iz resničnega okolja [4], ki se nato s pomočjo zmogljivih centralno-procesnih enot dopolnijo z navideznimi objekti, vse skupaj pa se na koncu izriše na zaslonu. Za obogateno resničnost na prenosnih napravah uporabljamo izraz mobilna obogatena resničnost (ang. Mobile Augmented Reality, MAR) [1,4].

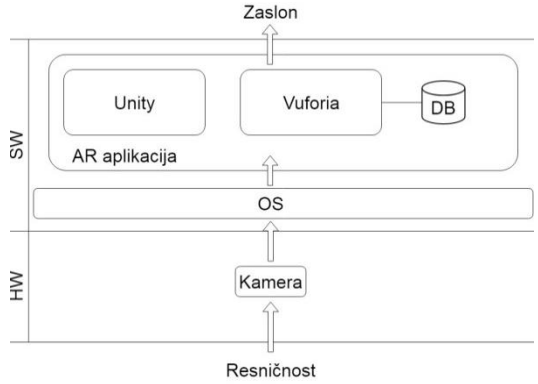
Ena izmed poglavitnih funkcionalnosti aplikacij obogatene resničnosti je prepoznavanje resničnega sveta v okolici, ki se v osnovi deli na dva principa [4]. Prvi je prepoznavanje okolja s pomočjo strojnega učenja, kjer algoritmi na dvodimenzionalni sliki lahko prepoznajo posamezne objekte (npr. človek, drevo, avtomobil) in jih medsebojno združijo po ustreznih kategorijah. Tovrstni princip se uporablja predvsem za prikazovanje splošnih informacij o okolici, saj ta ni natančno določena skozi čas. Drugi princip delovanja pa sloni na prepoznavanju točno določenih vzorcev z lastnimi atributi, ki so predhodno določeni kot tarče za prepoznavanje v resničnem svetu. Takšen princip se uporablja v primeru, ko mora biti okolica natančno prepoznana in obogatena s točno določenimi objekti. Večina aplikacij obogatene resničnosti deluje na principu prepoznavanja slikovnih informacij. S tega vidika ločimo dve metodi [4]:

1. Metoda na podlagi značilnosti statičnih slik
2. Metoda na podlagi značilnosti specifičnih vzorcev

Poznamo tudi t.i. metode brez značilnosti (ang. Markerless Augmented Reality) [2], ki okolice ne prepoznavajo na podlagi slikovnih informacij, temveč se osredotočajo na podatke drugih vhodnih naprav (npr. zaznavanje koordinat v prostoru na podlagi GPS sprejemnika).

3 Vuforia

Vuforia je okolje za implementacijo aplikacij navidezne resničnosti na mobilnih terminalih [1]. Podprtost v operacijskih sistemih iOS in Android ter kompatibilnost z okoljem za razvoj interaktivnih aplikacij Unity, sta njeni glavni prednosti, ki omogočata razvijalcem izdelavo aplikacije za širok spekter mobilnih naprav. Osnovni namen okolja Vuforia je zaznavanje in interpretacija vizualnih informacij iz okolice. Poleg prepoznavanja vzorcev iz dvodimenzionalnih slik, lahko s pomočjo Vuforie zaznavamo kompleksnejše slikovne podobe v treh dimenzijah [1]. Zaznavanje slikovnih informacij v realnem času omogoča tudi sledenje spremembam v okolici, na podlagi česar se nato lahko spreminjajo tudi navidezni objekti na zaslonu.



Slika 1: Osnovna arhitektura sistema obogatene resničnosti na mobilni napravi

Prepoznavanje slikovnih informacij v aplikaciji deluje povsem neodvisno od izrisovanja navideznih objektov [1], ki se vrši v sklopu ločenega okolja za interpretacijo interaktivne računalniške grafike (npr. Unity). S tovrstnim okoljem je Vuforia povezana preko integriranih programskih knjižnic. Funkcionalno torej aplikacijo obogatene resničnosti lahko razdelimo na dva dela:

- prepoznava in sledenje slikovnim informacijam v resničnem okolju (Vuforia)
- izris in interakcija med ustreznimi 3D objekti na podlagi prejetih informacij o resničnem okolju

Informacije, ki so potrebne za pravilno razpoznavo slikovnih vzorcev, Vuforia pridobi iz podatkovne baze, ki je lahko bodisi lokalna (v napravi) ali oddaljena (v oblaku) [1]. Prva za delovanje aplikacije ne potrebuje internetne povezave in deluje hitreje od podatkovne baze v oblaku, ki pa na drugi strani omogoča shranjevanje večjega števila slik in metapodatkov za prepoznavo. Na sliki 1 je prikazana osnovna arhitektura sistema obogatene resničnosti na mobilni napravi, ki je implementirana s pomočjo okolij Vuforia in Unity.
Proces razvoja aplikacije s pomočjo okolja Vuforia lahko razdelimo v štiri poglavitne faze [1]:

- Izdelava podatkovne baze s slikami za prepoznavo
- Definicija konfiguracije posamezne slike v podatkovni bazi
- Uvoz podatkovne baze v okolje za razvoj aplikacije
- Prepoznava slik in prikaz 3D modelov v aplikaciji

Podatkovno bazo v okolju Vuforia preko spletnega vmesnika lahko ustvarimo zelo enostavno. Vsaka slika, ki jo naložimo v podatkovno bazo, se oceni s parametri, ki definirajo, kako kvalitetna je slika za prepoznavo s pomočjo tehnologije strojnega vida. V odvisnosti od kontrastnosti in števila ostrih robov na sliki [2], se določijo značilne točke na sliki (označene s simbolom križca), na podlagi katerih se ustvarijo in shranijo metapodatki o posamezni sliki.



Slika 2: Primer slabo prepoznavne [8] (levo) in dobro prepoznavne slike (desno)

Metapodatki o slikah v podatkovni bazi se med delovanjem aplikacije primerjajo z metapodatki slike, ki jo okolje Vuforia zazna preko kamere. V primeru ustrezne stopnje ujemanja metapodatkov, se v aplikaciji izvrši ustrezna logika za izris ustreznih navideznih objektov.

4 Implementacija aplikacije

Aplikacija, ki smo jo izdelali za operacijski sistem Android, v osnovi sestoji iz dveh delov (Unity aktivnosti): osnovnega menija, preko katerega vstopimo v glavno aktivnost, v kateri je implementirana tehnologija obogatene resničnosti.
Za potrebe prepoznavanja slik smo izdelali podatkovno bazo, ki smo jo v razvojno okolje Unity uvozili v obliki programskega paketa, ob namestitvi aplikacije pa se ta zapiše v lokalno shrambo naprave. V podatkovno bazo smo naložili eno sliko, ki je bila ocenjena kot dobro prepoznavna slika.
3D model robota, ki se prikaže ob prepoznavi slike, smo opremili z informativnimi točkami, katere označujejo posamezne sestavne dele robota. Vse točke so odzivne na dotik, ob pritisku na posamezno točko, pa se na zaslonu prikaže povečan model ustreznega sestavnega dela in z njim povezane informacije.
Na zaslon smo umestili tudi gumb, s katerim uporabnik lahko preklaplja med zaprtostjo/odprtostjo pokrova, ki pokriva jedrni del robota in pod katerim se skrivajo dodatne informativne točke.
Modalnost interakcije, ki smo jo uporabili pri implementaciji uporabniškega vmesnika, je zaznavanje pritiskov na zaslonu občutljivem na dotik. Poleg že omenjenih gumbov in informativnih točk na zaslonu, lahko uporabnik s pomočjo potegov prstov preko interaktivnih elementov, te vrti okoli svoje osi za 360 stopinj. Sprva smo zaznavo dogodkov za vrtenje 3D modelov nameravali uresničiti s pomočjo uporabe že predpripravljene programske knjižnice, vendar smo zaradi nekaterih njenih omejitev, ki bi spremenila interakcijo na nekoliko manj naraven in za uporabnika bolj zahteven način, dogodke za interakcijo z objekti raje napisali sami. Vrtenje objektov, smo namesto z dvema prstoma omogočili zgolj z enim prstom.
Poleg standardnih elementov grafičnega uporabniškega vmesnika, smo vanj umestili tudi nekaj komponent za pomoč in namige uporabniku (npr. simbol za vrtenje pod 3D modelom).
Pri implementaciji zaznavanja slike smo uporabili t.i. način razširjenega sledenja (ang. extended tracking), kjer 3D model ostane izrisan na zaslonu tudi takrat, ko pogled kamere umaknemo stran od prepoznane slike. Slika 3 prikazuje nekaj zaslonskih posnetkov aplikacije.



Slika 3: Zaslonski posnetki izdelane aplikacije

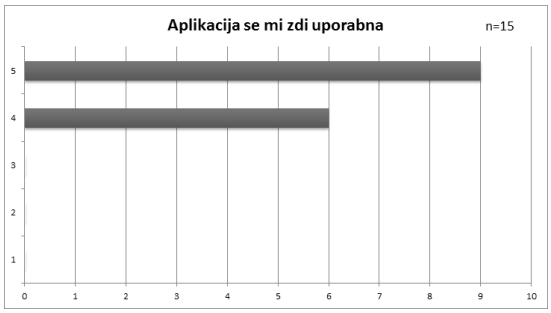
5 Testiranje in rezultati

Aplikacijo smo testirali na mobilni napravi s sledečimi karakteristikami:

Tabela 1: Karakteristike testne naprave

Ime naprave	Samsung Galaxy S8, SM-G950F
Operacijski sistem	Android 8.0.0 Oreo
Delovni pomnilnik	4 GB
Procesor	8-jedrni (2.3GHz Quad + 1.7GHz Quad), 64 bit
Kamera	12,5 MP
Zaslon	5,8″ (2960 * 1440)

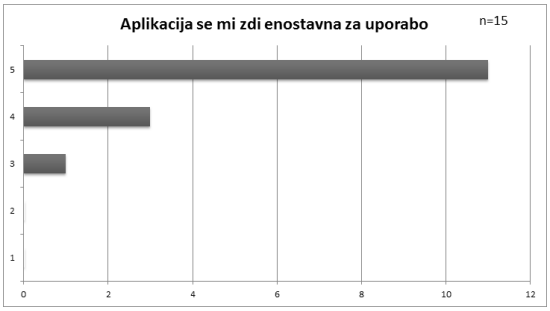
Aplikacijo smo testirali s pomočjo 15 oseb. Vsak uporabnik je imel na voljo nekaj časa, da se spozna z aplikacijo in preizkusi njeno delovanje, nato pa je izpolnil kratek vprašalnik [6]. Na podlagi vprašalnika smo predvsem želeli preveriti ali obogatena resničnost kot način za predstavitev informacij ustreza našemu primeru, hkrati pa smo želeli preveriti tudi, ali je aplikacija uporabniku prijazna, uporabniški vmesnik pa jasen in intuitiven. Vprašalnik, čigar rezultati so prikazani v spodnjih grafih, je bil sestavljen iz petih vprašanj, odgovor na vsakega pa je bilo potrebno podati v obliki številske ocene od 1 (najmanj) do 5 (največ).



Slika 4: Rezultati vprašanja o uporabnosti aplikacije

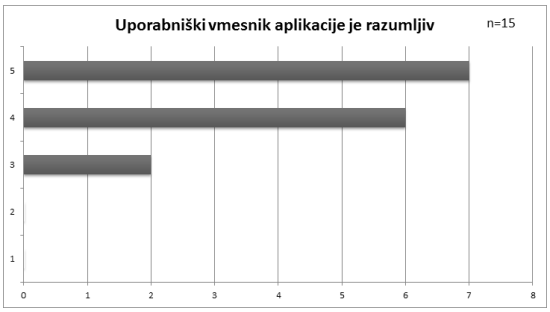
Slika 4 prikazuje rezultate odgovorov na vprašanje o uporabnosti aplikacije (ang. usefulness). Vidimo lahko, da je 9 uporabnikov na to vprašanje odgovorilo z oceno 5, preostalih 6 pa z oceno 4, kar v povprečju prinaša oceno 4.6, na podlagi katere lahko trdimo, da so vsi uporabniki ocenili aplikacijo kot uporabno.

Poleg ocen o uporabnosti aplikacije, so za nas veliko bolj pomembne ocene o uporabljivosti aplikacije (ang. usability), ki s subjektivnega stališča definirajo stopnjo prijaznosti rešitve do uporabnika. Pri sestavljanju ankete smo bili v dilemi glede razlikovanja med zgornjim in spodnjim vprašanjem s strani uporabnikov. Spraševali smo se namreč, ali le-ti dejansko poznajo razliko med pojmoma uporabnost in uporabljivost. S tem namenom smo vprašanji sestavili v relativno razumljivih oblikah, ki že sami po sebi namigujeta na bistveno razliko.



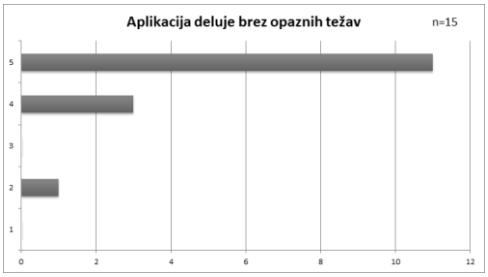
Slika 5: Rezultati vprašanja o enostavnosti uporabe aplikacije

Na sliki 5 so prikazani rezultati odgovorov na vprašanje o enostavnosti uporabe same aplikacije. Zopet je večina uporabnikov (11) podala najvišjo oceno, medtem ko so trije uporabniki izbrali oceno 4, eden pa oceno tri. Povprečna ocena enostavnosti uporabe tako znaša 4,7. Tudi tukaj lahko na podlagi visokega povprečja označimo aplikacijo kot enostavno za uporabo.



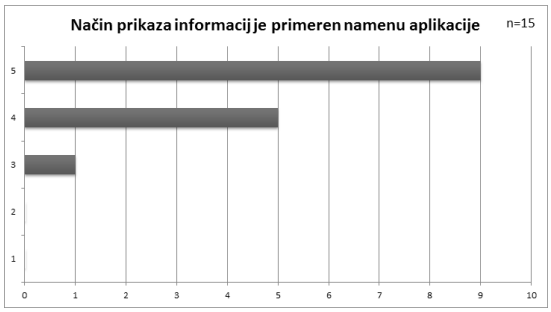
Slika 6: Rezultati vprašanja o razumljivosti uporabniškega vmesnika

Podoba grafičnega uporabniškega vmesnika aplikacije predstavlja velik faktor na področju interakcije aplikacijskih rešitev s končnimi uporabniki. 7 uporabnikov je razumljivosti uporabniškega vmesnika namenila oceno 5, 6 uporabnikov oceno 4 in 2 uporabnika oceno 3. Tukaj lahko opazimo nekoliko nižje skupno povprečje (4.3) in večji standardni odklon kot pri ostalih vprašanjih. Med testiranjem smo opazili, da nekaj uporabnikov ni takoj opazilo interaktivnih točk na robotu, saj so le te precej majhnih dimenzij, kontrast med njimi in 3D modelom pa morda nekoliko premajhen. Ker smo se pri razvoju aplikacije bolj osredotočili na samo implementacijo tehnologije navidezne resničnosti, smo načrtovanju uporabniškega vmesnika namenili nekoliko manj časa, kar se kaže tudi v odgovorih na zgornje vprašanje.



Slika 7: Rezultati vprašanja o performančnosti aplikacije

Tudi tekoče delovanje rešitve vpliva na uporabnikovo subjektivno mnenje o aplikaciji. 11 oseb med uporabo aplikacije ni opazilo nobenih težav, 3 osebe skoraj nič, ena oseba pa je zaznala nekoliko več težav, zato je na vprašanje odgovorila z oceno 2. Morda se razlog za to skriva v nekoliko manj intuitivnem uporabniškem vmesniku, kar smo izpostavili že v prejšnjem odstavku.



Slika 8: Rezultati vprašanja o primernosti načina prikaza informacij

V zadnjem vprašanju smo uporabnike povprašali o primernosti načina prikaza informacij glede na namen aplikacije. Želeli smo preveriti, ali je tehnologija obogatene resničnosti ustrezna za predstavitev oprijemljivih produktov v navidezni obliki. Na podlagi povprečne ocene 4.5 lahko trdimo, da so si uporabniki s pomočjo aplikacije uspeli ustvariti dobro predstavo o izdelku. Iz zgornjih grafov lahko razberemo, da princip obogatene resničnosti ustreza osnovnemu namenu naše aplikacije – predstavitev fizičnega produkta v navidezni obliki. S tem smo dosegli cilje, ki smo si jih zadali v prvi fazi načrtovanja aplikacije.

S stališča delovanja aplikacije večjih težav sami nismo opazili. Razpoznava vzorca iz slike se je v različnih svetlobnih pogojih (sončna svetloba, umetna luč, senca) izvršila podobno hitro – praktično skoraj takoj, ko smo kamero usmerili proti sliki. 3D objekt se je na zaslonu izrisal brez opaznih zamud in morebitnih popačitev in nepravilnosti. Ob premikanju naprave po prostoru, je izrisovanje objekta v realnem času tekoče sledilo trenutni poziciji naprave (približevanje/oddaljevanje, kroženje okoli navideznega objekta, premikanje podlage), kar predstavlja velik doprinos k zmanjševanju razlik med resničnim in navideznim svetom. Edina slabost, ki smo jo opazili med uporabo aplikacije, je fizično segrevanje mobilne naprave, saj je frekvenca izrisovanja objektov v realnem času precej visoka, posledica tega pa je visoka poraba strojnih zmogljivosti naprave.

6 Sklep

Kot smo lahko videli, smo z našim delom izpolnili cilje, ki smo si jih zadali na začetku. Kljub uspešnim rezultatom testiranja in tekočemu delovanju aplikacije, bi le-to lahko še izboljšali. V prvi vrsti je tu zagotovo optimizacija porabe strojne opreme, ki bi jo lahko zmanjšali z drugačnim načinom izrisovanja 3D objekta na zaslonu. Sedaj se v grafičnem cevovodu, ki skrbi za pravilno prikazovanje objekta, procesirajo vse njegove komponente, tudi tiste, ki trenutno niso vidne na zaslonu. Če bi izrisovanje omejili zgolj na vidne komponente v ospredju, bi zagotovo nekoliko razbremenili obremenitev centralno procesnih enot v napravi.

Tudi z vidika tehnologije obogatene resničnosti, bi aplikacijo lahko še nadgradili. V prvi fazi bi lahko implementirali hkratno prepoznavanje več slik, nadaljevali pa bi s prepoznavanjem tridimenzionalnih objektov v okolici.

Nekaj izboljšav bi lahko naredili tudi na intuitivnosti uporabniškega vmesnika aplikacije.

Literatura

[1] C. Xiao, Z. Lifeng, »Implementation of Mobile Augmented Reality Based on Vuforia and Rawajali,« http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6933713&isnumber=6933501

[2] D. Adrianto, M. Hidajat, V. Yesmaya, »Augmented Reality Using Vuforia for Marketing Residence,« http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8052642&isnumber=8052634

[3] D.H.Marjury, B. C. Karen, M. M. Diana in L. F. Gabriel, »Offline Mobile Application for Places Identification with Augmented Reality,« http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7962546&isnumber=7962495

[4] F. Peng, J. Zhai, »A Mobile Augumented Reality System for Exhibition Hall Based on Vuforia,« http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7984714&isnumber=7984442

[5] M. Di Capua, G. Costagliola, M. De Rosa in V. Fuccella, »Rapid prototyping of mobile applications for augumented reality interactions,« http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6070416&isnumber=6070364

[6] G. Sato, G. Hirakawa in Y. Shibata, »Push Typed Tourist Information System Based on Beacon and Augumented Reality Technologies,« http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7920923&isnumber=7920866

[7] Kaj je obogatena resničnost in kaj nam prinaša, DNE, http://dne.enaa.com/Inovacije/kaj-je-obogatena-resnicnost-in-kaj-nam-prinasa.html

[8] »Natural Features and Image Ratings,« vuforia, [Online], Dostopno: https://library.vuforia.com/articles/Solution/Optimizing-Target-Detection-and-Tracking-Stability.htm

Primerjava načrtovalskih vzorcev programske kode v mobilnih aplikacijah

Vid Ribič, Matevž Pogačnik

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: ribvid@gmail.com

Comparison of design patterns used in the development of mobile applications

Abstract. *In this paper, we examine some of the more common design patterns used in the development of mobile applications for the iOS. With their help, the application is divided into multiple components, which has certain key advantages. It is easier to reuse user interface (UI) elements and we thus follow the don't repeat yourself (DRY) concept, making the code optimized, more understandable and easier to test. The use of a suitable design pattern quickens development and simplifies bug fixing.*

We also describe the model-view-controller (MVC) architecture, which seems an excellent starting point for most applications. In recent years, however, certain qualms have arisen regarding its suitability for more demanding programs as the use of MVC requires caution in order to avoid the problem of massive view controller.

With the MVP concept, the controller is moved to a part of the view, while man-in-the-middle becomes a new element – the presenter.

Also described is the MVVM pattern, where the view owns the view model, as well as VIPER, which divides individual tasks of the application into five minor components.

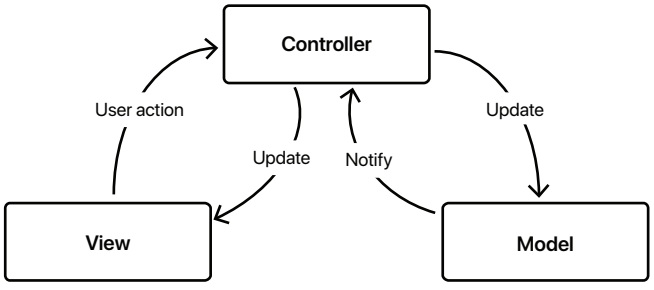
1 Uvod

Pri programiranju že od nekdaj pomembno vlogo igrajo načrtovalski vzorci (angl. design patterns). Z njihovo pomočjo razbijemo programsko kodo na več različnih komponent, vsako s svojo zadolžitvijo. To prinaša nekaj pomembnih prednosti. Modularno napisano kodo je lažje ponovno uporabiti in tako slediti konceptu DRY (angl. don't repeat yourself), s čimer zmanjšamo količino napisane kode in posledično verjetnost napak v njenem delovanju. Pomembna prednost je enostavnejše testiranje, ker ga lahko izvajamo nad vsako komponento posebej. Pri monolitno napisani kodi, ki opravlja več nalog hkrati, je testiranje veliko večji, če sploh rešljiv, izziv. Možnost enostavnega testiranja olajšuje spreminjanje obstoječih delov kode in dodajanje novih funkcionalnosti, prav tako kot boljša preglednost, organiziranost in jasnost, ki jo dosežemo z razbitjem kode na več smiselnih enot. Uporaba arhitekturnih vzorcev rezultira v hitrejšem razvoju, predvsem pa v manj hroščati in bolje optimizirani kodi.

Nič drugače ni pri razvoju programske opreme za mobilne naprave. Sprva so bile mobilne aplikacije enostavne in primitivne, zato so se večinoma uporabljali samo najosnovnejši arhitekturni koncepti, kot je model-pogled-kontroler oziroma MVC (angl. model-view-controller). S kompleksnejšimi in zahtevnejšimi programi so se pojavile težave tega pristopa in začelo se je razmišljati o vpeljavi novih arhitekturnih zasnov. V razvijalskih krogih se zadnja leta čedalje pogosteje zastavlja vprašanje, ali je MVC danes še vedno najprimernejši načrtovalski vzorec ali pa je za sodobno mobilno aplikacijo preveč enostaven, zato bomo v tem članku na kratko predstavili nekaj arhitekturnih idej, primernih za razvoj mobilnih aplikacij za platformo iOS.

2 MVC

Najpreprostejši, najpogostejši in tudi eden izmed najstarejših načrtovalskih vzorcev je model-pogled-kontroler ali MVC. Kratico je leta 1979 prvič uporabil Trygve Reenskaug za programe napisane v Smalltalku [1]. Applova različica, ki jo predstavljamo v nadaljevanju, je prilagojena posebnostim mobilnih aplikacij.



Slika 1: Pri idejni zasnovi MVC je most med pogledom in modelom kontroler.

Arhitektura je trodelna. Na eni strani je model, ki je abstrakten opis vsebine naše aplikacije. Model je popolnoma neodvisen od aplikacijskega ogrodja (angl. framework) [1]. Na drugi strani imamo pogled (angl. view), ki predstavlja vidni del aplikacije, uporabniški vmesnik. Pogled in model nista nikoli v neposredni povezavi, s čimer preprečujemo njuno soodvisnost. To je pomembno, saj lahko le tako poglede ponovno uporabimo tudi z drugimi modeli, kar je pri razvoju mobilnih aplikacij dokaj pogosto. Most

med modelom in pogledom je kontroler (angl. controller). Ta dobi informacijo o spremembi modela in ustrezno posodobi pogled. Ter obratno: ko uporabnik interaktira z aplikacijo in izvede določeno akcijo, kot je pritisk na gumb, lahko kontroler posodobi model, če je to seveda potrebno [2].

Po drugi strani pa to pomeni, da ima kontroler veliko zadolžitev in prav to je ključna težava ter kritika te arhitekturne zasnove. Pri kompleksnih aplikacijah se namreč hitro zgodi, da postane kontroler zelo obsežen, čemur se, kot zapisano v uvodu, skušamo izogniti. Pogled v dokumentacijo razkrije, kakšne zadolžitve lahko dodelimo kontrolerju (izpeljemo ga iz razreda UIViewController) [3]. Med njimi so skrb za pravilen prikaz pogleda, nadzor nad njegovim življenskim ciklom, spremljanje morebitnih rotacij zaslona, zaznava in ustrezna obdelava uporabnikovih akcij, lahko se povezuje s strežnikom in obdeluje odgovore, ki mu jih ta vrne, posluša obvestila (angl. notifications) in tako dalje. Seveda vseh funkcij ni potrebno implementirati, če niso potrebne, kljub temu pa se moramo zavedati, da se lahko hitro ujamemo v past preobsežne kode. Tisti, ki trdijo, da gre za preživet koncept, prav zaradi preobsežnosti kontrolerja MVC ironično pojasnujejo kot "massive view controller". Delno rešitev te težave predstavlja koncept kontrolerjev otrokov (angl. child controllers), na katere lahko prenesemo kakšno izmed nalog in jih nato po potrebi vključujemo v druge kontrolerje. Primer take uporabe bi bila ločena kontrolerja za obravnavo nalaganja vsebine in prikaz napake, ki ju nato lahko vključimo v vse tiste kontrolerje, ki obdelujejo strežniške odgovore [4]. Z uporabo kontrolerjev otrokov se lahko izognemo nevarnosti masivnega kontrolerja, obenem pa jih lahko uporabimo na več različnih mestih in s tem sledimo ideji DRY.

Druga težava kontrolerja, ki je povezan s pogledom in modelom, je oteženo testiranje poslovne logike (angl. business logic). Testiranje sicer je mogoče, vendar zahteva nekaj več dela in truda.

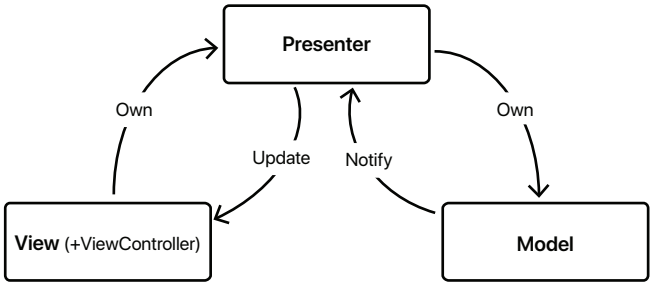
Kljub temu MVC še vedno ostaja pogosta arhitekturna izbira. Prvi razlog je njegova enostavnost – MVC-ja se je najlažje in najhitreje naučiti, hkrati pa v aplikacijo ne vnaša preveliko nepotrebne kode (angl. overheada). MVC je arhitektura, ki jo priporoča in v svojih aplikacijah uporablja Apple. Marsikdo, ki zagovarja uporabo MVC-ja, je prav zaradi tega prepričan, da je aplikacijsko ogrodje napisano z mislijo na to zasnovo (ali vsaj njene neposredne izpeljanke) in so vsi elementi, potrebni za razvoj, že del obstoječih knjižnic.

Četudi se nam zdijo drugi načrtovalski vzorci primernejši od MVC-ja, pa ta velja za osnovo večini ostalih idej. Marsikakšna izmed njih se pravzaprav zdi zelo podobna MVC-ju, zaradi česar se nemalokdo sprašuje, ali ne gre le za drugačen način implementacije istega vzorca in drugačno ime (MV-karkoli, angl. MV-whatever). Masiven kontroler za zagovornike MVC-ja ni odraz slabosti tega načrtovalskega vzorca, pač pa programerja, ki ga ne zna pravilno implementirati [5].

3 MVP

Eden izmed vzorcev, izpeljan iz MVC-ja, je model-pogled-prikazovalnik ali MVP (angl. model-view-presenter). Leta 1996 ga je zasnoval Mike Potel [6]. Ker vsevedni kontroler vodi v obsežno kodo in težavno testiranje, se pri tej zasnovi pomakne v del pogleda; gledano z vidika razredov lahko rečemo, da je poleg razreda UIView sedaj del pogleda tudi UIViewController. Ta sicer še vedno skrbi za UIView, vendar nima več neposredne povezave z modelom, s čimer se izognemo številnim težavam. Prikazovalnik zasede vlogo moža v sredini med modelom in pogledom. Pogled delegira uporabnikove interakcije prikazovalniku, ki vsebuje logiko za njihovo obravnavo. Hkrati je tudi tisti, ki prevzame komunikacijo z modelom, podatke, ki jih od njega prejema, ustrezno pripravi in nato pošlje naprej UIViewControllerju. Prikazovalnik je prav tako kot model neodvisen od aplikacijskega ogrodja UIKit [7].

UIViewController na ta način razbremenimo številnih nalog, hkrati pa tako kot UIView postane nevedni člen arhitekture. S tem ne dosežemo le manj obsežne kode, temveč tudi olajšujemo testiranje, saj lahko sedaj poslovno logiko aplikacije stestiramo neodvisno od UIViewControllerja. Po drugi strani pa MVP prinese nekaj dodatnega dela, saj je potrebno vpeljati nov element, torej prikazovalnik, in ga povezati z ostalimi členi arhitekture.

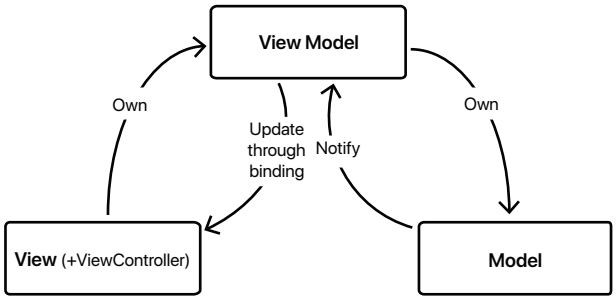


Slika 2: Pri MVP se kontroler pomakne v del pogleda, mož v sredini pa postane prikazovalnik.

4 MVVM

Pri načrtovalski zasnovi MVP ima pogled referenco na prikazovalnik in obratno. Če želimo prikazovalnik razbremeniti povezave s pogledom, lahko uporabimo vzorec model-pogled-pogled model ali MVVM (angl. model-view-view model). Izvorno ga je predstavil Microsoft leta 2005, da bi poenostavil dogodkovno usmerjeno programiranje (angl. event-driven programming) [2].

Še vedno ostaja ideja, da pogled predstavlja vidni del aplikacije in nima možnosti neposredne komunikacije z modelom. Ključna razlika pa je sledeča: pogled si lasti nov nivo, poimenovan pogled model, ta pa si lasti model. Pogled model je zadolžen za hrambo podatkov za pogled, ki si ga lasti. Za primer vzemimo aplikacijo za predvajanje glasbe. Pogled bo narisal seznam skladb, pogled model pa bo hranil informacije o njih (naslov, izvajalca, žanr, dolžino, ...).



Slika 3: Med pogledom in pogledom model vzpostavimo povezavo (angl. binding). Ker iOS nima nativnega pristopa k temu, se veliko razvijalcev odloči za vpeljavo reaktivnosti.

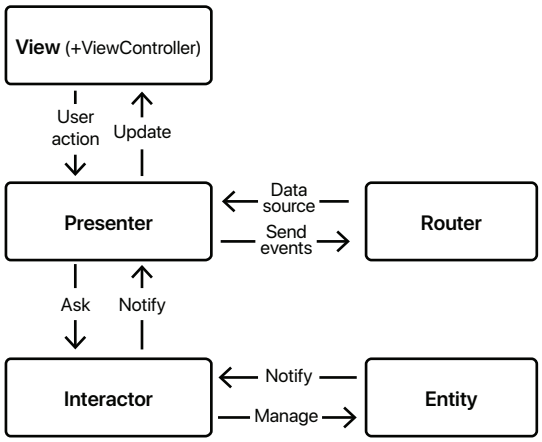
Tok aplikacije je pri MVVM nekoliko drugačen. Postavi se vprašanje, kateri nivo je zadolžen za pridobitev podatkov s strežnika ali baze? Povedano skozi gledišče zgornjega primera: kateri člen arhitekture bo s strežnika pridobil seznam skladb? Pri MVC in MVP sta za to zadolžena kontroler oziroma prikazovalnik, ki nato uporabita model in posodobita pogled. Pri MVVM pa za povezavo s strežnikom oziroma bazo skrbi pogled. Ta pridobi podatke in jih posreduje nivoju pogled model, ki jih sprocesira in pripravi v ustrezni obliki za uporabniški vmesnik ter vrne nazaj pogledu, s katerim je povezan (angl. binding). Težava je, da pri iOS ni pravega in nativnega pristopa za tovrstno povezovanje. Obstajajo sicer mehanizmi, kot so KVO (key-value observing), delegati ali notifikacije, vendar niso tako učinkoviti kot povezovalni mehanizmi pri drugih programskih jezikih. Osnovno aplikacijsko ogrodje ni bilo napisano z mislijo na to načrtovalsko zasnovo. Razvijalci se zato koncepta MVVM najpogosteje lotevajo z reaktivnim pristopom, čeprav ta seveda ni nujen. MVVM še vedno skrbi za razdelitev nalog med različne nivoje, reaktivnost pa za povezovanje (binding).

Čeprav sta pogled in pogled model med seboj tesno povezana, pa je pogled še vedno ločen od modela, s čimer se ohranja osnovna ideja MVC-ja. Hkrati je pri MVVM poslovna logika ločena od kontrolerja in s tem ostaja enostavnost testiranja, ki smo jo vpeljali z MVP-jem. Po drugi strani pa z uvajanjem MVVM-ja prihaja do novih izzivov. Prvo vprašanje, ki bega razvijalce, je razdelitev nalog. Kar nekoliko nenavadno se zdi, da je pogled tisti, ki je zadolžen za povezavo s strežnikom ali bazo, vendar tako narekuje izvorna ideja MVVM-ja. Morda bi bila ta naloga primernejša za pogled model? Primer uporabne vrednosti tega, da je pogled zadolžen za pridobivanje podatkov, je to, da sam ve, kdaj so podatki uspešno preneseni in kdaj je prišlo do napake. Če bi to nalogo preselili na nivo pogled model, bi to pomenilo, da moramo dodati še en nivo abstrakcije, za primer, ko podatkov ni in pogled model vrača pogledu napako.

Druga težava MVVM-ja pri iOS-u je manjko nativnega pristopa k povezovanju. Reaktivnost se sliši kot dobra rešitev, vendar je kompleksna, zahteva več učenja in posledično bolj nagiba našo kodo k možnosti napak. Zavedajoč se prednosti in slabosti tega koncepta, lahko sklenemo, da je MVVM zanimiva in premisleka vredna rešitev, ki pa zaradi svoje zahtevnosti zagotovo ni primerna za vse aplikacije.

5 VIPER

Še korak naprej od MVVM k večji modularnosti je VIPER (view-interactor-presenter-entity-router), ki aplikacijo razdeli na pet delov.



Slika 4: Viper razdeli aplikacijo na pet delov. Čeprav v teoriji dobro razdeli zadolžitve in olajša testiranje, se v praksi izkaže, da je tovrstna abstrakcija nemalokrat prevelika.

Pogled ostane zadolžen za prikaz uporabniškega vmesnika, uporabnikove akcije pa pošlje prikazovalniku. Ta je odgovoren za njihovo obdelavo in je popolnoma neodvisen od uporabniškega vmesnika. Tretji člen je interaktor. Vsebuje vso poslovno logiko, hkrati ima v lasti in upravljanju entiteto (angl. entity), ki je četrti nivo te zasnove, namenjen definiranju strukture za shranjevanje podatkov. Zadnja komponenta je usmerjevalnik (angl. router), ki skrbi za navigacijo na podlagi akcij, prejetih iz prikazovalnika [8].

VIPER razdeli zadolžitve aplikacije na več členov, s čimer še dodatno poenostavi testiranje. Po drugi strani pa uvedba toliko novih razredov aplikacijo zelo razdrobi in podaljšuje učenje ter razumevanje zgradbe in arhitekture programa. Zavedati se moramo tudi velike razlike med osnovno zasnovo MVC in VIPER. Se je res smiselno toliko oddaljiti od koncepta, ki ga zagovarja tisti, ki skrbi za aplikacijsko ogrodje?

6 Sklep

Poleg predstavljenih obstaja še mnogo drugih načrtovalskih zasnov (Redux, Cake, MVC+VS, MAVB, TEA, ...). Postavlja se vprašanje, katera je najprimernejša? Bistveno je zavedanje, da je namen vsake arhitekture programerju olajšati delo, pohitriti programiranje, odpravljanje napak in dodajanje novih funkcionalnosti, poenostaviti testiranje ter zmanjšati obseg kode. Če VIPER ponuja nadrobno razdelitev nalog med posameznimi deli, kar ima nedvomno svoje prednosti, pa je bistveno bolj kompliciran od MVC-ja. Obstaja pomislek, da se arhitekture, kot je VIPER, na papirju zdijo dobra rešitev, v praksi pa prinašajo preveč abstrakcije in posledično zmede. Ključna težava

MVC-ja je nevarnost preobsežnega kontrolerja, vendar to ne sme biti razlog, da se tej idejni zasnovi prehitro odpovemo. Obseg kontrolerja lahko zmanjšamo. Data source in delegate za collection view ali table view, ki sta pogosta elementa uporabniškega vmesnika, lahko ločimo od UIViewControllerja. Iz kontrolerja lahko vselej prestavimo mrežne klice in obdelavo datotek JSON, s čimer med drugim poenostavimo testiranje. Omenili smo tudi kontrolerje otroke, ki jih vključujemo, kjer jih potrebujemo. Težavo pretoka podatkov med posameznimi členi aplikacije lahko učinkovito rešimo z uporabo koordinatorjev. Upoštevajoč možnosti za razbremenitev kontrolerja in izogib njegove preobsežnosti, ugotovimo, da ima vendarle MVC še vedno široko uporabno vrednost. Ker je MVC najenostavnejši in Applov izbrani načrtovalski vzorec za platformo iOS, je najbolje, da se mu odpovemo šele takrat, ko predstavljajo njegove omejitve nepremostljivo oviro.

Literatura

[1] C. Eidhof, M. Gallagher, in F. Kugler, App Architecture, iOS Application Design Patterns in Swift. CreateSpace Independent Publishing Platform, 2018.

[2] A. Harbade, iOS design patterns, iOS design patterns — Part 1 (MVC, MVP, MVVM). Dostopno na: https://medium.com/swlh/ios-design-patterns-a9bd07818129.

[3] Apple, Documentation, UIKit, View Controllers, UIViewController. Dostopno na: https://developer.apple.com/documentation/uikit/uiviewcontroller.

[4] J. Sundell, Using child view controllers as plugins in Swift. Dostopno na: https://www.swiftbysundell.com/posts/using-child-view-controllers-as-plugins-in-swift.

[5] A. Vacić, Much ado about iOS app architecture. Dostopno na: http://aplus.rs/2017/much-ado-about-ios-app-architecture/.

[6] M. Potel, MVP: Model-View-Presenter The Taligent Programming Model for C++ and Java. Taligent, 1996.

[7] I. Agha, A dumb UI is a good UI: Using MVP in iOS with swift. Dostopno na: http://iyadagha.com/using-mvp-ios-swift/.

[8] A. Harbade, iOS design patterns — Part 2 (VIPER). Dostopno na: https://medium.com/@anup.harbade.iosdev/ios-design-patterns-part-2-viper-fa3522b22b6b.

Razvoj platforme za izmenjavo multimedijskih vsebin z uporabo zasebne kriptovalute

Uroš Zoretič, Grega Jakus

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, Ljubljana, Slovenija
E-pošta: uros.zoretic123@gmail.com, grega.jakus@fe.uni-lj.si

Abstract. *Blockchain is a key element for implementation of cryptocurrencies. It ensures the integrity, authenticity, irreversibility, privacy and timestamp of the recorded data without the need for a third-party trusted authority. Because of their potential, blockchains have found their ways also into the fields of decentralized autonomous organizations, supply chains and the Internet of Things. Due to the increasing popularity of blockchains, developers also benefit from ever more ready-available functionalities in form of web services and programming libraries, which enables developers, among other things, to use the existing cryptocurrencies in their applications, ensure time traceability of their data, or create their own blockchains and smart contracts.*

In order to evaluate the effectiveness of using libraries for cryptocurrency creation, we developed a content sharing application where the user gets an amount of digital money in exchange for providing digital content to the community. The cryptocurrency was created using the Naivecoin library and can be interacted through a web service. The platform can be used through web or mobile application. Based on our experience the use of blockchain libraries speeds up the development. However, in order to obtain reliable applications thorough testing is needed due to frequent bugs in libraries.

1 Uvod

Tehnologija veriženja blokov je bila razvita za izvedbo porazdeljene javne glavne knjige transakcij pri kriptovaluti Bitcoin. Glavna knjiga je ključni element kriptovalut, saj vsebuje seznam vseh izvedenih prenosov denarja s podatki o plačniku, prejemniku in znesku posameznega prenosa ter zaporedju izvršenih prenosov, iz katerega je mogoče sklepati o stanju na posameznem računu.

Blokovna veriga je sestavljena iz blokov, v katere zapisujemo podatke v obliki transakcije, ki ima praktično enako vlogo kot v tradicionalnih podatkovnih zbirkah, kjer služi spreminjanju zapisov. V transakcijo lahko zapišemo različne podatke, odvisno od namena uporabe verige. Pri kriptovalutah so to virtualna denarna sredstva, lahko pa transakcija vsebuje tudi izvedljivo programsko kodo v obliki pametnih pogodb (ang. smart contracts).

Poleg podatkov v transakcijah vsebuje blok tudi časovno oznako njegovega nastanka, javni ključ lastnika bloka, ki ga je potrdil, dokaz dela (ang. proof of work), ki služi kot orodje za doseganje soglasja o tem, kateri vozel v omrežju lahko v verigo doda naslednji blok, in izvleček (ang. hash), ki zagotavlja nespremenljivost podatkov v bloku. Ko so posamezni bloki ustvarjeni in potrjeni, se zapišejo v blokovno verigo, v kateri vsak blok vsebuje kazalec na izvleček prejšnjega bloka.

Pomembna lastnost blokovne verige je nespremenljivost v njej zapisanih podatkov, kar je pri kriptovalutah ključno za odpravo problema večkratne uporabe istega denarja. Ker lahko poleg tega nad blokovnimi verigami uporabimo tudi mehanizme, ki zagotavljajo verodostojnost, avtentičnost, neovrgljivost, časovno opredeljenost in tajnost podatkov, odpade potreba po zaupanja vrednem posredniku pri transakcijah, saj za navedene lastnosti skrbijo kar sodelujoči člani v omrežju. Bitcoin in druge kriptovalute so se zato lahko uveljavile kot alternativa klasični izmenjavi denarja s posredovanjem zaupanja vrednih avtoritet (bank).

Zaradi omenjenih lastnosti blokovnih verig, predvsem pa ker ne potrebuje zaupanja vrednega posrednika, se tehnologija širi tudi na druga področja, kot so kibernetska varnost [1], decentralizirani splet [2] in shranjevanje podatkov v porazdeljenem oblaku [3]. S pomočjo pametnih pogodb [4] je mogoče ustvariti tudi decentralizirane avtonomne organizacije [5] in upravljati oskrbovalne verige podjetij [6]. Pilotni projekti potekajo tudi v zdravstvu [7] in internetu stvari [8].

Naraščajoča priljubljenost in razširjenost tehnologije veriženja blokov pa ima še eno pomembno prednost, to je razširjenost že pripravljene funkcionalnosti, ki razvijalcem olajša in pohitri razvoj aplikacij, ki uporabljajo to tehnologijo. Namen prispevka je podati pregled razpoložljive funkcionalnosti in predlagati način uporabe enega izmed ogrodij pri izgradnji aplikacije, ki uporablja blokovno verigo za izvedbo zasebne valute.

2 Razpoložljiva ogrodja in storitve

Obstoječe funkcionalnosti zasledimo v obliki programskih knjižnic in spletnih storitev, ki omogočajo izgradnjo lastnih blokovnih verig ali pa integracijo obstoječih v končne aplikacije. Na ta način se razvijalcu ni potrebno ukvarjati z nizkonivojsko izvedbo blokovne verige, ampak pripravljeno ogrodje le prilagodi svojim potrebam.

2.1 Spletne storitve

Spletna storitev je storitev v skupni rabi, do katere odjemalci dostopajo z uporabo komunikacijskih protokolov, najpogosteje protokola HTTP (ang. HyperText Transfer Protocol). Storitev se izvaja ločeno od aplikacije, ki jo uporablja, pogosto na oddaljenem strežniku. Prednost uporabe storitve v primerjavi s programsko knjižnico je možnost sočasne interakcije z več odjemalci, ki storitev uporabljajo (vključno z njenimi podatki).

Uporaba kriptovalut v aplikacijah. Spletne storitve se najpogosteje uporabljajo za vključitev plačil s kriptovalutami v različne aplikacije. Med bolj uporabljenimi

tovrstnimi storitvami je Blockchain.info [9], ki omogoča plačila s kriptovaluto Bitcoin. Podobne storitve so še Coinbase [10], Gem [11] in BitGo [12].

Časovna sledljivost. Storitev Tierion [13] omogoča časovno sledljivost podatkov. Lahko se uporabi za shranjevanje zgodovine poslovnih procesov, časovno žigosanje in podpisovanje dokumentov ter zagotavljanje integritete podatkov, pridobljenih z naprav interneta stvari. Za zagotavljanje verodostojnosti podatkov uporablja protokol Chainpoint [14], ki je postal *de-facto* standard za časovno dokazljivost nastanka podatka.

Verižna omrežja v oblaku. Amazon Web Services (AWS) Blockchain [15] nudi oblačna ogrodja, ki temeljijo na priljubljenih verigah Ethereum ali Hyperledger, za izdelavo lastnih blokovno-verižnih omrežjih, ki se izvajajo znotraj okolja AWS. AWS Blockchain ponuja predloge, ki jih lahko razvijalec prilagodi in s tem ustvari lastne porazdeljene aplikacije.

2.2 Programske knjižnice

Programska knjižnica je deljena programska koda, ki se skupaj z izvorno kodo aplikacije, v kateri jo uporabimo, prevede v izvršno kodo. Razvijalec lahko do funkcionalnosti programske knjižnice dostopa s klici funkcij njenega aplikacijskega programskega vmesnika. Programska koda knjižnice se izvaja v istem izvajalnem okolju (računalniku) kot aplikacija, ki knjižnico uporablja.

Dostop do storitev. Programske knjižnice pogosto omogočajo le lažjo izvedbo klicev storitev, v okviru katerih je uporabljena funkcionalnost dejansko udejanjena. Veliko takih knjižic omogoča dostop do iste storitve v različnih programskih jezikih. Dostop do storitve Blockchain.info tako omogočajo knjižnice za python, PHP [16] in javascript [17]. Za interakcijo s storitvama Blockchyper in BitGo pa sta na voljo knjižnici za PHP [18, 19].

Zaradi njihove popularnosti se veliko knjižnic osredotoča na izmenjavo kriptovalut, še posebej Bitcoina. Za uporabo te kriptovalute sta tako na voljo knjižnici BitcoinJ [20] v javi in Bitcore-lib [21] v javascriptu.

Ustvarjanje pametnih pogodb. Naprednejše knjižnice omogočajo ustvarjanje pametnih pogodb, pri katerih namesto podatkov v blokovno verigo shranjujemo izvedljivo programsko kodo, ki olajša izvrševanja sporazuma med pogodbenimi strankami. Ukazi v programski kodi se izvedejo samodejno, ko so izpolnjeni določeni pogoji. S pametnimi pogodbami je tako mogoče ustvariti decentralizirane avtonomne organizacije ali pa upravljati oskrbovalno verigo podjetij.

Trenutno najbolj razširjena blokovna veriga na tem področju je Ethereum s programskim jezikom Solidity [22]. Programska knjižnica Web3j poenostavi pisanje pametnih pogodb za Ethereum [23] v javi in javascriptu. Pametne pogodbe je mogoče napisati tudi za blokovno verigo Hyperledger [24] v javi, pythonu in javascriptu. V primerjavi s knjižnicami za izvedbo transakcij v kriptovalutah je izbire pri ustvarjanju pametnih pogodb manj, saj zahtevajo več napora pri razvoju.

Zasledimo tudi bolj specializirane knjižnice, na primer takšne, ki izvajajo semantično in skladenjsko analizo pametnih pogodb. Primera sta pythonska knjižnica Mythril [25] in njena različica za javascript Solium [26].

Knjižnica Embark [27] po drugi strani služi kot orodje za razvoj porazdeljenih aplikacij. Temelji na verigi Ethereum in decentraliziranemu shranjevanju dokumentov IPFS (ang. InterPlanetary File System) [28].

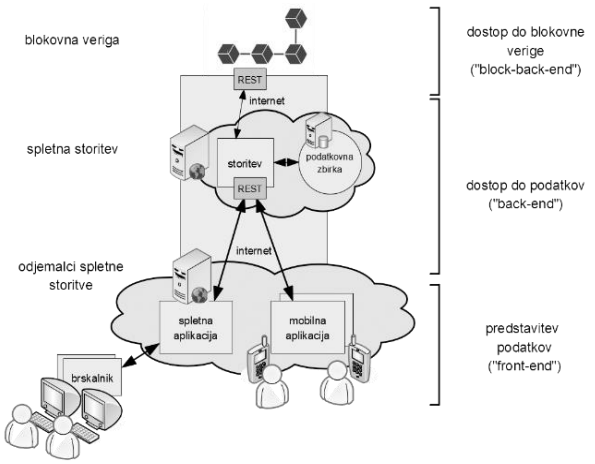
Knjižnice za izdelavo lastnih blokovnih verig. Večina knjižnic za izvedbo lastne blokovne verige je odprtokodnih in so na voljo v mnogih programskih jezikih. Mnogo tovrstnih programskih knjižnic je podanih v obliki spletnega tečaja. V programskem jeziku python sta tako na voljo "A simple Blockchain in Python" [29] in "Jbc" [30]. V javi je na voljo "NoobChain" [31], v javascriptu pa "BrewChain" [32] in "Naivecoin" [33]. Slednji služi tudi kot osnova za razvoj platforme z lastno kriptovaluto, predstavljene v naslednjem poglavju.

3 Platforma za izmenjavo vsebin z uporabo zasebne kriptovalute

Platforma MMShare je namenjena izmenjavi vsebin, pri čemer spodbuja uporabnike k sodelovanju. Naložene vsebine so zato dostopne samo tistim uporabnikom, ki so v skupnost že prispevali lastne vsebine. Uravnoteženost izmenjave dosežemo z uporabo zasebne kriptovalute, ki smo jo poimenovali MMCoin.

Ob registraciji dobi vsak uporabnik javni ključ, ki identificira njegov račun in se hrani v zalednem delu v podatkovni zbirki. Zasebni ključ, ki potrdi lastništvo prenesenih sredstev, se iz gesla vsakič sproti izračuna**.**

Uporabnik dobi ob registraciji pet tisoč kovancev. Nadaljnja sredstva pridobi z deljenjem lastnih vsebin in jih lahko porabi za dostop do vsebin drugih uporabnikov.



Slika 1. Struktura platforme MMShare

Vsak prenos sredstev se zabeleži v obliki transakcije v blokovni verigi, s čimer je zagotovljeno, da nihče ne

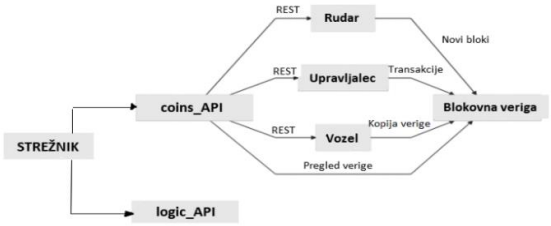
more porabiti istih sredstev večkrat ter tako dostopati do več vsebin, kot jih je sam dal na voljo ostalim. Digitalna sredstva so prenešena neposredno iz enega javnega naslova na drugega brez vmesnega posrednika.

Če uporabnik ne želi več uporabljati platforme, lahko svoj račun izbriše. Pri tem ostane zgodovina njegovih nakupov zapisana v blokovni verigi, se pa zaklenejo njegova digitalna sredstva, ki jih ne more porabiti nihče drug, saj nima dostopa do zasebnega ključa.

Slika 1 prikazuje strukturo platforme. Ta je sestavljena iz zalednega in čelnega dela. Zaledni del sestavljata blokovna veriga ter preostala aplikacijska logika, čelni del pa spletna in mobilna aplikacija.

3.1 Zaledni del

Zaledni del platforme skrbi za shranjevanje in zagotavljanje varnosti podatkov in dostop do blokovne verige. Za izvedbo slednje smo uporabili odprtokodno knjižnico Naivecoin, ki ponuja pripravljeno ogrodje blokovne verige. Ogrodje sestavljajo *Rudar*, ki skrbi za dodajanje novih blokov v verigo, *Upravljalec*, ki skrbi za javne ključe in bilance stanj na računih ter ustvarja nove transakcije za prenos sredstev med računi uporabnikov, in *Vozli*, ki predstavljajo sodelujoče člene omrežja. Struktura zalednega dela je prikazana na Sliki 2.



Slika 2. Struktura zalednega dela platforme MMShare

```
//ustvarimo novo transakcijo s podanimi parametri
let newTransaction =
  operator.createTransaction(walletId,
    fromAddress, toAddress, amount, changeAddress);
//preverimo veljavnost transakcije
newTransaction.check();
//transakcijo dodamo v blokovno verigo
blockchain.addTransaction
  (Transaction.fromJson(newTransaction));
```

Slika 3. Izvedba transakcije s knjižnico Naivecoin

Slika 3 prikazuje izvedbo transakcije z uporabo knjižnice Naivecoin. V prvem koraku transakcijo ustvarimo, pri čemer je potrebno določiti naslov denarnice prejemnika ter naslova pošiljatelja in prejemnika denarnih sredstev (uporabnik bi lahko imel v denarnici več naslovov ali pa bi imel več denarnic), količino nakazanih sredstev in naslov, kamor se prenesejo morebitni stroški transakcije. V drugem koraku preverimo veljavnost transakcije (pravilnost izvlečka, veljavnost podpisa itd.), na koncu pa transakcijo dodamo v blokovno verigo, v kateri je zapisana v formatu JSON (Slika 4).

Za ustvarjanje javnih in zasebnih ključev sodelujočih uporabnikov, namenjenih podpisovanju transakcij za prenos sredstev, uporablja Naivecoin asimetrično šifriranje z eliptičnimi krivuljami, imenovano ECDSA (Elliptic Curve Digital Signature Alghoritm).



Slika 4. Transakcija v blokovni verigi v formatu JSON

Blokovno verigo iz knjižnice Naivecoin smo preoblikovali v spletno storitev REST in jo od ostalega zaledja ločili s komunikacijskim vmesnikom. Če bi bilo to potrebno (na primer za čim večjo izkoriščenost strojne opreme za potrebe rudarjenja), bi se tako storitev, ki skrbi za blokovno verigo, lahko izvajala na lastnem fizičnem strežniku, dostop do verige pa se ne bi spremenil. Knjižnico smo prilagodili tudi tako, da lahko nove zapise v verigo dodaja (rudari) samo strežniški vozel. Vsak rudar je namreč nagrajen z nekaj virtualnimi sredstvi, če je on tisti, ki uspe dodati naslednji blok v verigo. To pa ni v skladu z namenom platforme, saj želimo, da lahko ostali uporabniki sredstva pridobijo le, če nalagajo vsebine.

Tabela 1. REST vmesnik blokovne verige

HTTP metoda	**URI**	**parametri**	**učinek**
POST	/wallets	geslo	dodajanje novega uporabnika
POST	/wallets/ *{walletID}* /transactions	geslo in naslov pošiljatelja, naslov prejemnika, količina prenesenih sredstev	izvedba transakcije za prenos sredstev med dvema računoma
GET	*{addressID}*/ balance	/	pridobivanje stanja na računu
POST	wallets/ *{walletID}*	staro in novo geslo	menjava gesla

Tabela 1 prikazuje dejanja, ki jih je mogoče preko vmesnika REST izvesti v blokovni verigi. Spletna storitev je dostopna na glavnem naslovu */coinsAPI*. Za prenos sredstev z enega uporabnikovega računa na drugega tako na naslov */coinsAPI/wallets/{walletID}/transactions* pošljemo HTTP zahtevek po metodi POST, v telo zahtevka pa vključimo geslo in naslov pošiljatelja, naslov prejemnika in količino prenesenih sredstev. Geslo pošiljatelja je pomembno, ker se iz njega izračuna pošiljateljev zasebni ključ, s katerim se transakcija digitalno podpiše. S tem se zagotovi pristnost transakcije, saj lahko digitalna sredstva porabi le njihov lastnik. Uspešno izvedena transakcija se nato zapiše v blok, ki se doda v blokovno verigo (Sliki 3 in 4).

Preostala aplikacijska logika je izvedena v okviru spletne storitve MMShare. Ta komunicira s storitvijo blokovne verige, obenem pa nudi dostop do funkcionalnosti platforme čelnemu delu s pomočjo lastnega vmesnika

REST, podobnega tistemu za interakcijo z blokovno verigo. Odjemalci lahko z zahtevki na naslov */logic_API* s predpisano HTTP metodo in dodatkom k URI naslovu shranjujejo, prenašajo, brišejo in kupujejo vsebine.

Zaledni del je napisan v javascriptu, izvaja pa se v izvajalnem okolju Node.js s pomočjo ogrodja Express.js. Za shranjevanje vsebin in ostalih podatkov smo uporabili entitetno-relacijsko podatkovno zbirko MongoDB, saj ta v primerjavi s klasičnimi relacijskimi zbirkami nudi večjo fleksibilnost strukture in povezav med podatki.

3.2 Čelni del

Za uporabo platforme MMShare lahko uporabnik izbira med spletno in mobilno aplikacijo. Funkcionalnosti obeh aplikacij vključujejo:

- registracijo in prijavo uporabnika,
- nalaganje novih vsebin in brisanje obstoječih,
- nakup vsebine in njen prenos na svojo napravo,
- filtriranje vsebin po temah ali lastniku.

Izsek iz uporabniškega vmesnika spletne aplikacije prikazuje Slika 5, na kateri so razvidne tudi pravice uporabnika nad posamezno datoteko.



Slika 5. Izsek iz uporabniškega vmesnika spletne aplikacije

Spletna aplikacija je razvita po principu aplikacije na eni strani (ang. Single Page Application, SPA) s programsko knjižnico ReactJS. Mobilna aplikacija uporablja v telefone vgrajen fotoaparat in tako omogoča hitro ustvarjanje in nalaganje datotek na strežnik.

4 Zaključek

Čeprav je tehnologija veriženja blokov razmeroma nova, v zadnjih letih pridobiva na popularnosti in razširjenosti. Njen razmah je spodbudilo zavedanje, da zaupanje v posrednika s seboj prinese tudi določena tveganja, če je ta kompromitiran. Veriženje blokov se poskuša uveljaviti na različnih področjih, vendar zaradi relativne novosti še ne moremo z gotovostjo trditi, na katerih se bo dokončno uveljavila. Na trgu se mesečno pojavi veliko novih aplikacij, ki ponujajo rešitve z uporabo blokovne verige. Novim aplikacijam je mogoče slediti in jih podpreti na platformi za množično financiranje ICO Drops [34].

Za razvoj rešitev, predvsem pri manjših projektih, je smiselno uporabiti odprtokodne knjižnice, saj pohitrijo razvoj aplikacij in nudijo dodaten sloj abstrakcije tako, da se razvijalcem ni potrebno ukvarjati z delovanjem tehnologije veriženja blokov. Pri razvoju platforme za deljenje datotek MMShare smo tako enostavno uporabili knjižnico Naivecoin, jo prilagodili lastnim potrebam in se v nadaljevanju lahko osredotočili na funkcionalnosti same aplikacije.

Po drugi strani so programske knjižnice pogosto podvržene programskim hroščem in je zato potrebno veliko testiranja in prilagajanja programske kode. Pri uporabi odprtokodnih knjižnic v industriji je zato potrebno biti posebej pazljiv, saj mora biti njihovo delovanje brezhibno. Razvijalci zahtevnejših projektov se tako pogosto odločajo za razvoj lastnih komponent ali pa njihov nakup, da pridobijo vso potrebno podporo.

Literatura

[1] Namecoin, https://namecoin.org/
[2] Blockname, https://github.com/telehash/blockname
[3] Storj.io, https://storj.io/
[4] K. Delmolino, M. Arnett, A. Kosba, A. Miller, and E. Shi, "Step by step towards creating a safe smart, contract: Lessons and insights from a cryptocurrency lab," v International Conference on Financial Cryptography and Data Security, 79-94, Springer, 2016
[5] Decentralized Autonomous Organisation – Organisations on the Blockchain, https://blog.codecentric.de/en/2017/09/decentralized-autonomous-organization-blockchain/
[6] IBM Pushes Blockchain into the Supply Chain, https://www.wsj.com/articles/ibm-pushes-blockchain-into-the-supply-chain-1468528824
[7] MedRec, https://medrec.media.mit.edu/
[8] IBM Reveals Proof of Concept for Blockhain-Powered Internet of Things, https://www.coindesk.com/ibm-reveals-proof-concept-blockchain-powered-internet-things/
[9] Blockchain.info, https://www.blockchain.com/en/explorer
[10] Coinbase, https://www.coinbase.com/
[11] Gem, https://gem.co/
[12] BitGO, https://www.bitgo.com/info/
[13] Tierion, https://www.bitgo.com/info/
[14] Chainpoint, https://chainpoint.org/
[15] AWS Blockchain, https://aws.amazon.com/about-aws/whats-new/2018/04/introducing-aws-blockchain-templates/
[16] Blockchain.info, https://github.com/blockchain/
[17] blockchain.info, https://libraries.io/npm/blockchain.info
[18] Blockcypher, https://github.com/blockcypher/php-client
[19] BitGOSDK-PHP, https://github.com/neto737/BitGoSDK-PHP
[20] The Top 3 Blockchain Libraries for Java Devs, https://dzone.com/articles/the-top-3-blockchain-libraries-for-java-devs
[21] Bitcore-lib, https://libraries.io/npm/bitcore-lib
[22] Solidity, http://solidity.readthedocs.io/en/v0.4.24/
[23] Web3j, https://github.com/ethereum/web3.js/
[24] HyperLedger Fabric, https://hyperledger-fabric.readthedocs.io/en/release-1.2/
[25] Mythril, https://libraries.io/pypi/mythril
[26] Solium, https://libraries.io/npm/solium
[27] Embark, https://libraries.io/npm/embark
[28] IPFS, https://ipfs.io/
[29] Blockchain, https://github.com/dvf/blockchain
[30] Jbc, https://github.com/jackschultz/jbc
[31] Noochain, https://github.com/CryptoKass/NoobChain-Tutorial-Part-2
[32] BrewChain, https://github.com/dbjsdev/BrewChain
[33] Naivecoin, https://github.com/conradoqg/naivecoin
[34] ICO Drops, https://icodrops.com/

Primerjava uporabniške izkušnje različnih uporabniških vmesnikov v vozilu

Tomaž Čegovnik, Jaka Sodnik[1]

[1] *Univerza v Ljubljani/ Fakulteta za elektrotehniko, Ljubljana, Slovenija*
E-pošta: tomaz.cegovnik@fe.uni-lj.si, jaka.sodnik@fe.uni-lj.si

Abstract. *In a user study we compared three interfaces in an In-Vehicle Infotainment System in terms of user experience. Free hand gestural interaction and touchpad based interaction were compared to a conventional input interface consisting a button on the steering wheel. The user experience was evaluated with the User Experience Questionnaire (UEQ). The results are showing significant differences, where the conventional buttons on the steering wheel resulted with the best user experience.*

1 Uvod

Področje interakcije človek-stroj v vozilih je hitro razvijajoče se področje, saj je soočeno s hitrim razvojem tehnologij. Informacijsko-komunikacijski sistem v vozilu (ang. *In-Vehicle Infotainment System*, IVIS) v vozilu se razvija v smeri zaslonov na dotik in prostoročno komunikacijo od do sedaj uveljavljene interakcije s tipkami in gumbi. Poleg tega se število funkcionalnosti IVIS povečuje in jih je treba vozniku predstaviti na praktičen in jasen način. Nekateri avtomobilski proizvajalci v svojih vozilih implementirajo celoten IVIS na zaslon na dotik, nekateri pa vpeljujejo kombinacijo gumbov, tipk in zaslona na dotik [1,2]. Uvajajo se tudi različni načini vhodnih naprav kot so sledilne ploščice, prepoznavanje govora ali prepoznavanje prostoročnih kretenj. Proizvajalci niso enotni, katera vrsta uporabniškega vmesnika je najboljša za IVIS. Veliko dejavnikov je potrebno upoštevati - uporabnost, preprostost, estetika, odvračanje pozornosti od vožnje in kakovost uporabniške izkušnje.

Loehmann v svojem prispevku pokaže, da se pomembnost uporabniške izkušnje v avtomobilski industriji stalno povečuje. Navaja, da bi lahko zlasti prostoročna interakcija izboljšala uporabniško izkušnjo brez povečanega vpliva na varnost v prometu. Prav tako se uvajajo različni koncepti, na kakšen način izvajati prostoročne kretnje v vozilih, predvsem z namenom zmanjševanja vizualnega odvračanja voznika od prometa [3]. Raziskave raziskovalne skupine iz podjetja BMW kažejo, da lahko prostoročne kretnje prepoznamo s preprosto sodobno strojno opremo in najpreprostejše kretnje uporabimo za enostavne interakcije z IVIS [4]. BMW je v IVIS vozila serije 7 v letu 2016 [5] vgradil preprost sistem za prostoročno interakcijo.

Zelo obetavni so rezultati uporabe sledilne plošče kot vhodne naprave v IVIS. Burnet je pokazal, da je uporaba sledilne ploščice za preproste naloge kot je nastavitev temperature v vozilu veliko boljša kot zaslon na dotik ali vrtljivi gumbi [6]. Norberg in Rahe sta v svoji študiji ugotovila, da bi lahko s sledilno ploščico razvili vsestransko uporaben uporabniški vmesnik, ki bi bil dobro sprejet med uporabniki [8]. Proizvajalec avtomobilov AUDI v nekaterih svojih vozilih že uporablja različico sledilne ploščice za interakcijo z IVIS v nekaterih prestižnih modelih [7].

Da bi zmanjšali vizualno odvračanje pozornosti pri uporabi IVIS, smo se v tej študiji odločili za uporabo transparentnega prikazovalnika (ang. Head-Up Display, HUD). Študija, ki sta jo izvedla Liu in Wen, kaže tudi, da je pri uporabi HUD odzivni čas voznikov boljši kot pri uporabi običajnih prikazovalnikov. Pri uporabi HUD je tudi manj variance hitrosti in uporaba povzroča manj stresa [9].

Glavno raziskovalno vprašanje te študije je – Ali prostoročna interakcija ali interakcija s sledilno ploščico ponuja boljšo uporabniško izkušnjo kot uporaba tipk na volanih? Izvedli smo uporabniško študijo, ki se osredotoča na ocenjevanje uporabniške izkušnje interakcije z IVIS preko treh različnih vhodni načinov interakcije. Primerjali smo dva sorazmerno nova pristopa, to sta sledilna ploščica in prostoročna interakcija z načinom interakcije, ki je v praski zelo razširjen - interakcija z gumbom na volanu. Primerjava vhodnih vmesnikov je bila opravljena pri dveh različnih težavnostih vožnje.

2 Metodologija

2.1 Simulator vožnje

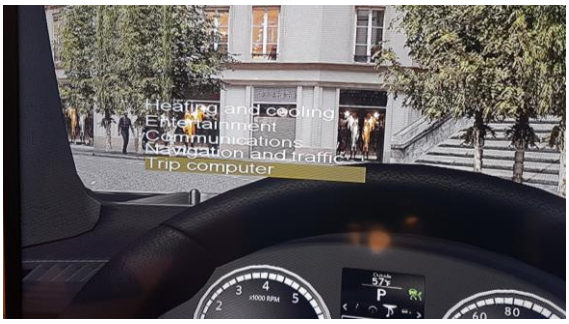
Študija je bila izvedena na zmogljivem simulatorju vožnje (Slika 1). Simulator je sestavljen iz pravega avtomobilskega sedeža, vhodnega avtomobilskega krmilnega sistema Fanatec [10], z modelom volana ClubSport Porsche 918 RSR [11], pedali Fanatec ClubSport V2 [12] in s tremi ukrivljenimi televizijskimi zasloni Samsung z diagonalo 48" [13]. Ta konfiguracija je zelo podobna dejanskemu voznemu okolju in ekrani pokrivajo več kot 120° horizontalnega vidnega kota. Programska oprema uporabljena za simulacijo vožnje je bila OKTAL SCANeR Driver Training [14], ki nam jo je posredovalo podjetje Nervteh [15]. Omogoča ustvarjanje poljubnih scenarijev vožnje in implementacijo IVIS neposredno v programsko opremo simulatorja.



Slika 1: Simulator vožnje.

2.2 Uporabniški vmesnik

Za to uporabniško študijo smo izdelali preprost vmesnik s hierarhično menijsko strukturo, ki je sestavljena iz petih elementov na vsakem nivoju. Odvisno od izbrane naloge ima hierarhični meni 3-5 stopenj. Vozniku je bil meni predstavljen prek HUD. Izbrani element je bil poudarjen z nevtralno zeleno barvo.



Slika 2: HUD: pet vrstic predstavlja možne izbire v meniju, z zeleno bravo je označen trenutni izbran element.

IVIS je bil sestavljen s tremi različnimi načini vnosa (tj. gumbi na volanu, sledilna ploščica in prostoročni vmesnik). Prvi vmesnik je bil volan Fanatec, kjer je bil kot vhodni vmesnik uporabljen gumb na desni strani volana. Ta gumb je možno potisniti navzgor in navzdol – na tak način se premika izbiro vrstice gor ali dol. Potisk v levo je izvedel izbiro označenega elementa, potisk v desno pa je izbiro vrnil v prejšnji nivo menija.

Drugi vhodni vmesnik je sledilna ploščica, ki je bila realizirana na pametnem Android telefonu. Aplikacija na pametnem telefonu je sledila premikanju prsta na zaslonu in pošiljala podatke v programsko opremo simulatorja. Telefon je bil nameščen na desni strani sedeža na ergonomsko ustrezni lokaciji. To lokacijo smo izbrali na podlagi opazovanj v obstoječih vozilih, v katerih so nameščene podobne naprave. Ta položaj omogoča, da voznik komolec nasloni na naslon za roko, sledilno polje pametnega telefona pa doseže brez posebnega napora. Premikanje po zaslonu naprej in nazaj je predstavljalo premikanje izbire po meniju gor in dol. Drsenje po plošči v desno je predstavljalo izbiro trenutno izbrane opcije, drsenje v levo pa vrnitev na prejšnji nivo menija

Za tretji vhodni vmesnik je bil uporabljen krmilnik Leap Motion, ki zaznava prostoročne kretnje. Sistem je sledil položaju in usmeritvi dlani. Posamezen element v meniju je bilo moč izbrati s spreminjanjem usmeritve dlani. Usmeritev dlani navzgor pomeni izbiro zgornje vrstice, usmeritev navzdol izbiro spodnje, položaji vmes pa zvezno izbiro vmesnih vrstic. Za potrditev izbire je moral uporabnik držati dlan še v želenem položaju 1000 ms. Da bi se vrnili na prejšnjo raven menija, je moral uporabnik zavrteti dlan za 90 stopinj v desno.

2.3 Načrt poskusa

V študiji je sodelovalo 16 moških in 14 žensk. Udeleženci so bili razdeljeni v dve skupini. Skupina A je vozila zelo preprost scenarij vožnje po podeželski cesti. V tem scenariju ni bilo križišč in naloga uporabnika je bila le sledenje sprednjemu vozilu. Skupina B je vozila v zahtevnejšem mestnem okolju z visoko gostoto prometa. Navigacijski sistem je vodil voznika po prej določeni trasi skozi križišča.

Med vožnjo je moral vsak udeleženec opraviti sedem različnih nalog (npr. Nastavite temperaturo na 23°C, spremenite radijsko postajo, preverite sporočilo itd.). Prometne razmere in vrsta vmesnika sta bili neodvisni spremenljivki v študiji. Vsak udeleženec je izvedel komplet sedmih nalog po tri krat, vsakič z različnim vhodnim vmesnikom in nekoliko drugačnim naborom nalog, vedno z enakimi prometnimi pogoji. Zaporedje uporabljenih vmesnikov je bilo naključno.

2.4 Naloge

Pripravili smo tri različne sklope nalog, saj je vsak udeleženec opravljal naloge na treh različnih vmesnikih. Vsak sklop je vseboval tri preproste naloge, pri katerih je bilo za dokončanje nalog potrebno manj kot pet posameznih korakov. Zadnji, potrditveni korak je bil vedno na drugem ali tretjem nivoju menija, npr. `Temperature->Seat warmers->On`. Poleg tega so bile v vsakem sklopu tudi štiri težje naloge, ki so potrebovale več korakov, končen korak pa je bil na najnižjem nivoju hierarhičnega menija, npr. `Entertainment->Music->Authors->Justin Bieber->Baby`.

2.5 Ocenejvanje uporabniške izkušnje

Uporabniško izkušnjo smo ocenjevali z vprašalnikom o uporabniški izkušnji (ang. *User Experience Questionnaire,* UEQ) [16]. Celotna študija je bila izvedena v slovenskem jeziku in uporabljena je bila slovenska različica UEQ. Udeleženec je UEQ izpolnil trikrat, vedno takoj po zaključku enega sklopa nalog z uporabo enega vhodnega vmesnika.

UEQ je nabor 26 vprašanj z odgovori od -3 do 3, kjer ekstremi predstavljajo dva nasprotna opisa za en parameter. Na primer, pari opisov za vrednotenje privlačnosti so privlačen-neprivlačen, prijazen-

neprijazen ali prijeten-neprijeten. UEQ ocenjuje šest različnih kategorij: privlačnost, preglednost, učinkovitost, zanesljivost, stimulacija in novost. Nabor vprašanj in orodje za vrednotenje je dosegljivo na spletu [16].

3 Rezultati

Rezultati so bili izračunani z orodjem za analizo podatkov UEQ, ki je na voljo na spletni strani UEQ. Uporabniško izkušnjo smo izračunali za vsako skupino in vsako vhodno napravo posebej, kar nam da skupno šest rezultatov. Slike 3-5 predstavljajo merilo uspešnosti posamezne naprave. Črna črta je rezultat UEQ za posamezen sklop vprašanj. Barvni histogrami predstavljajo referenčne vrednosti meril uspešnosti. Referenčne vrednosti so pridobljene iz primerjalne podatkovne baze, ki vsebuje podatke od 9905 oseb iz 246 različnih študij in je na voljo v orodju za analizo podatkov UEQ [16]. Območje *Excellent* vsebuje 10% najboljših rezultatov; *Good* pomeni, da je 10% rezultatov boljših, 75% slabših; *Above average* pomeni 25% boljših, 50% slabših; *Below average* pomeni 50% boljših rezultatov in 25% slabših; *Bad* je v območju 25% najslabših rezultatov.

4 Razprava

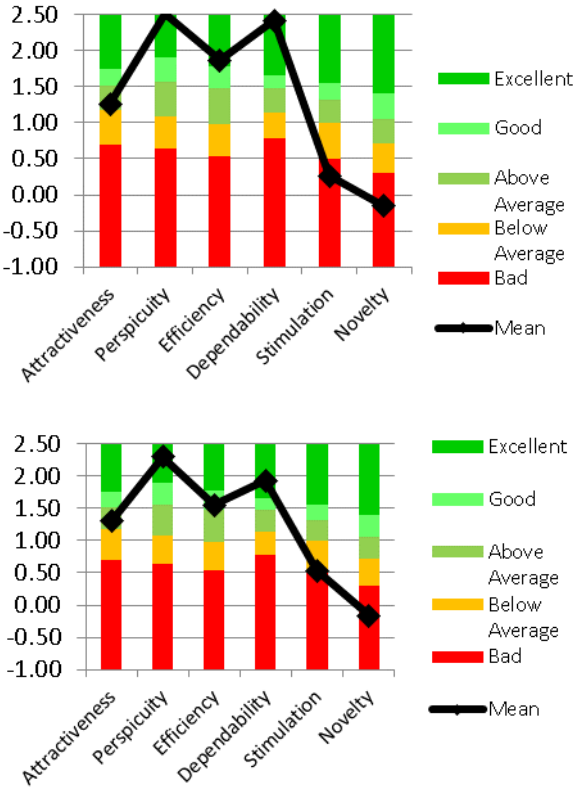
Ocenili smo uporabniško izkušnjo uporabe različnih načinov vnosa v uporabniški vmesnik za uporabnike vozil z UEQ.

Rezultati kažejo pričakovane in jasne razlike med različnimi vhodnimi napravami, ki smo jih na tak način eksperimentalno potrdili. Praktična kakovost, ki je kombinacija preglednosti, učinkovitosti in zanesljivosti ter opisuje kakovost, povezano z opravljanjem nalog, je najvišja za uporabo gumba na volanu. Ta rezultat je pričakovan, saj je takšen vhodni vmesnik v zelo široki uporabi v avtomobilski industriji. Praktična kakovost je za interakcijo s sledilno ploščico nad povprečjem in zelo slaba za interakcijo z prostoročnimi kretnjami.

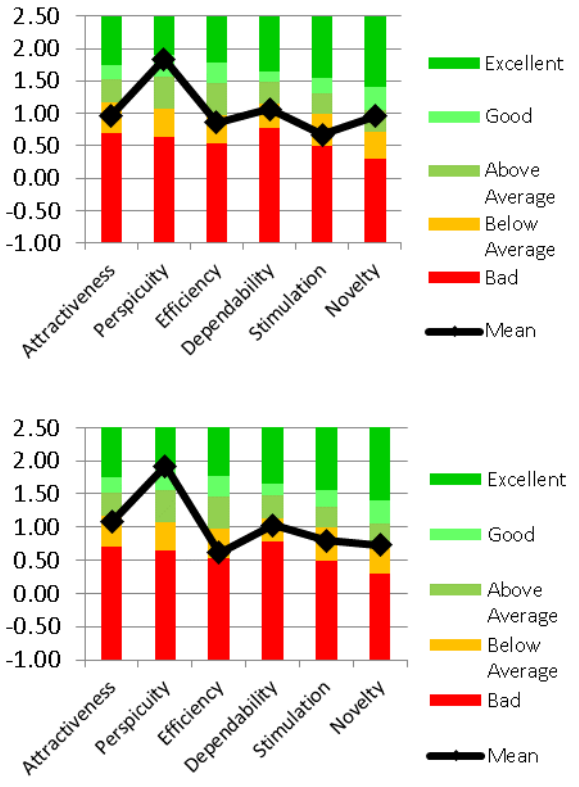
Kakovost, ki je povezana s stimulacijo in novostjo, predstavlja vse vidike kakovosti, ki niso povezani z opravljanjem nalog. Rezultati tega vrednotenja kažejo ravno nasprotje praktična kakovosti. Uporaba gumba na volanu ima slab rezultat, medtem ko ima prostoročna interakcija dober rezultat. Sledilna ploščica je spet med obema drugima dvema rezultatoma.

Pričakovali smo boljše rezultate v preprosti prometni simulaciji, ker tu ni dodatnih prometnih motenj in ni potreben velik napor za vožnjo. Rezultati so v nasprotju s temi pričakovanji, saj dve različni prometni situaciji ne vplivata na kakovost uporabniške izkušnje.

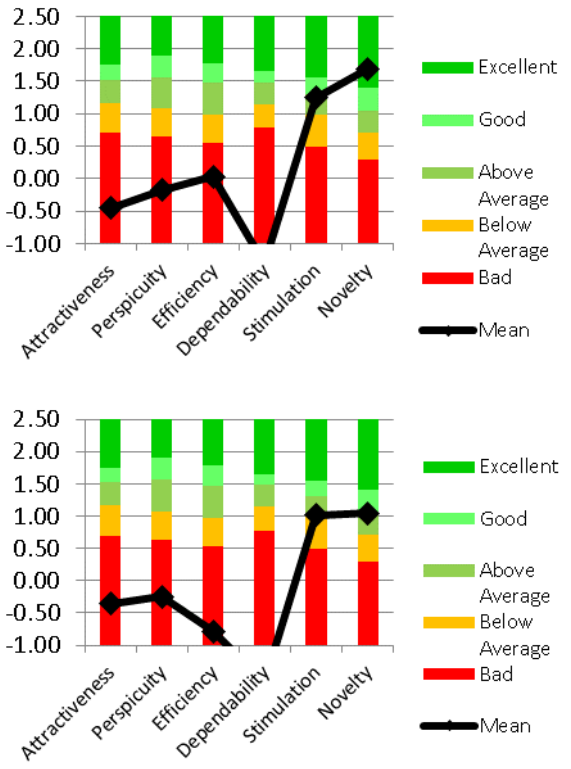
Splošno uveljavljena interakcija s tipko na volanu, ki se večinoma uporablja v sedanjih avtomobilih, kaže najboljše rezultate v smislu uporabnosti in kakovosti povezano z dejanskim opravljanjem nalog na napravi. Interakcija s prostoročnimi kretnjami ima zelo dobre ocene za svojo izvirnost in ima perspektivo za uporabo v vozilih.



Slika 3: Rezultati UEQ za gumbe na volanu. Zgoraj – enostaven promet, Spodaj – težaven promet.



Slika 4: Rezultati UEQ za sledilno ploščico. Zgoraj – enostaven promet, Spodaj – težaven promet.



Slika 5: Rezultati UEQ za prostoročno interakcijo.
Zgoraj – enostaven promet, Spodaj – težaven promet.

Trenutna izvedba sistema prostoročne interakcije v tej študiji sicer ni dovolj učinkovita, da bi se jo lahko uporabilo za glavno vhodno napravo IVIS. Za uporabo v pravih vozilih bi morali uporabiti drugačne gibe in izbrati nabor lažjih funkcij pri interakciji z IVIS. Sledilna ploščica ni pokazala odličnih rezultatov, ampak je pa nadpovprečna v taki izvedbi v tej študiji. Verjetno ni popolna za uporabo kot glavna vhodna naprava za IVIS, vendar bi jo lahko uporabili kot del več-modalnega uporabniškega vmesnika.

Ugotavljamo, da je gumb na volanu vmesnik, ki je najbolj učinkovit glede na uporabniško izkušnjo in ima najboljšo praktično kakovost. Sledilna ploščica in prostoročne kretnje sta zelo privlačna in z izvedbo, ki bi upoštevala omejitve teh vmesnikov, bi taki vmesniki bili lahko dobro sprejeti med vozniki.

Zahvala

Raziskava je bila podprta s strani ARRS. Avtorji se zahvaljujemo podjetju NERVteh za zagotavljanje merilne opreme.

Literatura

[1] Riener, A., Rossbory, M. (2011) Natural and Intuitive Hand Gestures: A Substitute for Traditional Vehicle Control?. In Adjunct Proc. of AutomotiveUI'11.

[2] Liu, Y.-C. (2001) Comparative Study of the Effects of Auditory, Visual and Multimodality Displays On Drivers' Performance in Advanced Traveller Information Systems. Ergonomics 44(4), 425-442.

[3] Loehmann, S.; Diwischek, L.; Schröer, B., „The User Experience of Freehand Gestures", In: *User Experience in Cars Workshop at INTERACT 2011 - 13th IFIP TC13 Conference on Human-Computer Interaction*, pp. 54-58, 2011.

[4] Althoff, F.; Lindl, R.; Walchshäusl, L., "Robust Multimodal Hand- and Head Gesture Recognition for Controlling Automotive Infotainment Systems", In: VDI-Tagung: Der Fahrer im 21. Jahrhundert, 2005.

[5] Bmw Series 7 (2017) http://driving.ca/bmw/7-series/auto-news/news/how-it-works-bmw-gesture-control

[6] G. Burnett, G. Lawson, L. Millen, and C. Pickering, 'Designing touchpad user-interfaces for vehicles: which tasks are most suitable?', Behaviour & Information Technology, vol. 30, no. 3, pp. 403–414, May 2011.

[7] Audi MME Touch (2018) https://www.audi-technology-portal.de/en/electrics-electronics/controls/mmi-touch_en

[8] Norberg, S., & Rahe, U. (2009). Touchpad as interaction input control for use of In-Vehicle infotainment system. In *1st International Conference on Driver Distraction and Inattention (DDI 2009) Chalmers University of Technology, SwedenSAFER Vehicle and Traffic Safety CentreINRETS-ARCUEIL, FRANCE.*

[9] Y.-C. Liu and M.-H. Wen, 'Comparison of head-up display (HUD) vs. head-down display (HDD): driving performance of commercial vehicle operators in Taiwan', International Journal of Human-Computer Studies, vol. 61, no. 5, pp. 679–697, Nov. 2004.

[10] FANATEC (2017) https://www.fanatec.com/

[11] Fanatec ClubSport Steering Wheel Porsche 918 RSR (2017) https://www.fanatec.com/eu-en/steering-wheels/clubsport-steering-wheel-porsche-918-rsr-eu.html

[12] Fanatec ClubSport Pedals V2 (2017) https://www.fanatec.com/eu-en/support/product;old_product_id;252

[13] Samsung Curved TV (2017) http://www.samsung.com/uk/tvs/full-hd-j6300/UE48J6300AKXXU/

[14] Oktal (2017) http://www.oktal.fr/en/automotive/range-of-simulators/software

[15] NERVteh (2017), http://nerv-teh.com/.

[16] User Experience Questionnaire (2018) http://www.ueq-online.org/

Uporaba prilagojenih vprašalnikov UEQ za povečanje dostopnosti za gluhe in naglušne

Ana Berkopec[1], Matevž Pogačnik[1], Klemen Pečnik[1]

[1]*Univerza v Ljubljani, Fakulteta za elektrotehniko, Laboratorij za multimedijo, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: ana.berkopec@ltfe.org, matevz.pogacnik@fe.uni-lj.si, klemen.pecnik@ltfe.org

Adaptation of the UEQ for the Deaf and Hard of Hearing

Abstract. *The number of Deaf and Hard of Hearing people worldwide is estimated at a few hundred million. Even though there are already some applications and other products designed specifically for or adapted for them, there is still no standardized questionnaire for evaluation of user experience of the Deaf and Hard of Hearing. Since the vast majority has trouble with reading and writing, the questionnaire should be adapted for them by translating questions into a specific or an international sign language followed by interviews about users' experiences with the help of an interpreter, virtual avatar or written sign language (SignWriting). The article also includes suggestions for improved translations of UEQ (User Experience Questionnaire) terms to Slovenian language and introduces translations into SSL (Slovenian Sign Language).*

1 Uvod

Po statutu Mednarodne zveze naglušnih oseb pojem "naglušen" opredeljuje osebe z izgubo sluha z omejeno sposobnostjo poslušanja in razumevanja, ki običajno komunicirajo z govorom [1]. Gluhe osebe pa opredeljuje izguba sluha določenih frekvenc na ravni 91 dB ali več [2]. S tem se pojavi težava tako pri komunikaciji z drugimi, kot tudi pri uporabi informacijsko-komunikacijskih tehnologij (IKT). Po podatkih svetovne zdravstvene organizacije (WHO) je danes gluhih že 466 milijonov ljudi, oziroma več kot 5% svetovne populacije. Vsi ti imajo pri uporabi aplikacij težave z dostopnostjo. Čeprav nekatere aplikacije za gluhe in naglušne že obstajajo, se premalo poudarka namenja tako razvoju ustreznih uporabniških vmesnikov, kot tudi merjenju same uporabniške izkušnje. Tudi metode razvoja uporabniških vmesnikov in merjenje uporabniške izkušnje niso standardizirane za gluhe in naglušne uporabnike, podobne težave pa obstajajo tudi za slepe in slabovidne. V tem članku bo izveden pregled možnih rešitev za prilagoditev vprašalnikov za merjenje uporabniške izkušnje, ter možna uporaba novih tehnologij pri prilagajanju televizijskega programa gluhim in naglušnim. Podan je tudi predlog prevodov pojmov UEQ vprašalnika, ki bo služil kot osnova za izdelavo vprašalnika s pomočjo navideznega animiranega tolmača znakovnega jezika.

2 Načini komunikacije

Komunikacijske sposobnosti gluhih in naglušnih oseb so odvisne od njihove stopnje naglušnosti. Poleg tega se pojavlja večja razlika med tistimi, ki so bili že rojeni gluhi ali naglušni, ter tistimi, ki so sluh izgubili kasneje. Prvi običajno kot materni jezik uporabljajo znakovni jezik (v Sloveniji je to SSL – slovenski znakovni jezik, angl. Slovenian Sign Language), ter jim branje in pisanje povzročata težave, saj se pisane besede učijo kot drugi jezik. Velik vpliv na način komunikacije ima tudi družina, v kateri je taka oseba rojena. Če je tudi ostala družina gluha ali naglušna, se taka oseba že od malega srečuje z znakovnim jezikom in lahko razvije boljše komunikacijske sposobnosti, kot tisti, ki se sprva pogovarjajo z uporabo gest, ki si jih izmislijo sami, ter se pravilnega znakovnega jezika naučijo šele v šoli. Medtem pa se tisti, ki so sluh izgubili kasneje, običajno ne zanašajo le na znakovni jezik, saj brez težav uporabljajo tudi zapisano besedo [3]. Približno 30%-40% gluhih in naglušnih ljudi zna govor sogovornika brati iz ustnic [4].

3 Dostopnost

Uporabniki, ki so gluhi ali naglušni že od rojstva, imajo pogosto težave z branjem in pisanjem, zato tudi aplikacije in programi, ki temeljijo izključno na besedilu niso primerni za njih, še posebej, če je v njih uporabljen kompleksen jezik [3].

3.1 Prilagoditev uporabniškega vmesnika

Za omogočanje dostopnosti posameznega izdelka gluhim in naglušnim se uporabljajo različni pripomočki:

- znakovni jezik,
- preprost jezik,
- video,
- napisi,
- grafično hiperbesedilo (linki),
- izpisi,
- široka struktura pri pogosti uporabi in
- globoka struktura pri redkejši uporabi.

Od zgoraj naštetih se prvi dve osredotočata na človeka, ostale pa na tehnologijo. Zadnji dve se nanašata tudi na navigacijo.

Pri aplikacijah, ki so namenjene gluhim in naglušnim je potrebno omogočiti jasen nadzor nad glasnostjo, povratnimi informacijami z vibriranjem ali vidnim opozorilom, ter vizualne ali taktilne indikatorje za tipkovnico [5].

Razvoj uporabniških vmesnikov danes večinoma temelji na množični uporabi, smernice in priporočila za povečanje dostopnosti pa so pogosto prezrta in neupoštevana. Veliko razvijalcev ali skupin iniciative W3C WCAG (angl. Web Content Accessibility Guidelines) ne jemlje kot smernice ampak kot nujno zlo, zato so pogosto tudi rešitve za povečanje dostopnosti implementirane samo deloma, tako da zgolj zadostijo minimalnim zahtevam.
Povečanje oziroma izboljšanje dostopnosti in razširitev uporabnosti na vse uporabnike za razvijalce namreč predstavlja kar nekaj dodatnega dela in finančnega vložka v razvoj.

3.2 Namenske aplikacije

Za omogočanje pogovora z nekom, ki ne zna znakovnega jezika, je gluhim ali naglušnim na voljo tolmač znakovnega jezika, ki prevaja govor v znakovni jezik in obratno. Prisotnost tretje osebe ni vedno mogoča, zaradi različnih omejitev, zato se danes razvija in ponekod že uporablja virtualne tolmače, ki omogočajo računalniško generiran prikaz znakovnega jezika brez prisotnosti tretje osebe [3]. Z njihovo uporabo je mogoče besedilo ali zvok prevesti v izbrani znakovni jezik, so pa tovrstni sistemi dobro podprti in delujoči predvsem za bolj razširjene jezike kot npr. britanski, ameriški in japonski znakovni jezik.

4 Vprašalniki za merjenje uporabniške izkušnje

Za merjenje uporabniške izkušnje se uporablja različne načine, ki poskušajo čim bolj celovito zajeti uporabnikov odziv in občutke ob uporabi produkta. V nadaljevanju je podan pregled nekaterih vprašalnikov, ki se uporabljajo za merjenje uporabniške izkušnje.

4.1 Vprašalnik za uporabniško izkušnjo (UEQ)

UEQ (angl. User Experience Questionnaire) je standardiziran vprašalnik, ki je sestavljen iz 26 ocen. Vsaki oceni pripadata 2 opisna pojma – pozitivni in negativni. Te lahko uporabnik oceni od -3 (popolno strinjanje z negativnim pojmom) do 3 (popolno strinjanje s pozitivnim pojmom) na 7-stopenjski Likert lestvici [6].
Vprašanja se nanašajo na 6 dimenzij ocenjevanja:

1. atraktivnost,
2. jasnost,
3. učinkovitost,
4. zanesljivost,
5. stimulacija in
6. inovativnost [7].

4.2 AttrakDiff

Tudi ta vprašalnik uporablja 7-stopenjsko skalo ocenjevanja in omogoča izdelavo vprašalnika za oceno posameznega izdelka, ali za primerjavo dveh.
Loči štiri področja:

1. kvaliteto, zagotovljeno s strani oblikovalca,
2. subjektivna percepcija in ocena kvalitete,
3. neodvisne pragmatične in hedonične lastnosti, ter
4. vedenjske in čustvene posledice [8].

4.3 Lestvica za uporabnost sistema (SUS)

SUS (angl. System Usability Scale) temelji na ISO 9241-11 standardu, ki priporoča upoštevanje učinkovitosti in zadovoljstva med razvojem posameznega izdelka. Sestavljen je iz 10 vprašanj, ki jih uporabnik ocenjuje na 5-stopenjski Likert lestvici in efektivno ločuje uporabne sisteme od neuporabnih [9].

5 Prilagoditev vprašalnikov gluhim osebam

Vsi vprašalniki opisani v prejšnjem poglavju niso prilagojeni za gluhe in naglušne. Merjenje uporabniške izkušnje za aplikacije, ki so namenjene njim, še ni standardizirano in tudi jasne smernice za ustrezne prilagoditve še ne obstajajo. Zato bi bil potreben razvoj standardiziranega vprašalnika za gluhe, ali pa ustrezen prevod že obstoječih vprašalnikov v znakovni jezik. Ker se znakovni jeziki med seboj razlikujejo (slovenski, ameriški, japonski, nemški,... znakovni jezik) bi lahko za poenotenje vprašalnika tudi med državami uporabljali mednarodni znakovni jezik (IS – International Sign) [10].
Vprašalnike in pojasnila glede merjenja uporabniške izkušnje bi lahko uporabnikom predstavili s pomočjo Stokoeve notacije, znakovnega pisanja (angl. »SignWriting«), tolmača znakovnega jezika ali virtualnega tolmača znakovnega jezika.

5.1 Stokoeva notacija

Stokoeva notacija je bil prvi zapis, uporabljen za znakovni jezik. Temelji na fonemih, ter latinskih črkah in številkah z dodanimi simboli za zapis pozicij, gibanja in orientacije rok. Notacijo sicer uporabljajo predvsem jezikoslovci in akademiki [11].

5.2 Znakovno pisanje

Sistem znakovnega pisanja je bil razvit leta 1974 in omogoča zapis znakovnega jezika v slikovnem notacijskem sistemu. Uporablja ikone, ki predstavljajo oblike rok, orientacije dlani, obrazne izraze, ter indikatorje gibanja in kontaktov [12]. Čeprav je sistem intuitiven za uporabnike znakovnega jezika, je potrebno nekaj dodatnega znanja in prepoznavanja simbolov za gladko branje ali pisanje. Za razliko od Stokoeve notacije, ki se jo bere od leve proti desni, so znaki pri znakovnem pisanju urejeni vertikalno in se jih bere od zgoraj navzdol.

5.3 Tolmač znakovnega jezika (interpreter)

Tolmač znakovnega jezika je prevajalec med govorjenim in znakovnim jezikom. V primeru pomoči tolmača pri merjenju uporabniške izkušnje z gluhimi ali naglušnimi uporabniki, je potrebna stalna prisotnost tolmača ter sprotno prevajanje v govorjeni jezik.

5.4 Virtualni tolmač

Ker zapisana oblika znakovnega jezika izgubi dinamičnost, ki je za znakovni jezik sicer značilna, je za boljše razumevanje in podporo bolj primerna uporaba znakovnega jezika v gibajoči obliki. Z uporabo virtualnega tolmača (avatarja), oziroma 3D animacije virtualnega človeka, ki tolmači znakovni jezik, je mogoče gluhim in naglušnim olajšati komunikacijo s pomočjo tehnologije. Z uporabo znakovnega pisanja, ter dodatnih informacij o gibanju je možno avtomatsko generirati modeliran znakovni jezik, ki ga prikazuje 3D animirani navidezni tolmač znakovnega jezika [12].
V LMMFE (Laboratorij za Multimedijo Fakultete za elektrotehniko) v okviru diplomskih, magistrskih in doktorskih nalog ter projektnega sodelovanja poteka razvoj virtualnega tolmača slovenskega znakovnega jezika, ki bo omogočal lažjo komunikacijo gluhim in naglušnim tako pri izvajanju meritev uporabniške izkušnje, kot pri prilagajanju televizijskega programa gluhim in naglušnim. Na Slika 5.1 je prikazan 3D model virtualnega tolmača.



Slika 5.1: 3D model za virtualnega tolmača znakovnega jezika

5.5 Prilagoditev UEQ za gluhe in naglušne

Vprašalnik UEQ sicer obstaja tudi v slovenskem prevodu [13], vendar bi lahko za posamezna področja uporabili tudi bolj ustrezne prevode. Zato je bil izdelan predlog z manjšimi prilagoditvami prevodov v slovenski jezik. Znakovni jeziki imajo besedišče bolj omejeno kot govorjeni jeziki, saj npr. ne poznajo sklanjatev in mnogo drugih oblik besed, kot jih poznamo pri pisanih besedilih, zato je prevod v znakovni jezik še toliko težji in bolj tog. Na podlagi slovarja slovenskega znakovnega jezika je bil izdelan tudi predlog vprašalnika UEQ za gluhe in naglušne v slovenskem znakovnem jeziku. Vprašanja oz. tabela pojmov, skupaj z uradnimi, predlaganimi prevodi in prevodi v Slovenski znakovni jezik so razvidna iz Tabela 6.1. Pri tem je potrebno poudariti, da je poleg samega prevoda vprašalnika zelo pomembna tudi ustrezna predstavitev namena vprašalnika in poteka merjenja tako pred pričetkom testiranja, kot tudi pred pričetkom izpolnjevanja vprašalnika. Na začetku tolmač uporabnika najprej seznani s potekom testiranja ter z načinom poteka meritev. Po uvodu uporabnik prične s testiranjem uporabniškega vmesnika/aplikacije in po končanem testiranju tolmač na kratko predstavi še vprašalnik in sam potek. Vprašalnik je sestavljen iz videa virtualnega tolmača znakovnega jezika ter grafično ponazoritvijo 7 stopenjske Likert lestvice z enostavnim upravljanjem. Video je razdeljen na 2 dela, v levem delu tolmač prikazuje negativni pojem (-3), v desnem pa pozitivni pojem (+3) posameznega vprašanja. Video se ponavlja do potrditve odgovora uporabnika. Po potrditvi uporabnika se pojavi v sredini okna video tolmača, ki uporabnika pozove k potrditvi odgovora in nato program nadaljuje na naslednje vprašanje. Po izpolnitvi vprašalnika se virtualni tolmač uporabniku še zahvali za sodelovanje.

5.6 Uporaba prilagoditev

Zaradi boljše ustreznosti je bila predlagana metoda izvedbe meritev na napravah, na katerih se izvaja tudi testiranje. V primeru testiranja na TV sprejemnikih se meritve prav tako izvede na TV ekranu s pomočjo HbbTV (hibridna televizija - angl. Hybrid Broadcast Broadband Television) aplikacije, ki omogoča ustrezno izvedbo predlagane rešitve, tako s stališča predstavitve, kot tudi izvedbe meritev in shranjevanja podatkov v ustrezno bazo. V kolikor izvedba meritev ni mogoča na testirani napravi, se vprašalnik predstavi in izvede na računalniku, pri čemer se lahko uporabi isto ogrodje, saj tudi HbbTV aplikacija temelji na spletnih tehnologijah.
Poleg opisane uporabe, bo virtualnega tolmača mogoče uporabiti tudi pri televizijskih oddajah, ki vsebujejo omejen nabor besedišča – npr. vremenska napoved. Lete bi lahko tolmač iz pisnega besedila prevajal v slovenski znakovni jezik, ter samodejno generiral primerno animacijo kretenj.

6 Zaključek

Vsekakor bi bilo potrebno pri razvoju uporabniških vmesnikov in merjenju uporabniške izkušnje ustrezno upoštevati tudi smernice za povečanje dostopnosti. Prvi korak za povečanje dostopnosti je tudi prilagoditev standardnih metod in vprašalnikov, ki bi jih bilo potrebno prilagoditi tudi za osebe s posebnimi potrebami, še posebno, če so aplikacije namenjene prav njim. Zgolj osnovne prilagoditve niso zadostne, potrebne so spremembe že v samem načrtovanju, kjer je potrebno pri postavljanju uporabnika v ospredje upoštevati vse uporabnike. Sčasoma, ob ustrezni standardizaciji in zakonodaji lahko tak pristop k načrtovanju vmesnikov in merjenju uporabniške izkušnje bistveno pripomore k izboljšanju stanja na področju dostopnosti in enakosti za vse uporabnike. Razvoj tehnologij in storitev danes omogoča velik napredek pri zagotavljanju dostopnosti, potrebno jih je le ustrezno povezati in uporabiti, ter storitve in aplikacije približati tudi manjšinam, ki so zaradi takih ali drugačnih omejitev sedaj izločene.
Čeprav se članek osredotoča na prilagoditev za gluhe, je pristop in predloge mogoče razširiti tudi pri prilagoditvah za slepe in slabovidne, ter morda tudi za osebe z govorno-jezikovnimi motnjami.

	SLOVENSKI ZNAKOVNI JEZIK	PREDLOG NOVEGA SLOVENSKEGA POJMA	SLOVENSKI POJEM	ANGLEŠKI POJEM	NEMŠKI POJEM	NEMŠKI POJEM	ANGLEŠKI POJEM	SLOVENSKI POJEM	PREDLOG NOVEGA SLOVENSKEGA POJMA	SLOVENSKI ZNAKOVNI JEZIK
1	zoprno	zoprn	nerazveseljiv	annoying	unerfreulich	erfreulich	enjoyable	razveseljiv	prijeten	prijeten
2	ne razumljiv	nerazumljiv	nerazumljiv	not understandable	unverständlich	verständlich	understandable	razumljiv	razumljiv	razumljiv
3	uporabiti fantazija / je domislica	kreativen	kreativen	creative	kreativ	phantasielos	dull	brez fantazije	nedomiseln	brez ideja
4	enostaven za učenje	enostaven za učenje	enostaven za učenje	easy to learn	leicht zu lernen	schwer zu lernen	difficult to learn	težak za učenje	težak za učenje	težko za učenje
5	koristen	dragocen	osvežilen	vaulable	wertvoll	minderwertig	inferior	uspavajoč	manjvreden	ne koristen
6	dolgočasen	dolgočasen	dolgočasen	boring	langweilig	spannend	exciting	napet	napet	napet
7	ne zanimiv	nezanimiv	nezanimiv	not interesting	uninteressant	interessant	interesting	zanimiv	zanimiv	zanimiv
8	ne možno predvideti	nepredvidljiv	nepredvidljiv	unpredictable	unberechenbar	voraussagbar	predictable	predvidljiv	predvidljiv	možno predvideti
9	hiter	hiter	hiter	fast	schnell	langsam	slow	počasen	počasen	počasen
10	ima avtentičnost	izviren	nov	inventive	originell	konventionell	conventional	star	običajen	običajno
11	ovirati uporabnik	ovira uporabnika	se ne da upravljati	obstructive	behindernd	unterstützend	supportive	se z lahkoto upravlja	pomaga uporabniku	pomagati uporabnik
12	dober	dober	dober	good	gut	schlecht	bad	slab	slab	slabo
13	kompliciran	kompliciran	kompliciran	complicated	kompliziert	einfach	easy	enostaven	enostaven	enostaven
14	ne privlačen	neprivlačen	odbijajoč	unlikable	abstoßend	anziehend	pleasing	privlačen	privlačen	privlačen
15	običajno	običajen	zastarel	usual	herkömmlich	neuartig	leading edge	moderen	novost	novost
16	ne prijeten	neprijeten	neprijeten	unpleasant	unangenehm	angenehm	pleasant	prijeten	prijeten	prijeten
17	možno računati na njega	zanesljiv	napovedljiv	secure	sicher	unsicher	not secure	nenapovedljiv	nezanesljiv	ne možno računati na njega
18	spodbujati	spodbujajoč	raznolik	motivating	aktivierend	einschläfernd	demotivating	enoličen	zavirajoč	ustaviti motiviranje
19	ustrezati pričakovati	ustreza pričakovanjem	zanesljiv	meets expectations	erwartungskonform	nicht erwartungskonform	does not meet expectations	nezanesljiv	ne ustreza pričakovanjem	ne ustrezati pričakovati
20	ne učinkovit	neučinkovit	ni učinkovit	inefficient	ineffizient	effizient	efficient	učinkovit	učinkovit	učinkovit
21	jasen	jasen	pregleden	clear	übersichtlich	verwirrend	confusing	ustvarja zmedo	zapleten	ne jasen
22	ne praktično	nepraktičen	deluje zatikajoče	impractical	unpragmatisch	pragmatisch	practical	deluje tekoče	praktičen	praktično
23	urejen	organiziran	prostoren	organized	aufgeräumt	überladen	cluttered	prenatrpan	natrpan	razmetan
24	lep	lep	lep	attractive	attraktiv	unattraktiv	unattractive	grd	grd	grd
25	prijazen za uporabnik	prijazen za uporabnike	simpatičen	friendly	sympatisch	unsympathisch	unfriendly	nesimpatičen	neprijazen za uporabnika	ne prijazen za uporabnik
26	zastarelo	konzervativen	nevpadljiv	conservative	konservativ	innovativ	innovative	vpadljiv	inovativen	je inovacija

Tabela 6.1: Prevodi pojmov vprašalnika UEQ za merjenje uporabniške izkušnje

Literatura

[1] „Naglušnost - Zveza-GNS.si“. [Na spletu]. Dostopno na: http://www.zveza-gns.si/o-zvezi/o-naglusnosti/. [Pridobljeno: 12-jul-2018].

[2] „Gluhost - Zveza-GNS.si“. [Na spletu]. Dostopno na: http://www.zveza-gns.si/o-zvezi/o-gluhoti/. [Pridobljeno: 10-jul-2018].

[3] J. Cromartie, B. Gaffey, in M. Seaboldt, „Evaluating Communication Technologies for the Deaf and Hard of Hearing“, Worcester Polytechnic Institute, An Interactive Qualifying Project submitted to the Faculty of WORCESTER POLYTECHNIC INSTITUTE in partial fulfillment for the requirements for the Degree of Bachelor of Science, mar. 2012.

[4] A. Young in R. Hunt, „Research with d/Deaf people“. Social School for Social Care Researh, National Institute for Health Research.

[5] G. Yeratziotis in D. Van Greunen, „Making ICT Accessible for the Deaf“, *IST-Afr. Conf. Proc.*, let. 2013.

[6] „Likert scale“, *Wikipedia*. 31-maj-2018.

[7] M. Schrepp, „User Experience Questionnaire Handbook“. 01-jan-2018.

[8] „AttrakDiff“. [Na spletu]. Dostopno na: http://attrakdiff.de/index-en.html. [Pridobljeno: 12-jul-2018].

[9] „System usability scale“, *Wikipedia*. 13-okt-2017.

[10] „International Sign“, *Wikipedia*. 08-jul-2018.

[11] „Stokoe notation“, *Wikipedia*. 03-apr-2018.

[12] Y. Bouzid in M. Jemni, „An Avatar based Approach for Automatically Interpreting a Sign Language Notation“, *2013 IEEE 13th Int. Conf. Adv. Learn. Technol.*, let. 2013.

[13] „UEQ“. [Na spletu]. Dostopno na: https://www.ueq-online.org/. [Pridobljeno: 07-jul-2018].

Nadgradnja izobraževalnega okolja E-CHO

Martin Kozmelj[1], Marko Papić[1]

[1]Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: martin.kozmelj@ltfe.org

Upgrade of learning environment E-CHO

Abstract. *Laboratory of telecommunications at the Faculty of Electrical Engineering, University of Ljubljana, developed second generation of virtual learning environment E-CHO 7 years ago. It was based on widely adopted technologies of that time, such as PHP and Javascript. Recent changes in the learning paradigm, introducing myriad of improved functionalities, such as collaboration, social learning, and multimedia into state-of-the-art learning environments as well as technological limitations of the existing E-CHO system (e.g. termination of support for PHP 5.6, taking place at the end of 2018) were the main reasons behind substantial upgrade onto E-CHO+ platform, that will be presented in this article and commented upon.*

1. Uvod

Selitev izobraževalnega procesa iz klasičnih oblik v sodobne tehnološko podprte oblike je že nekaj let realnost, tako v izobraževalnih in akademskih, kot tudi v poslovnih okoljih. Takšna izobraževanja imenujemo e-izobraževanja. Eden izmed pomembnih razlogov za razvoj e-izobraževanj je možnost strukturiranega vključevanja tako obstoječih elektronskih gradiv (npr. dokumenti, slike, kvizi, ankete in preizkusi znanja), kot novejših, bolj interaktivnih in multimedijsko podprtih gradiv (video posnetki, video konference, webinarji in interaktivne vsebine).V okviru e-izobraževanj se je razvilo veliko novih vrst e-gradiv, ki v klasičnih oblikah izobraževanj niso obstajale. Sodobno e-izobraževalno okolje mora tako vključevati vse oblike gradiv klasičnega izobraževanja, kot so dokumenti, slike, kvizi, ankete in preizkusi znanja, ter modernega izobraževanja, kot so video posnetki, video konference, webinarji in interaktivne vsebine.

Pri izobraževanju je pomembna komunikacija med učiteljem in učencem in komunikacija ter sodelovanje med učenci, zato sodobna okolja za e-izobraževanje vključujejo tudi tehnologije, s katerimi se lahko učenci in učitelji med seboj sporazumevajo in tako še nadgradijo svoje znanje.

Za učence in učitelje je zelo pomembna tudi povratna informacija, zato morajo e-izobraževalna okolja vsebovati tudi povratno informacijo o lastnemu uspehu za učence in poročila o napredku za učitelje. Le z učinkovitim sistemom povratnih informacij in poročil lahko okolje za e-izobraževanje doseže odlično uporabniško izkušnjo.

Od leta 2011, ko je nastala prva različica izobraževalnega okolja E-CHO druge generacije, se je v svetu e-izobraževanja zamenjalo kar nekaj trendov. E-CHO se je z novimi različicami postopno prilagajal spremembam, kot so uvajanje video posnetkov, preizkusov znanja in interaktivnih vsebin. V okolju E-CHO je bil vključena tudi sinhronizacija video predavanj s pripadajočimi prosojnicami. S koncem leta 2018 preneha podpora za PHP 5.6, kar pomeni, da bi morali E-CHO nadgraditi na novo verzijo jezika PHP. Ker je okolje E-CHO vseboval še nekaj ne najbolj sodobnih tehnologij, je bila sprejeta odločitev za nadgradnjo obstoječega izobraževalnega okolja. Novo izobraževalno okolje E-CHO+ temelji na odprtokodni rešitvi Meteor, ki jo je razvilo slovensko podjetje Media interactive v sodelovanju z Laboratorijem za telekomunikacije. Prehod na novo okolje prinaša veliko prednosti in izboljšav, ki so opisane v nadaljevanju.

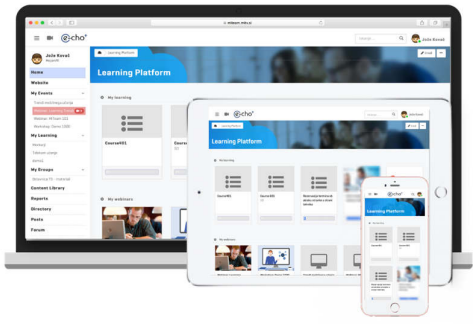
2. Prilagajanje vsem vrstam zaslonov

V letu 2014 je imelo 95.5% ljudi na svetu mobilni telefon, od tega je 30% bilo pametnih telefonov [1]. V zadnjih štirih letih je število pametnih telefonov strmo naraslo in danes ima skoraj vsak posameznik svoj pametni mobilni telefon. Zato je pomembno, da ima e-izobraževalno okolje možnost prilagajanja vsem vrstam mobilnih naprav. Velika pomanjkljivost okolja E-CHO je slabo prilagajanje mobilnim napravam. Novo okolje E-CHO+ se prilagaja vsem vrstam mobilnih naprav, od mobilnih telefonov in tablic do velikih zaslonov namiznih računalnikov.

2.1. Prilagajanje velikosti zaslonov

Pri prilagajanju različnim vrstam mobilnih naprav imamo najprej v mislih predvsem prilagajanje velikostim zaslonov. Vsebina na namiznih računalnikih mora biti drugače prikazana kot tista na mobilnih napravah. V letih pred prihodom mobilnih naprav je v razvoju spletnih platform prevladoval način oblikovanja, kjer so bili vsi gradniki fiksne velikosti, ne glede na velikost zaslona naprave, na kateri se je vsebina ogledovala. Z rastjo uporabe mobilnih telefonov se je povečala tudi težnja po prilagajanju različnim velikostim zaslonov. Nov način oblikovanja velikost posameznih elementov prilagaja velikosti zaslona. Razvoj standardov HTML5 in CSS3 je

ta način postavil v ospredje, saj je za spletne platforme pomembno, da je vsebina ustrezno prikazana na vseh napravah.

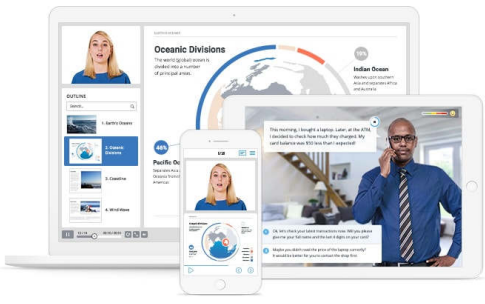


Slika 1: Prilagajanje okolja E-CHO+ različnim velikostim zaslonov

Ker se v svetu uporaba mobilnih telefonov veča, se povečuje tudi njihova uporaba v izobraževalne namene. Pojavilo se je tako imenovano e-učenje na mobilnih napravah, ki se pogosto imenuje tudi mobilno učenje [1]. Zavedati se je potrebno, da vsebina ne more biti prikazana na enak način na vseh napravah zato je pri izobraževalnih okoljih pomembno, da se tudi velikost in postavitev izobraževalnih vsebin prilagajata različnim velikostim zaslonov.

Okolje E-CHO+ že vključuje osnovno prilagajanje vsebin, kot so slike, datoteke, video posnetke in ostale multimedijske vsebine. Za potrebe prilagajanja interaktivnih gradnikov kot so kvizi in preverjanja znanja je v okolju E-CHO+ za razvoj takih vsebin priporočljiva uporaba orodja iSpring [2]. Orodje omogoča izdelavo kvizov in preizkusov znanja, ki se prilagajajo različnim zaslonom.

Za mobilno učenje je pomembna tudi kakovost vsebine, ki jo uporabnik gleda. Boljša kot je kvaliteta, večja je tudi velikost datoteke, ki se mora prikazati. Zato okolje samodejno prilagodi kvaliteto multimedijskih vsebin in z neprekinjenim predvajanje izboljša uporabniško izkušnjo.



Slika 2: Primer prikaza vsebine, narejene z orodjem iSpring, na različnih napravah

3. Multimedijska knjižnica

Za potrebe predvajanja video posnetkov, ki postopno pridobivajo vse večjo vlogo v e-izobraževanju okolje E-CHO+ vsebuje multimedijsko knjižnico. Za predvajanje video posnetkov in prenos v živo okolje E-CHO+ uporablja tehnologijo Wowza [3]. Okolje podpira dinamično prilagajanje strujanja preko http (ang. Dynamic Adaptive Streaming over HTTP) ali MPEG-DASH [4], ki omogoča tekoč in nemoten ogled posnetkov tudi v primeru slabših povezav. Okolje samo prilagodi kvaliteto videa glede na zmogljivosti povezav, ki je na voljo. Da tehnologija MPEG-DASH deluje, je potrebno izvorni video pretvoriti v nižje kvalitete. To v okolju E-CHO+ ob naložitvi videa avtomatsko opravi odprto kodno rešitev FFmpeg [5] Video posnetek se pretvori v 6 različnih profilov (1080p, 720p, 480p, 360p, 240p in 144p).



Slika 3: Tehnologije, ki delujejo v ozadju multimedijske knjižnice v okolju Echo+

Video posnetki predstavljajo osnovni gradnik e-izobraževanj. Zato okolje E-CHO+ vključuje ogled videa na zahtevo in prenos v živo.

3.1. Video na zahtevo

V multimedijski knjižnici so na voljo video posnetki, ki si jih je mogoče ogledati kadarkoli in na kateri koli napravi. Ob naložitvi na E-CHO++ se s pomočjo FFmpeg izvorni video posnetek pretvori v posnetek nižje kvalitete. Okolje nato uporabniku samostojno prilagodi kvaliteto videa glede na povezavo, ki je na voljo. To omogoča boljšo uporabniško izkušnjo ne glede na povezavo, ki jo ima uporabnik v določenem trenutku.

3.2. Prenos v živo

Okolje uporablja Applov standard Http Live Streaming (HLS) za oddajanje videov. Podprt je na iOS, Android in Flash okoljih [6]. Standard je podprt v večini programov za obdelavo in strujanje video in avdio vsebin, kot sta Wirecast in Open Broadcast Software. Po zaključku prenosa v živo, se video naloži v multimedijsko knjižnico in je na voljo kot video na zahtevo.

4. Komunikacijsko okolje

Komunikacijska orodja, kot so Skype, Facebook in podobne rešitve omogočajo komunikacijo na vsakem koraku. Okolje E-CHO+ podpira Web Real-Time Communication (WebRTC) tehnologijo, ki omogoča komunikacijo med spletnimi brskalniki v realnem času.

Real-Time Communication (RTC), ki vključuje zvočno in video komunikacijo, ima nekaj prednosti, vendar ima predvsem zaradi dragih avdio in video licenc, številne pomanjkljivosti, ki so privedle razvijalce do novega standarda [7]. The Worlde Wide Web Consortion (W3C) in Internet Engineering Task Force (IETF) sta zato razvila nov standard, ki se imenuje WebRTC. Standard zagotavlja, da zvočni in video klici delujejo brez dodatnih vtičnikov. Kmalu po standardizaciji so ga v svoje brskalnike vključili vsi brskalniki razen Applovega Safarija, ki je standard vključil šele z različico 11 v sredini leta 2017 [8].
Na standardu WebRTC v okolju E-CHO+ temeljijo tudi videokonference in webinarji, ki postajajo vedno pogosteje uporabljeni v svetu e-izobraževanj.



Slika 4: WebRTC

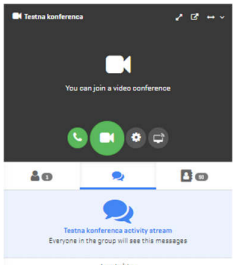
4.1. Webinar

Webinar je namenjen večjemu številu udeležencev, pri čemer je eden (ali več) pobudnik webinarja in prevzame aktivno vlogo v webinarju. Ostali udeleženci samo spremljajo webinar in lahko v klepetu postavljajo vprašanja.

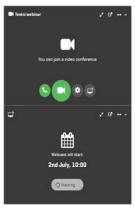
4.2. Videokonference

Videokonference predstavljajo sestanke na daljavo in so posebej priročne, kadar vsi udeleženci ne morejo biti fizično prisotni na sestanku. V okolju E-CHO+ je poleg videokonferenc na voljo tudi navaden zvočni klic, saj je potrebno upoštevati tudi tiste uporabnike, ki nimajo na voljo kamere in bi vseeno želeli aktivno sodelovati v razpravi. Okolje podpira tudi deljenje zaslona z ostalimi uporabniki, ki omogoča predstavitev vsebine na daljavo.

Pri videokonferencah v okolju E-CHO+ lahko aktivno sodeluje do 10 ljudi. To je omejitev konferenčnega strežnika, ki lahko hkrati izdajaja samo 10 konferenčnih licenc. Če je število udeležencev večje okolje E-CHO+ omogoča kombinacijo videokonferenc in webinarjev, kjer v videokonferenci aktivno sodelujejo predavatelj(i), ostali udeleženci pa so v večji meri pasivno prisotni na webinarju.



Slika 5: Videokonfenčni del okolja Echo+



Slika 6: Kombinacija videokonference in webinarja

5. Poročila

Za dobro uporabniško izkušnjo so v e-izobraževalnem okolju potrebna čim bolj pregledna poročila o delu napredku učencev. Poročila so pomembna za učitelje, ki lahko tako spremljajo napredek svojih učencev, in učence, ki lahko spremljajo svoj napredek.

5.1. Experience API

Okolje za beleženje napredka uporablja standard Experience API (ali xAPI). Za razliko od starejših oblik beleženja (npr. SCORM) ima xAPI precej manj omejitev in vključuje beleženje različnih aktivnosti, kot so mobilno učenje, učenje brez povezave in sodelovalno učenje [9].



Slika 7: Koncept delovanja xAPI-ja

5.2. Statements in Learning Record System

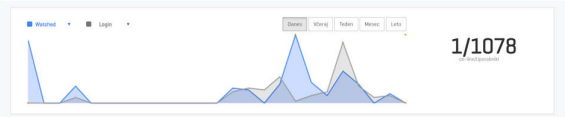
Stanja (ang. Statements) so zapisi o aktivnostih, ki jih opravi uporabnik in so del beleženja napredka. Ko se zgodi neka aktivnost, se stanje zapiše v Learning Record System (LRS). LRS je sistem za shranjevanje učnih rezultatov, ki so vezane na posamezne učne vsebine, in je del vseh okolij za e-izobraževanje. Vloga poročil je prikazati oziroma vizualizirati stanja, ki so shranjena v LRS.
Beleži se lahko stanja, ki so vezana na vsebino, komunikacijo ali uporabnika samega. V okolju E-CHO+ tako beležimo naslednja stanja: ogledano, prijava, odjava, opravljeno, ni opravljeno, v delu, zaključeno, potrjeno in zavrnjeno.



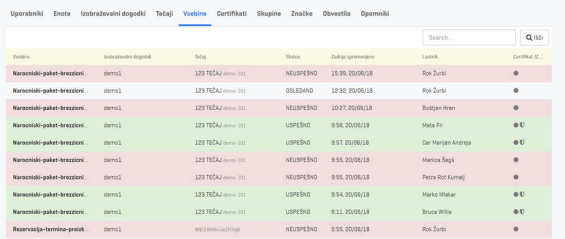
Slika 8: Stanja v okolju Echo+

5.3. Prikaz stanj

V okolju E-CHO+ se stanja prikazujejo na dva načina: grafično in tabelarično. V grafičnem prikazu okolje različna stanja prikaže na časovni osi, medtem ko pri tabelaričnem vsako stanje podrobno predstavimo. Stanja lahko filtriramo po uporabnikih ali vsebinah.



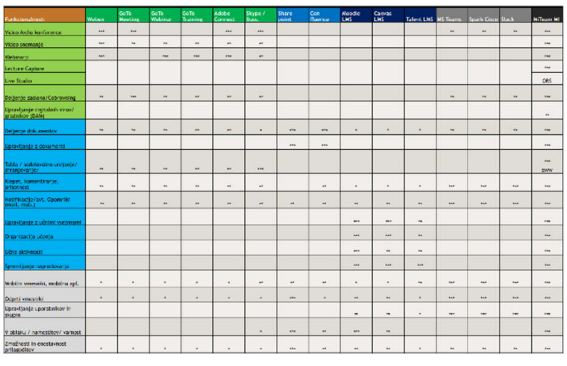
Slika 9: Grafičen prikaz stanj v okolju Echo+



Slika 10: Tabelaričen prikaz stanj v okolju Echo+

6. Primerjava

Okolje E-CHO+, ki temelji na platformi MiTeam, ki jo je razvilo slovensko podjetje Media Interactive, je e-izobraževalno okolje, ki vsebuje tudi lastnosti sodelovalnih okolij, kot so Webex, GoTo Meeting, Adobe Connect ali Skype for Business. Glavne lastnosti sodelovalnih okolij, ki jih vsebuje tudi okolje E-CHO+ so video in avdio konference, webinarje, deljenje zaslonov, snemanje predavanj (ang. lecture capture). Na področju deljenja dokumentov in upravljanja z njimi se lahko okolje E-CHO+ primerja z okolji Sharepoint in Confluence. Glavni del okolja E-CHO+ pa je namenjen izobraževanju, kjer glavno konkurenco predstavljajo okolja Moodle, Talent in Canvas. Na spodnji sliki je prikazana primerjava omenjenih okolij.



Slika 11: primerjava okolij

7. Zaključek

Okolje E-CHO+ je celovito e-izobraževalno okolje, ki vključuje vse potrebne funkcionalnosti za izvedbo izobraževanja z uporabo informacijsko komunikacijsko tehnologijo. Okolje je prilagojeno interaktivnim vsebinam, ki postajajo čedalje bolj pomemben gradnik e-izobraževanja. Interaktivne vsebine predstavljajo preplet vseh vsebin, ki so do sedaj sestavljale izobraževalne sisteme.

Ker uporabniki vse več uporabljajo mobilne naprave, okolje E-CHO+ spretno prilagaja velikost elementov in vsebin velikosti zaslonov. Pri tem upošteva tudi hitrost povezave, ki je na voljo, in po potrebi zmanjša kvaliteto vsebin, da zagotovi boljšo uporabniško izkušnjo.

Video vsebine danes predstavljajo najpomembnejši del vsebin pri izvedbi e-izobraževanj in predvsem pri izobraževanju na daljavo. Okolje E-CHO+ podpira video na zahtevo in prenos v živo. S tehnologijo dinamičnega prilagajanja kvalitete izboljša uporabniško izkušnjo, saj se video posnetek, v primeru manjše hitrosti internetne povezave, ne ustavi, ampak se zgolj zmanjša njegova kakovost.

Za izvedbo izobraževanja na daljavo izobraževalno okolje potrebuje tudi komunikacijsko orodje. Okolje E-CHO+ vsebuje tehnologijo WebRTC, ki omogoča klepet in video ter avdio klice.

Kar daje okolju E-CHO+ pomembno dodano vrednost, so poročila. Okolje s tehnologijo Experience API beleži vse aktivnosti, ki se dogajajo na platformi in jih vizualizira na grafičen ali tabelaričen način.

Literatura

[1] V. Bhuttoo, K. Soman, in R. K. Sungkur, „Responsive design and content adaptation for e-learning on mobile devices“, v *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, 2017, str. 163–168.

[2] iSpring, https://www.ispringsolutions.com/

[3] Wowza Live Streaming Software, https://www.wowza.com/

[4] H. Kwan, C. Roy, in D. Das, „Demonstration of a multimedia player supporting the MPEG-DASH protocol“, v *2012 Visual Communications and Image Processing*, 2012, str. 1–1.

[5] FFmpeg, https://www.ffmpeg.org/

[6] M. Kozmelj, N. Pušnik, in K. Pečnik, „Tehnologije za prenos v živo in ogled videa na zahtevo“, 2017

[7] Cristian Cola in Honoriu Valean, „On multi-user web conference using WebRTC“, v *Proceedings of the 18th International Conference on System Theory, Control and Computing*, Sinaia, Romania, 2014, str. 430–433.

[8] WebRTC support in Safari 11, https://webrtc.ventures/2017/06/webrtc-support-in-safari-11/

[9] *Experience API,* https://xapi.com/overview/

Avtomatika in robotika

Automatic Control and Robotics

Comparative analysis of MES and ERP systems

Zerina Hasić[1], Amila Dubravić[2]

[1]Tuzla, Bosnia and Herzegovina
[2] Faculty of Electrical Engineering University of Tuzla, Tuzla, Bosnia and Herzegovina
E-mail: zerinahasic92@gmail.com, amila.dubravic@untz.ba

Comparative analysis of MES and ERP systems

Abstract. *The paper provides a theoretical and practical overview of the basic parameters when transforming the classical factory from the production plant into a modern service center. This means that the decisive potentials of companies are not so much in their productive capacities as in their process capabilities.*

The process of implementation and testing of ERP and MES software is described, which will be used for managing the production processes, collecting and processing data, where the advantages of the MES system in relation to the ERP system can be determined.

1 Introduction

Increased competitive pressures and high quality demands force companies to take advantage of all possible ways to increase their economic efficiency in the industry.

In order to achieve efficient value creation in production, equipment that can meet the new requirements in general is needed. Existing Enterprise Resource Planning (ERP) systems are mainly administrative and accounting systems. The new systems must include planning, logging and control functions that operate in real time. This concept is the Manufacturing Execution Systems (MES). Integrating the right software across the company is key to increasing efficiency and gaining business insight, essential to securing competitive advantage and maximizing profitability. Software applications that manage one business aspect provide partial information that can lead to separation of the organization. Companies with such systems often lack the flow of information they expected to achieve. For these reasons, it is necessary to use appropriate systems, such as MES system which is one of the most important current resources for achieving economic efficiency in a digital factory.

2 Defining the conditions - Levels, concepts and technologies

The world of IT, especially the subject of production systems, is marked with special terms. An overview of the terms in Fig. 1 show the classification of concepts at the level of Instrumentation, Systems and Automation Society – ISA (Figure 1).

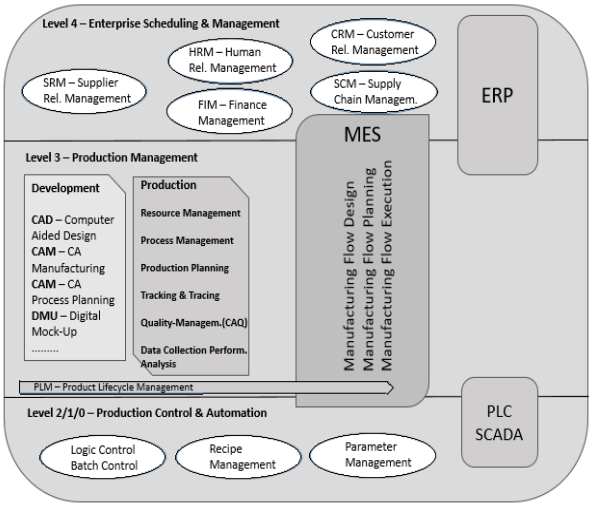


Figure 1 Overview of terms according to ISA level model [1]

Figure 2 is a horizontally illustrated product life cycle, from idea to implementation – Product Lifecycle Management (PLM), and the workflow of MES is displayed vertically. It can be noticed that there is an overlap between both systems at the production stage, which can be solved through a parallel and harmonized implementation of the system (Figure 2).

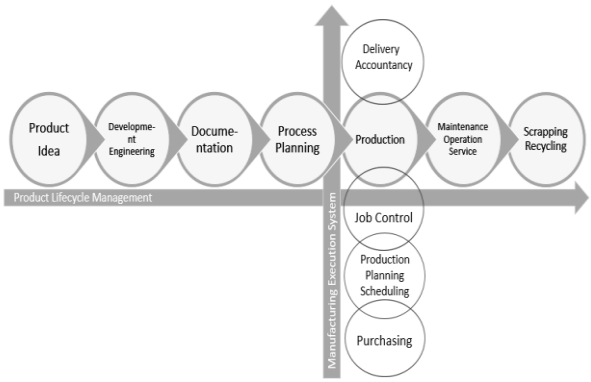


Figure 2 Product Lifecycle Management Concept [1]

ERP system is an information system that supports business operations and is used in almost all organizational parts of a company. One of the world's leading ERP systems is the SAP system. ERP is used to create and manage the basic layout of the plant, including manufacturing, use of materials and delivery, which

makes the information the center of the operation. ERP focuses on work within a time period such as months, weeks and days. MES works within time frames such as dates, shifts, hours, and minutes.

3 Construction of MES systems as a new class of IT applications

3.1 Software architecture and interfaces of MES systems

The basic functions of a MES system are a collection of software functions that enable the development of products using unified structure modules.

The MES system is divided into three layers of system architecture (Figure 3).

The data layer is part of the MES application responsible for defining the database structure, as well as for the data stored therein.

The application layer depends on the data layer and functionally makes available the process mapping.

The process mapping has the task of recreating the actual business logic, based on the methods and objects of the application layer. The process mapping receives data over a graphical interface that is processed using the if-then conditions and using the objects and methods of the application layer.

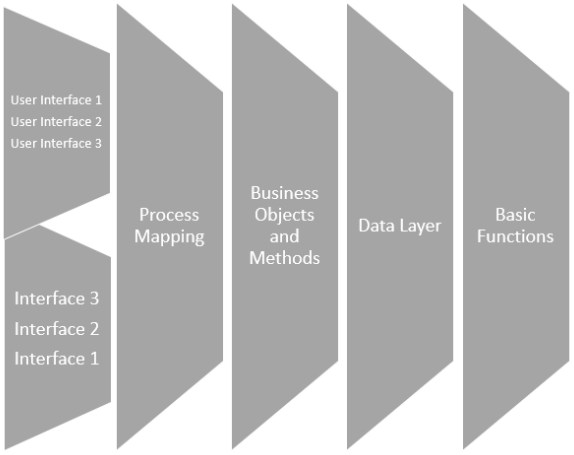


Figure 3 The software architecture of a MES system [2]

The rule is that the product is changed or expanded only in one layer, which ensures stability and reduces changes. [2]

For the MES system it is important to maintain some of its interfaces open, especially interfaces and functions related to product integration. User interface has the user-friendly interface flexibility, fast and efficient data acquisition and real-time display.

The database (Figure 4) is a central part and, from a technical point of view, the most important element of the MES. [1] Migrating to another database system is very difficult and it is difficult to perform when the system starts up. It is therefore important to select the appropriate database before the system is implemented.

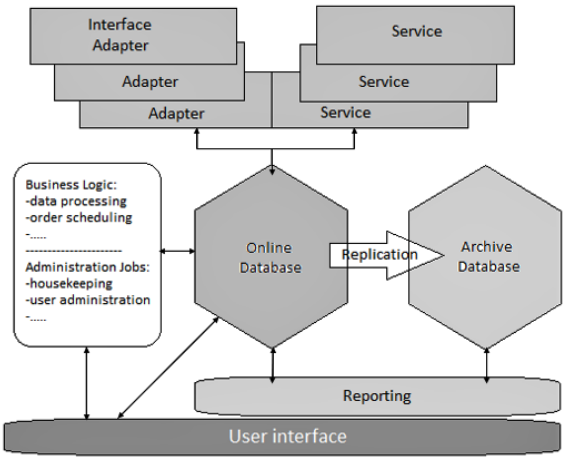


Figure 4 Central software components as an overview [1]

Software database architecture has two approaches:

- Application-oriented systems: The application controls the reservation functions in the database and the business logic of the system.
- Systems based on the database: With this approach, the database is not only the data memory but also the pivot of the whole system. [1]

3.2 Preparation for the implementation of the MES system in production

In the context of the MES implementation project, it is important to create the most realistic model of a future production organization.

If new software is needed, then the question arises as to whether to purchase standard software or to develop customized software. Software should be evaluated based on criteria: functionality, quality, performance, documentation, technology (Table 1.).

Table 1 Advantages and Disadvantages of Standard and Customized software [1]

Standard Software	Customized Software
Is often less expensive than customized software.	Software is tailored precisely to the needs of the organization.
Software has already proved itself in use.	It is likely that no interface problems will arise.
Functions that are not required must be purchased with the software.	Only the required functions are paid for.
Interface problems could potentially arise.	Quality of the software is reduced by an inexperienced development team and time pressure.
Extensive documentation generally is provided.	Documentation is often neglected.

Implementation can be performed on the basis of three strategies: "Big Bang", step by step in individual areas of work, step by step replacing individual business processes. Implementation of the manufacturing execution systems (MES) is a complex project and requires strong project management.

The main team should perform the following activities in the implementation project:

• Project management throughout the implementation period
• Achieving a fundamental decision: "MES: yes or no" (Table 2)
• Appointment of project manager and project team for actual implementation
• Securing funding
• "Release" of employees to perform duties during implementation
• Monitoring the progress of the project, including costs and monitoring of the deadline
• Review and track success after implementation
• Creating a work concept for sustainable use of the system. [1]

Table 2 Decision matrix for the fundamental decision: "Is a MES necessary?" [1]

Criteria / Importance:	Low	Medium	High
Number of process operations	<5	5–10	>10
Number of products	<10	10–50	>50
Number of product versions	0	1–10	>10
Added value in production	Low	Medium	High
MES Required?	No	Recomme-nded	Yes

4 Comparative analysis of MES and ERP systems implemented on the open source platform

Two SW packages are going to be analyzed and compared, the Qcadoo MES software compared to Dolibarr ERP CRM software based on the open source platform.

In the case of MES implementation, using qcadoo software the possibilities of more detailed analysis are limited due to the complexity of the program. For more detailed analysis, it is necessary to buy software commercial versions. For this purpose, comparison can be made with Dolibarr ERP CRM software, part of the analysis related to the administrative and accounting systems for which the ERP software is predicted.

Qcadoo MES is an open source project designed in Poland, which is intended for production management in small and medium-sized enterprises. It is a software that aims to create an easy and modular IT solution for the manufacturing industry and allows you to calculate costs and plan production in real time.

The features of qcadoo MES software are:

-User interface is easy to understand and use for both professionals and workers in production;

-Online Availability: all that is needed to use qcadoo MES is an internet browser and internet connection.

Qcadoo MES comes in two different versions:

• Community version - Open source version that serves as a trial version under the GNU AGPLv3 license.
• Commercial version - version developed and sold by Qcadoo Limited company, which provides additional analysis functions and full support for production management.

After qcadoo is installed, registration is followed, and opening the application will give a home screen display and data entry screen for factory module (Figure 5, Figure 6 and Figure 7).

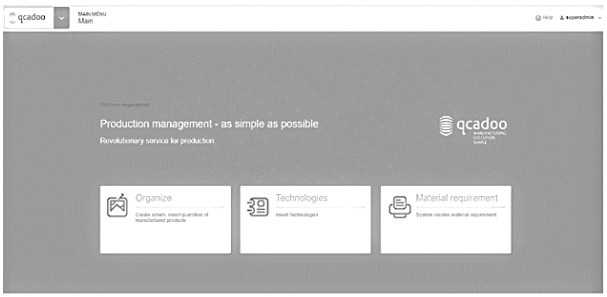


Figure 5 View the start screen of qcadoo MES [3] [4]

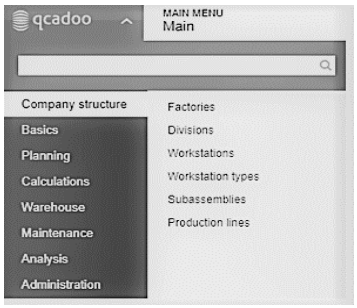


Figure 6 Qcadoo MES module and submodule [3] [4]

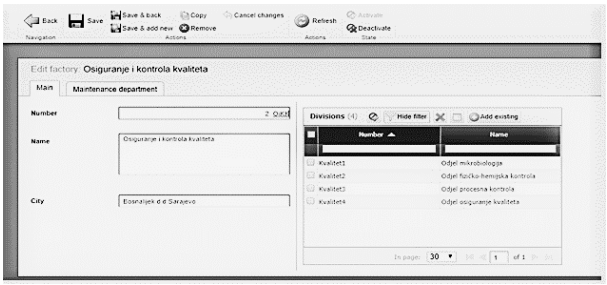


Figure 7 Display the data entry screen for the qcadoo MES factory module [3] [4]

The modules in the application are as follows: Company structure, Basics, Planning, Calculations, Warehouse, Maintenance, Analysis, Administration (Figure 6). Each module has submodules for data entry

or options for executing planning, calculations and analyzes based on the entered data. In the submodule Factories, it is possible to enter the parts/sectors of the factory or the company, company departments, workstation production lines and types of work stations, subsystems and production lines. In the representation of each submodule and the entered data, the pharmaceutical and chemical industries are shown as an example.

Dolibarr ERP CRM is an open source, free software package for small and medium enterprises, designed in France. Includes various Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM).

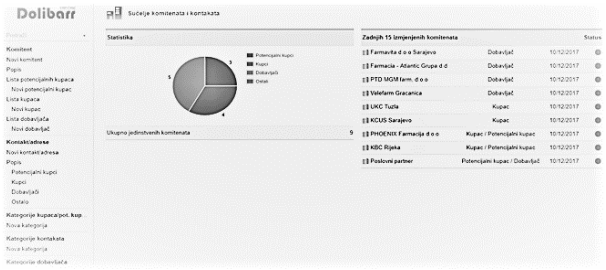


Figure 8 Display of the Dolibarr ERP CRM screen of analysis and planning modules [3] [5] [6]

There are several functional modules that can be enabled or disabled, as needed. It represents a web-based application and aims to offer free open source ERP and CRM functions, providing a simple solution (Figure 8).

Dolibarr is written in PHP and uses MySQL databases. It is possible to install an application on a smartphone, to track the company's activities remotely, and the ability to save data through the cloud service.

According to the comparative analysis, the following conclusions can be made: ERP knows "why", while MES knows "how". ERP is used to create and manage the basic layout of the plant, while MES is used to guide the production operation, as an event in real time. ERP and MES are overlapping in some areas, but each ERP could be supported by MES (Table 3).

Table 3 Functions of analyzed MES and ERP systems

Qcadoo MES	Dolibarr ERP CRM
Open source software for small and medium enterprises.	Open source, free software package for small and medium enterprises.
User interface is easy to understand and use.	Several functional modules that can be enabled or disabled.
Online Availability.	Providing a simple solution.
Community version and commercial version.	Track the company's activities remotely.
Modules for executing planning, calculations and analyzes.	Save data through the cloud service.
MES knows "how".	ERP knows "why".

There is an equal possibility of using both software by the employees themselves, not just administrators, from which it can be concluded that both systems are equally useful, where MES represents an innovative solution for all parts of the company.

5 Conclusion

The success of MES software shows examples of implementations in the most eminent and most developed manufacturing companies in the field of automation, robotics and electrical engineering, as well as in pharmaceutical industries. Based on the research conducted, it can be concluded that the companies that are dealing with discrete production or production by order ERP applications can provide rich functionality for managing these types of operations. Manufacturers in these industries should consider whether additional functions of the MES system will be excessive in certain situations. There is no standard list for deciding whether a company needs a MES or ERP system or both systems. The key to successful software implementation is that the company is confident in defining the requirements and options that are important for the business, before the appropriate software is selected. The method of collecting and processing data using MES will be used in the future, which will bring flexibility, transparency and economic efficiency in production systems.

References

[1] H. Meyer, F. Fuchs i K. Thiel, Manufacturing Execution Systems - Optimal Design, Planning and Deployment, McGraw Hill Professional, 2009.

[2] J. Kletti, Manufacturing Execution System – MES, Springer, 2007.

[3] Z. Hasić, Master's thesis - Comparative analysis MES and ERP systems implemented on open source platform, Tuzla, 2018.

[4] »qcadoo MES - friendly SaaS manufacturing management software,« 2017, Poland. [Mrežno]. Available: https://qcadoo.com/en/.

[5] »Dolibarr ERP CRM - Open Source ERP CRM for business,« 2017, France. [Mrežno]. Available: https://www.dolibarr.org/.

[6] »Community where developers can discover, use and contribute to projects,« [Mrežno]. Available: https://github.com.

Izvedba sistemov vodenja na osnovi modelov za avtonomno vožnjo miniaturnih robotskih vozil

Marko Tavčar, Urban Skalar, Jernej Perko, Andrej Zdešar

Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana

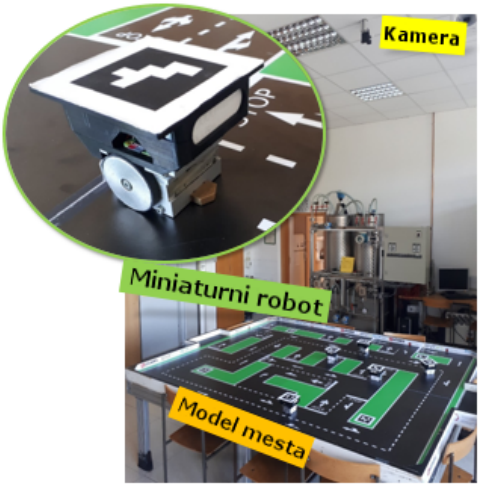
E-pošta: mt6480@student.uni-lj.si, us1673@student.uni-lj.si, jp2352@student.uni-lj.si, andrej.zdesar@fe.uni-lj.si

Implementation of model-based control systems for autonomous driving of miniature robotic vehicles

In this work we focus on implementation of algorithms for localisation, path planning and path tracking on a cyber-physical system of miniature mobile robots. We implemented Kalman filter to combine the data from two different sensor sources. An A algorithm was used to find the optimal path in the map. Robot was controlled in such a manner that it is capable of following the reference path, which leads to the desired goal.*

I. Uvod

V svetu zanimanje za avtonomne mobilne sisteme narašča. V tem delu predstavljamo, kako smo implementirali algoritme, ki so ključni za avtonomno delovanje robotskih vozil. V ta namen smo uporabili poenostavljen fizični model mesta z miniaturnimi mobilnimi roboti, ki je prikazan na sliki 1. Na fizičnem modelu smo implementirali algoritme za lokalizacijo, načrtovanje poti in vodenje po poti [1] tako, da se mobilni robot lahko avtonomno vozi po cestah med poljubnimi cilji in pri tem vozi varno ter upošteva cestno-prometne predpise.



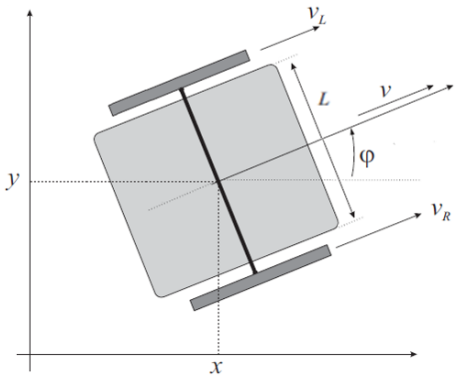
Slika 1: Fizični model mesta s cestami in miniaturni mobilni roboti z značkami ter kamera

Delo, predstavljeno v tem članku, je bilo opravljeno v okviru projekta InMotion, ki ga sofinancira program Evropske unije Erasmus+, št. 573751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP.

II. Predstavitev sistema

A. Kolesni mobilni robot

V našem sistemu smo uporabili 5 kolesnih mobilnih robotov z diferencialnim pogonom. Na vsakem mobilnem robotu se nahaja mini računalnik *Raspberry Pi Zero W*, ki omogoča izvedbo algoritmov za avtonomno vožnjo, in mikrokrmilnk *Arduino Mini*, ki se uporablja za krmiljenje motorjev in branje podatkov iz enkoderjev. Maksimalna hitrost gibanja robota znaša $0{,}3\,\mathrm{m/s}$. Na vsakem motorju kolesa ima robot nameščen en enkoder za merjenje zasuka kolesa s frekvenco vzorčenja $100\,\mathrm{Hz}$.



Slika 2: Prikaz mobilnega robota od zgoraj z označenimi osnovnimi vektorji gibanja

Model kolesnega mobilnega sistema z diferencialnim pogonom je prikazan na sliki 2. Vpliv vhodne tangencialne hitrosti $v(t)$ in kotne hitrosti $\omega(t)$ na spremembo lege mobilnega sistema podaja kinematični model (1).

$$\begin{aligned} \dot{x}(t) &= v(t)\cos\varphi(t) \\ \dot{y}(t) &= v(t)\sin\varphi(t) \\ \dot{\varphi}(t) &= \omega(t) \end{aligned} \tag{1}$$

B. Kamera

Nad poligonom se nahaja kamera tako, da je celotni poligon v njenem vidnem polju. Kamera preko WiFi-povezave sporoča vsem sistemom v lokalnem omrežju izmerjene lege robotov na poligonu glede na globalni koordinatni sistem, ki se nahaja v enem izmed robov poligona. Sistem za vizualno sledenje značk omogoča obdelavo slike s frekvenco $10\,\mathrm{Hz}$.

III. Lokalizacija mobilnega robota

Za merjenje lege $\boldsymbol{x}(k) = [x(k)\ y(k)\ \varphi(k)]^T$ mobilnega robota imamo na voljo dva nepovezana senzorska vira: enkoderje na kolesih in kamero. Na podlagi enkoderjev na kolesih lahko določimo lego mobilnega sistema s postopkom odometrije. Merilni sistem s kamero podaja lego mobilnega robota glede na globalni koordinatni sistem, in v tem smislu deluje kot globalni satelitski sistem za določanje položaja.

A. Odometrija

Kinematični model (1) lahko z Eulerjevo integracijsko metodo pretvorimo v diskretno obliko (2). Za kratke čase vzorčenja T je napaka, ki jo dobimo zaradi integracijske metode zanemarljiva glede na ostale vire napak (npr. zdrsovanje koles).

$$\boldsymbol{x}(k+1) = \boldsymbol{x}(k) + \begin{bmatrix} \cos\varphi(k) & 0 \\ \sin\varphi(k) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v(k)T \\ \omega(k)T \end{bmatrix} \tag{2}$$

Vpeljemo relativni tangencialni premik $\Delta d(k) = v(k)T$ in relativni kotni zasuk $\Delta\varphi(k) = \omega(k)T$ mobilnega robota. Relaciji med relativnimi premiki mobilnega sistema in relativnimi premiki levega in desnega kolesa, $\Delta d_L(k)$ in $\Delta d_D(k)$, sta:

$$\begin{aligned} \Delta d(k) &= \frac{\Delta d_D(k) + \Delta d_L(k)}{2}\,, \\ \Delta\varphi(k) &= \frac{\Delta d_D(k) - \Delta d_L(k)}{L}\,, \end{aligned} \tag{3}$$

kjer je L razdalja med kolesoma. Relativne premike koles lahko merimo z relativnimi spremembami vrednosti enkoderjev na kolesih ($\Delta\alpha_D(k)$ in $\Delta\alpha_L(k)$):

$$\boldsymbol{u}(k) = \begin{bmatrix} \Delta d_L(k) \\ \Delta d_D(k) \end{bmatrix} = \begin{bmatrix} K_L\Delta\alpha_L(k) \\ K_D\Delta\alpha_D(k) \end{bmatrix} \tag{4}$$

Konstante $K_L = K_D = 9{,}537\,\mu\mathrm{m}$ in $L = 0{,}075\,\mathrm{m}$ smo določili s postopkom kalibracije. Po vstavitvi (3) in (4) v (2), dobimo model, ki omogoča spremljanje lege mobilnega sistema na podlagi merjenja relativnih premikov koles z enkoderji.

B. Nedeterministični model kolesnega robota

Za združevanja informacije o legi na podlagi odometrije in kamere smo uporabili razširjeni Kalmanov filter, ki je pogosta izbira pri ocenjevanju lege v mobilni robotiki [2]. Za ta namen smo definirali nedeterministični model mobilnega robota, ki je predstavljen v (5) in (6).

Različne vplive motenj, ki nastopajo pri merjenju položaja na podlagi odometrije, smo v modelu prehajanja stanj (5) modelirali z normalno porazdeljenim šumom $\boldsymbol{w}$ s srednjo vrednostjo nič in varianco $\boldsymbol{Q}$.

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{f}(\boldsymbol{x}(k), \boldsymbol{u}(k), \boldsymbol{w}(k)) = \\ &= \boldsymbol{x}(k) + \begin{bmatrix} \frac{\cos\varphi(k)}{2} & \frac{\cos\varphi(k)}{2} \\ \frac{\sin\varphi(k)}{2} & \frac{\sin\varphi(k)}{2} \\ -\frac{1}{L} & \frac{1}{L} \end{bmatrix} (\boldsymbol{u}(k) + \boldsymbol{w}(k)) \end{aligned} \tag{5}$$

$$\boldsymbol{w}(k) = \begin{bmatrix} w_L(k) \\ w_D(k) \end{bmatrix}\,, \quad \boldsymbol{Q} = \begin{bmatrix} \sigma_L^2 & \sigma_L\sigma_D \\ \sigma_L\sigma_D & \sigma_D^2 \end{bmatrix}$$

Predpostavljamo, da meritve s kamere $\boldsymbol{z}(k)$ vsebujejo normalno porazdeljen šum $\boldsymbol{v}(k)$ s srednjo vrednostjo nič in varianco $\boldsymbol{R}$, kot to opisuje (6).

$$\boldsymbol{z}(k) = \boldsymbol{h}(\boldsymbol{x}(k), \boldsymbol{v}(k)) = \boldsymbol{x}(k) + \boldsymbol{v}(k) \tag{6}$$

$$\boldsymbol{v}(k) = \begin{bmatrix} v_x(k) \\ v_y(k) \\ v_\varphi(k) \end{bmatrix}\,, \quad \boldsymbol{R} = \begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_\varphi^2 \end{bmatrix}$$

C. Razširjeni Kalmanov filter

Razširjeni Kalmanov filter smo izpeljali na podlagi nedeterminističnega modela mobilnega robota, ki je predstavljen v poglavju III-B. Algoritem vsebuje predikcijski in korekcijski korak, ki ju izvajamo zaporedno s frekvenco vzorčenja podatkov z odometrije.

V predikcijskem koraku s (7) napovemo trenutno lego mobilnega sistema $\boldsymbol{x}(k)$ na podlagi meritev iz enkoderjev $\boldsymbol{u}(k-1)$ in zadnje ocene lege $\boldsymbol{x}(k-1)$. Z (8) določimo še varianco napovedi.

$$\hat{\boldsymbol{x}}(k) = \boldsymbol{f}(\boldsymbol{x}(k-1), \boldsymbol{u}(k-1), \boldsymbol{0}) \tag{7}$$

$$\hat{\boldsymbol{P}}(k) = \boldsymbol{A}_{k-1}\boldsymbol{P}(k-1)\boldsymbol{A}_{k-1}^T + \boldsymbol{F}_{k-1}\boldsymbol{Q}\boldsymbol{F}_{k-1}^T \tag{8}$$

Če imamo na voljo novo meritev s kamere, izračunamo ojačenje $\boldsymbol{K}_k$:

$$\boldsymbol{K}_k = \hat{\boldsymbol{P}}(k)\boldsymbol{C}_k^T(\boldsymbol{C}_k\hat{\boldsymbol{P}}(k)\boldsymbol{C}_k^T + \boldsymbol{R})^{-1}\,, \tag{9}$$

sicer pa postavimo $\boldsymbol{K}_k = \boldsymbol{0}$. V korekcijskem koraku posodobimo oceno lege mobilnega robota $\boldsymbol{x}(k)$ na podlagi napovedi $\hat{\boldsymbol{x}}(k)$ in ob upoštevanju meritve s kamere $\boldsymbol{z}(k)$, kot to opisuje (10). Z (11) posodobimo tudi varianco ocene.

$$\boldsymbol{x}(k) = \hat{\boldsymbol{x}}(k) + \boldsymbol{K}_k(\boldsymbol{z}(k) - \boldsymbol{h}(\hat{\boldsymbol{x}}(k), \boldsymbol{0})) \tag{10}$$

$$\boldsymbol{P}(k) = \hat{\boldsymbol{P}}(k) - \boldsymbol{K}_k\boldsymbol{C}_k\hat{\boldsymbol{P}}(k) \tag{11}$$

Matrike $\boldsymbol{A}_k$, $\boldsymbol{F}_k$ in $\boldsymbol{C}_k$, ki se pojavljajo v enačbah (7) do (11), določimo na podlagi linearizacije nelinearnega modela mobilnega sistema v okolici ocene v trenutku k:

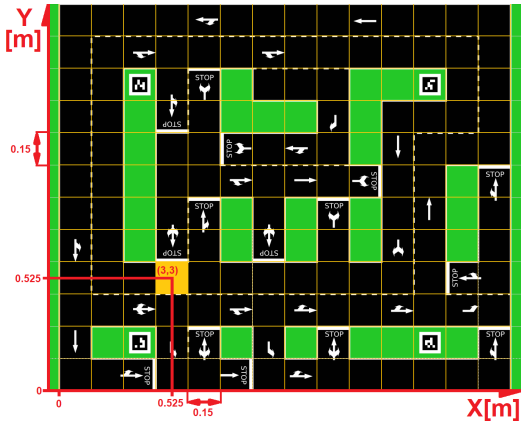
$$\begin{aligned} \boldsymbol{A}_k &= \left.\frac{\partial\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})}{\partial\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}(k),\boldsymbol{u}=\boldsymbol{u}(k),\boldsymbol{w}=\boldsymbol{0}} \\ \boldsymbol{F}_k &= \left.\frac{\partial\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})}{\partial\boldsymbol{w}}\right|_{\boldsymbol{x}=\boldsymbol{x}(k),\boldsymbol{u}=\boldsymbol{u}(k),\boldsymbol{w}=\boldsymbol{0}} \\ \boldsymbol{C}_k &= \left.\frac{\partial\boldsymbol{h}(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}(k),\boldsymbol{v}=\boldsymbol{0}} \end{aligned} \tag{12}$$

IV. Načrtovanje poti

Avtonomni mobilni roboti morajo biti sposobni načrtovanja optimalne poti od točke A do točke B. Pri implementaciji načrtovanja poti so bila upoštevana določena cestno-prometna pravila. Mobilni robot se lahko vozi le po desnem pasu ceste, vožnja po zelenici ni dovoljena. Pri vožnji mora upoštevati tudi stop znake, na križiščih pa so mu dovoljeni U-zavoji.

A. Modeliranje okolja

Da lahko mobilni robot najde optimalno pot, se mora zavedati okolja v katerem se nahaja. Okolje smo z mrežo razdelili na 14×12 celic (slika 3) in vsaki celici smo



Slika 3: Razdelitev poligona na celice z označenimi koordinatnimi sistemi in dimenzijami

priredili unikatno oznako v obliki dveh števil $(i,\ j)$. Preslikavo med položajem v globalnem koordinatnem sistemu in oznako pripadajoče celice podaja (13), kjer operator $\lfloor\cdot\rfloor$ predstavlja celoštevilsko zaokroževanje navzdol.

$$i = \left\lfloor \frac{x}{0{,}15} \right\rfloor, \quad j = \left\lfloor \frac{y}{0{,}15} \right\rfloor \tag{13}$$

Obratna transformacija (14) priredi vsaki celici položaj v globalnem koordinatnem sistemu, ki se nahaja v središču te celice.

$$x = 0{,}15i + 0{,}075\,, \quad y = 0{,}15j + 0{,}075 \tag{14}$$

Slika 3 ilustrira primer transformacije za eno celico.

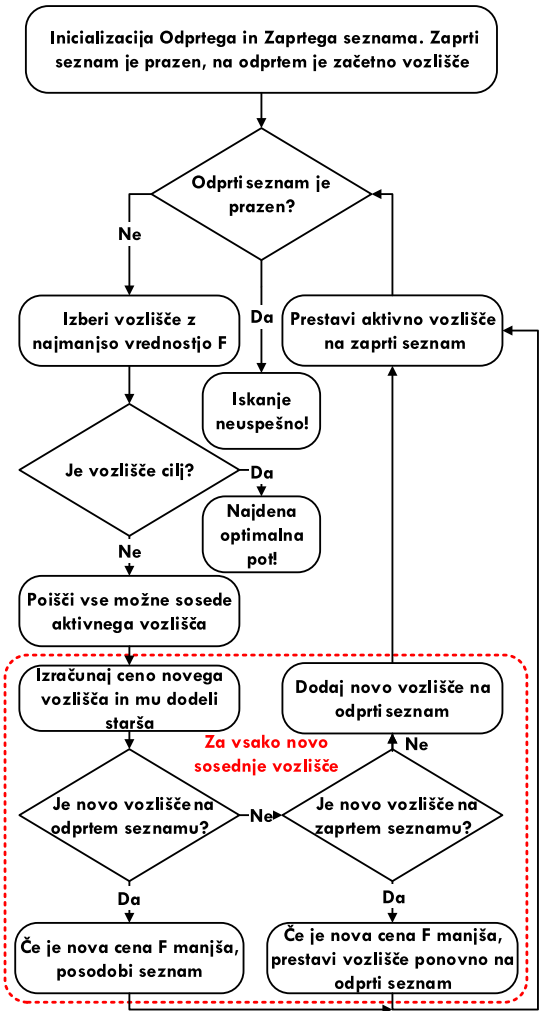
B. Algoritem iskanja poti

Cilj algoritma za iskanje poti je najti optimalno pot od starta do cilja. Pri tem vhod v algoritem predstavljata točki start (začetna lega robota) in cilj (želena lega robota). Izhod algoritma predstavlja urejen niz celic, ki sestavljajo pot od start do cilja. Optimalna pot ni nujno najkrajša pot. V naši implementaciji, algoritem posebej upošteva stop celice. Vsaka stop celica vzame mobilnemu robotu več časa kot pa ostale celice, saj se mora mobilni robot pred stop znakom ustaviti. Posledično imajo takšne celice manjši potencial, da se znajdejo na optimalni poti.

*1) Algoritem A**

V našem sistemu smo implementirali algoritem A* [3], ki je zaradi svoje optimalnosti in hitrosti izračuna, pogosta izbira za iskanje optimalne poti. Z našo implementacijo algoritma smo dosegli zadovoljivo hitro izvajanje algoritma, zato nismo iskali alternativnih algoritmov. A* je informiran algoritem iskanja poti, ki pri svojem delovanju uporablja hevristično oceno *cene do cilja*. S tem se iskanje po vozliščih omeji na celice, ki bolj verjetno sestavljajo optimalno pot, kar posledično poveča hitrost izračuna.

V algoritmu nastopa cena F, ki določa vrstni red odkrivanja celic in je definirana kot vsota dveh cen: $F = G + H$. G predstavlja ceno poti, ki vodi od začetne do trenutno izbrane celice. Določimo jo na podlagi vsote uteženih prehodov med celicami. V našem primeru smo uporabili različne uteži za prehode čez stop znake (utež 2) kot za običajne prehode med celicami (utež 1). Tako postanejo celice s stop znakom slabši kandidati za nadaljevanje poti, saj se mora tam mobilni robot ustaviti. H podaja hevristično oceno razdalje od trenutne celice do cilja. V našem primeru smo uporabili Manhattansko razdaljo.



Slika 4: Diagram izvajanja algoritma A*

Koraki algoritma A* so razvidni na sliki 4. V inicializaciji algoritma definiramo prazen odprti in zaprti seznam ter uvrstimo začetno celico na odprti seznam. Nato vstopimo v neskončno zanko, kjer v vsakem ciklu izvedemo naslednje korake. Iz odprtega seznama odstranimo celico, ki ima najmanjšo vrednost F. Nato poiščemo vse njene sosednje celice in jih uvrstimo na odprti seznam Za vsako novo odkrito celico izračunamo vrednosti G, H in F ter ji dodelimo starševsko (trenutno) celico. En cikel zanke končamo tako, da uvrstimo trenutno celico na zaprti seznam.

Ko po tem postopku pridemo do ciljne celice, lahko algoritem prekinemo, saj to pomeni, da obstaja pot do cilja. Optimalno pot poiščemo tako, da sledimo staršev-skim celicam, dokler ne prispemo do začetne celice. Izvajanje algoritma zaključimo tudi, če je odprti seznam prazen. To pomeni, da pot do želenega cilja ne obstaja.

V. Vodenje mobilnega robota po poti

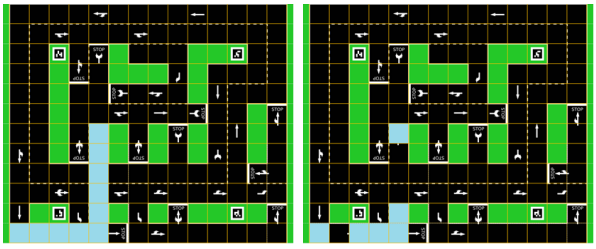
Lego mobilnega robota glede na podane vhodne hitrosti opisuje direktna kinematika (1). Pri načrtovanju vodenja po poti se ukvarjamo s problemom izračunavanja tangencialne in kotne hitrosti, ki sta potrebni, da se mobilni robot vozi po želeni poti. Sistem (1) ima neholonomično omejitev, zato je izračun invezne kinematike v splošnem zahtevna naloga. Preprosta rešitev inverzne kinematike obstaja za primere, ko se mobilni robot vozi le naravnost in obrača na mestu, a takšno gibanje robota običajno ni optimalno. V nadaljevanju je predstavljen princip vodenja mobilnega robota po referenčni poti, ki smo jo dobili s postopkom načrtovanja poti, predstavljenim v poglavju IV-A.

A. Poenostavljen zapis poti

Mobilni robot dobi navodila za pot v obliki urejenega seznama celic, ki jih mora obiskati, da doseže cilj. Ali so vse celice potrebne za popoln opis poti, ki jo želimo opraviti? Sosednje celice v urejenem seznamu celic nimajo dodane vrednosti, če le-te predstavljajo ravno pot. Takšne celice lahko s seznama odstranimo. Obdržati moramo torej le začetno in končno celico ter celice, kjer pride do spremembe smeri. Poleg tega moramo na seznamu obdržati še celice s stop znakom, saj se moramo na teh mestih ustaviti.

Za zaznavanje celic, kjer pride do spremembe smeri poti, smo uporabili naslednji algoritem. Sprehodimo se skozi urejen seznam celic, ki predstavljajo pot. Za vsako celico izvedemo preprost test, tako da primerjamo oznako trenutne celice (oznaka celice je definirana v poglavju IV-A) z oznako celice s seznama, ki je oddaljena za dve mesti. Če se obe vrednosti oznake razlikujeta, pomeni, da je na tem mestu zavoj, zato vse celice, ki tvorijo zavoj, obdržimo na seznamu.

Primer poenostavitve poti je prikazan na sliki 5.



(a) (b)

Slika 5: (a) Pot iz celice $(0,\ 0)$ v celico $(4,\ 5)$ in (b) poenostavljena pot

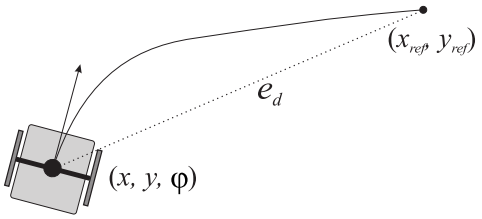
B. Algoritem za vodenje po poti

Oglejmo si algoritem za vodenje do točke. Definiramo pogrešek po razdalji (evklidsko mera) in kotu (slika 6):

$$e_d(k) = \sqrt{(x_{ref} - x(k))^2 + (y_{ref} - y(k))^2}\,, \qquad (15)$$

$$e_\varphi(k) = \arctan \frac{y_{ref} - y(k)}{x_{ref} - x(k)} - \varphi(k) + (n + 2m)\pi\,.$$

Pri izračunu pogreška po kotu moramo s pravilno izbiro parametra $n \in \{0, 1\}$ zagotoviti, da dobimo kot rezultat



Slika 6: Prikaz vodenja do referenčne točke

funkcije $\arctan$ kot, ki je v pravilnem kvadrantu. S parametrom $m \in \mathbb{Z}$ pa moramo zagotoviti, da je zaloga vrednosti vedno znotraj območja $e_\varphi(k) \in [-\pi, \pi)$.

Za izračun translatorne hitrosti $v(k)$ uporabimo enostaven P-regulator z ojačenjem K_v:

$$v(k) = \begin{cases} K_v e_d(k) \cos^2 e_\varphi(k)\,; & |e_\varphi(k)| < \frac{\pi}{2} \wedge e_d(k) > \xi\,, \\ 0\,; & \text{sicer}\,. \end{cases} \qquad (16)$$

Razdalja ξ je radij, ki predstavlja krog okoli referenčne točke, ki ga smatramo za cilj. Za izračun kotne hitrosti $\omega(k)$ uporabimo naslednji regulator z ojačenjem K_ω:

$$\omega(k) = \begin{cases} K_\omega e_\varphi(k)\,; & e_d(k) > \xi\,, \\ 0\,; & \text{sicer}\,. \end{cases} \qquad (17)$$

Ojačanje P-regulatorja predstavljenega v (16) je 1. Ojačanje regulatorja v (17) pa se spreminja, glede na oddaljenost od točke, v katero se robot pelje. Tako zagotovimo večjo stabilnost regulatorja. Vrednosti ojačanja sta bili določeni eksperimentalno in znašata $1{,}65$, ko je robot blizu cilja, in 70, ko je daleč od njega.

Vodenje po poti lahko izvedemo s prilagoditvijo algoritma za vodenje mobilnega robota do točke tako, da izvedemo primerno preklapljanje med točkami, ki sestavljajo pot. Ko je vrednost pogreška $e_d(k)$ manjša od ξ, za referenčno lego vzamemo naslednjo točko s seznama, ki predstavlja referenčno pot. Če referenčna točka pripada stop celici, se na njej ustavimo za določen čas in nato nadaljujemo z izvajanjem algoritma.

VI. Zaključek

Razvoj algoritmov za avtonomno vožnjo je zahtevna naloga. Fizični model mesta z avtonomnimi mobilnimi sistemi nam je omogočil enostavno implementacijo in študijo osnovnih algoritmov, ki so potrebni za avtonomno vožnjo. Tekom implementacije smo se srečali s številnimi izzivi, ki nastopajo pri prenosu teorije v prakso. V prihodnosti je pričakovati, da se bo področje avtonomnih vozil vedno bolj razvijalo.

Literatura

[1] G. Klančar et al., *Wheeled Mobile Robotics: From Fundamentals Towards Autonomous Systems*, Elsevier, 2017.

[2] G. Klančar, L. Teslić in I. Škrjanc, "Mobile-robot pose estimation and environment mapping using an extended Kalman filter,"*International Journal of Systems Science*, zv. 45, št. 12, str. 2603–2618, 2014.

[3] Z. Zhang in Z. Zhao, "A Multiple Mobile Robots Path planning Algorithm Based on A-star and Dijkstra Algorithm", *International Journal of Smart Home*, zv. 8, št. 3, 75–86, 2014.

A Simulation Study of Different Model Predictive Controllers

Amila Dubravić, Zenan Šehić

University of Tuzla, Faculty of Electrical Engineering
amila.dubravic@untz.ba

Keywords: MPC Controller, Orthonormal functions, Laguerre functions, Legendre functions, Kautz functions, GPC

Abstract

The analysis has been done for the set of processes from first order up to the eight order process controlled by GPC controller and MPC controllers based on orthonormal functions (Laguerre functions, Legendre functions and Kautz functions) for the fastest possible response without overshoot.

1 Introduction

In the late 1970s model predictive control (MPC) emerged as a way of controlling a wide range of processes. It is not one specific strategy of control, but represents wide range of control methods that explicitly use model of the process in obtaining the control law by optimizing the cost function. [1] Since then, many approaches of MPC were proposed, both linear and nonlinear. They are dynamic matrix control (DMC), predictive functional control (PFC), generalized predictive control (GPC), in more recent time hybrid fuzzy model [2,3], Wiener-model based on PWL [4], hybrid MPC based on genetic algorithms [5], probabilistic neural-network [6], orthonormal functions [7,8,9].
In the following sections, short simulation study for the processes, starting from the first order system up to the eight order system controlled by different MPC techniques, will be shown. Used techniques are GPC, an older MPC technique, and MPC based on orthonormal Laguerre, Legendre and Kautz functions.

2 GPC Generalized Predictive Control

GPC is a well known algorithm. Predictions of the output are based on CARIMA model:

$$A(z^{-1})y(t) = B(z^{-1})z^{-d}u(t-1) + C(z^{-1})\frac{e(t)}{\Delta} \tag{1}$$

GPC uses quadratic cost function in the form:

$$J(N_1, N_2, N_u) = \sum_{j=N_1}^{N_2} \delta(j)\left[\hat{y}(t+j \mid t) - w(t+j)\right]^2 + \sum_{j=1}^{N_u} \lambda(j)\left[\Delta u(t+j-1)\right]^2 \tag{2}$$

where weight sequences $\delta(j)$ and $\lambda(j)$ are usually constant or exponentially increasing. Refererece trajectory $w(t+j)$ can be generated using recursion, starting from the current output and exponentially approaches setpoint. [1]

3 MPC Based on Orthonormal Functions

The design of predictive control that uses orthonormal functions occured in recent years and can be considered as new compared to GPC, particularly the approaches that use the sets of exponential orthonormal functions such as Laguerre functions, Legendre functions and Kautz functions. [7,8]
Following introduction has been made based on [7].
In MPC, the future control trajectory is denoted by

$$\Delta u(k_i), \Delta u(k_i+1), \ldots, \Delta u(k_i+N_C-1), \tag{3}$$

Nc being the control horizon dictating the number of parameters used to capture the future control trajectory. The future state variables are

$$x(k_i+1|k_i), x(k_i+2|k_i), \ldots, x(k_i+m|k_i), \ldots, x(k_i+Np|k_i), \tag{4}$$

where $x(k_i+m \mid k_i)$ is the predicted state variable at $ki+m$ having current plant information $x(k_i)$. The control horizon N_c is chosen to be less than (or equal to) the prediction horizon N_p.
Receding horizon control principle, takes only the first sample of the sequence of vector ΔU containing the controls, written in (3), and implements it, ignoring the rest of the sequence. In the next sample period, the more recent measurement is used to form the state vector for the calculation of the new sequence of control signal.
The z-transforms of the discrete-time Laguerre networks are written as

$$\Gamma_1(z)=\frac{\sqrt{1-a^2}}{1-az^{-1}}$$
$$\Gamma_2(z)=\frac{\sqrt{1-a^2}}{1-az^{-1}}\frac{z^{-1}-a}{1-az^{-1}} \quad (5)$$
$$\vdots$$
$$\Gamma_N(z)=\frac{\sqrt{1-a^2}}{1-az^{-1}}\left(\frac{z^{-1}-a}{1-az^{-1}}\right)^{N-1}$$

$l_i(k)$ denotes the inverse z-transform of $\Gamma_i(z,a)$. The set of discrete-time Laguerre functions are expressed in a vector form as

$$L(k)=\left[l_1(k)\ l_2(k)\ldots l_N(k)\right]^T \quad (6)$$

A set of Laguerre functions, $l_1(k)$, $l_2(k)$, . . . , $l_N(k)$ are used to capture the dynamic response with a set of Laguerre coefficients that are to be determined from the design process. Based on this,

$$\Delta u(k_i+k)=\sum_{j=1}^{N}c_j(k_i)l_j(k) \quad (7)$$

where k_i is the initial time of the moving horizon window and k is the future sampling instant, N is the number of terms used in the expansion, c_j, $j = 1,2,...,N$, are the coeffitients being functions of the initial time of the moving horizon window, k_i.
Equation (7) can also be expressed in a vector form:

$$\Delta u(k_i+k)=L(k)^T\eta \quad (8)$$

Where η has N Laguerre coefficients:

$$\eta=\left[c_1 \quad c_2 \quad \ldots \quad c_N\right]^T \quad (9)$$

So, the coefficient vector η is optimized and computed in the design.
With the Laguerre functions, an alternative formulation of the cost function is used and formulated with a link to discrete-time linear quadratic regulators (DLQR). The task is finding the coefficient vector η to minimize the cost function:

$$J=\sum_{m=1}^{N_p}x\left(k_i+m|k_i\right)^T Qx\left(k_i+m|k_i\right)+\eta^T R_L\eta \quad (10)$$

$Q \geq 0$ and $R_L > 0$ being weighting matrices.
Having optimal parameter vector η, the receding horizon control law is realized as

$$\Delta u(k_i)=L(0)^T\eta \quad (11)$$

The z-transforms of the discrete-time Kautz networks are written as

$$K_i(z)=\sqrt{\left(1-a^2\right)\left(1-b^2\right)}\frac{\left(z^{-1}-a\right)^{i-1}\left(z^{-1}-b\right)^{i-1}}{\left(1-az^{-1}\right)^i\left(1-bz^{-1}\right)^i}$$
$$0\leq a<1 \qquad 0\leq b<1 \quad (12)$$

a and b are poles of discrete time Kautz network. [8]
These functions are also orthonormal as are Laguerre functions. It can be seen that Kautz functions can be written in the form

$$K_i(z)=K_{i-1}(z)\frac{\left(z^{-1}-a\right)\left(z^{-1}-b\right)}{\left(1-az^{-1}\right)\left(1-bz^{-1}\right)}$$
$$0\leq a<1 \qquad 0\leq b<1$$
$$K_1(z)=\frac{\sqrt{\left(1-a^2\right)\left(1-b^2\right)}}{\left(1-az^{-1}\right)\left(1-bz^{-1}\right)} \quad (13)$$

The set of discrete-time Kautz functions are expressed in a vector form as

$$K(n)=\left[k_1(n),k_2(n),\ldots,k_N(n)\right]^T \quad (14)$$

$$K(i+1)=GK(i)+FK(i-1) \quad (15)$$

matrices G and F are $N\times N$ and they are functions of parameters $\alpha=ab$, $\beta=(a+b)$ and $\gamma=\sqrt{\left(1-a^2\right)\left(1-b^2\right)}$.

Equation (7) can be used with Kautz functions too, accordingly, coefficients are the ones for Kautz functions and insted of Laguerre functions, Kautz functions are used.
Legendre orthonormal basis is:

$$\mathrm{B}_{n-1}(z)=\frac{\sqrt{1-a_n^2}}{z-a_n}\prod_{k=0}^{n-1}\left(\frac{1-a_kz}{z-a_k}\right) \quad (16)$$

$$a_k=\frac{2-\alpha(2k+1)}{2+\alpha(2k+1)}.$$

The procedure for generation of controller is the same as in Laguerre and Kautz case, the only difference is in the basis used.

4 Simulation

Simulation has been done for the process of type

$$G = \frac{1}{(1+s)^n}, \qquad (17)$$

n being [2 3 4 5 6 7 8].

Simulation parameters are: $N_p = 50$, $N_C = N = 3$, $Q = 1$, $R = 0.1$,

$Ts = [0.025 \quad 0.05 \quad 0.075 \quad 0.1 \quad 0.125 \quad 0.15 \quad 0.175]$.

Criteria is the fastest possible response without overshoot.

Parameters were tuned empirically, mainly starting with greater value of control parameter a (close to 1).

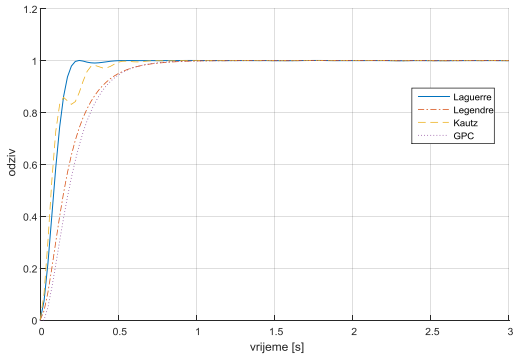


Figure 1. Step responses of the process for *n=2* controlled with Laguerre MPC, Legendre MPC, Kautz MPC and GPC

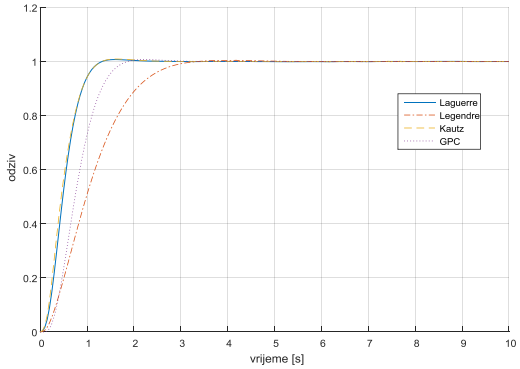


Figure 2. Step responses of the process for *n=3* controlled with Laguerre MPC, Legendre MPC, Kautz MPC and GPC

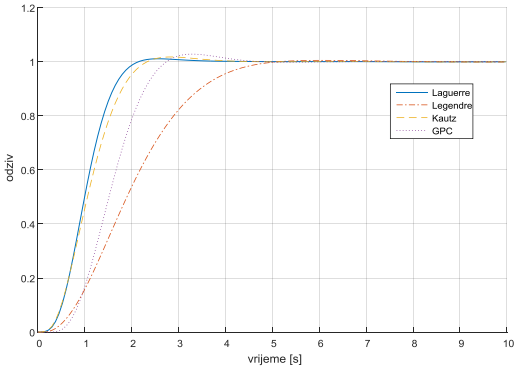


Figure 3. Step responses of the process for *n=4* controlled with Laguerre MPC, Legendre MPC, Kautz MPC and GPC

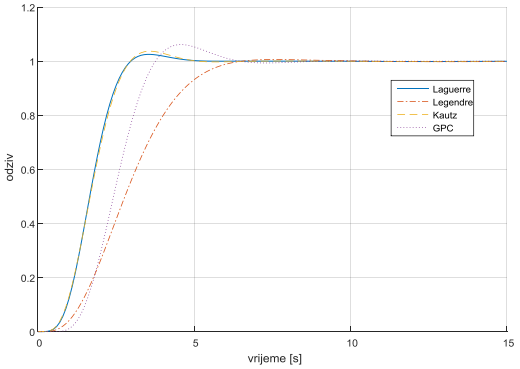


Figure 4. Step responses of the process for *n=5* controlled with Laguerre MPC, Legendre MPC, Kautz MPC and GPC

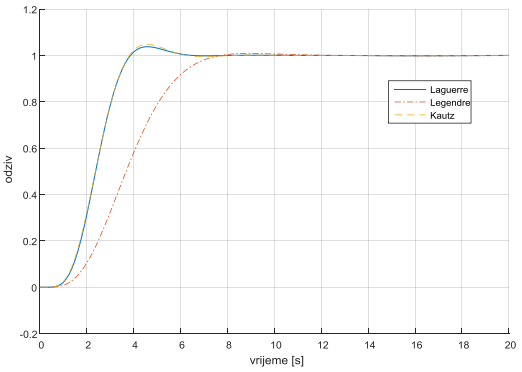


Figure 5. Step responses of the process for *n=6* controlled with Laguerre MPC, Legendre MPC, Kautz MPC

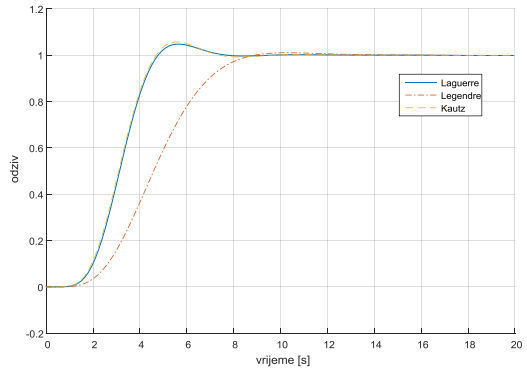


Figure 6. Step responses of the process for *n=7* controlled with Laguerre MPC, Legendre MPC, Kautz MPC

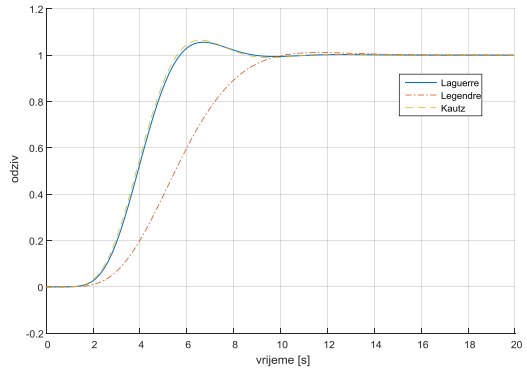


Figure 7. Step responses of the process for *n=8* controlled with Laguerre MPC, Legendre MPC, Kautz MPC

5 Conclusions

This paper presents results obtained in the set of simulations for the set of processes presented with equation (17), starting from the first order process, up to the eight order process, controlled with different types of MPC controllers. The idea was also to use small and equal control horizon for all the controllers. MPC controllers are GPC and MPC based on Laguerre orthonormal functions, Legendre orthonormal functions and Kautz orthonormal functions. The step setpoint was treated. The goal was the fastest possible response without overshoot. As can be seen from the figures, for the concrete set of processes, under the conditions as stated above, only in the case of Legendre functions based MPC, it is possible to get the response without overshoot for all values of *n*. For some values of *n*, the goal is also achieved with other types of MPC controllers with even faster response, mainly Kautz functions based MPC and Laguerre functions based MPC. In other cases the goal could be achieved by setting constraints. In the last three types of processes, *n* being *6*, *7* and *8* GPC did not give bounded output.

6 References

[1] E.F. Camacho, C. Bordons. Model Predictive Control, Springer-Verlag, (1999).

[2] D. Dovžan, I. Škrjanc . Predictive functional control based on an adaptive fuzzy model of a hybrid semi-batch reactor., Control Engineering Practice Vol. 18 Issue: 8 pp. 979-989 (2010).

[3] A. Nunez, D. Saez, S. Oblak et al. Fuzzy-model-based hybrid predictive control., ISA Transactions Vol. 48 Issue: 1 pp. 24-31 (2009).

[4] S. Oblak, I. Škrjanc. Continuous-time Wiener-model predictive control of a pH process based on a PWL approximation., Chemical Engineering Science Vol. 65 Issue: 5 pp. 1720-1728 (2010).

[5] J. Causa, G. Karer, A. Nunez et al. Hybrid fuzzy predictive control based on genetic algorithms for the temperature control of a batch reactor., Computers & Chemical Engineering Vol. 32 Issue:12 pp. 3254-3263 (2008).

[6] B. Potočnik, G. Mušič, I. Škrjanc et al. Model-based predictive control of hybrid systems: A probabilistic neural-network approach to real-time control., Journal of Intelligent & Robotic Systems Vol. 51 Issue: 1 pp. 45-63 (2008).

[7] L. Wang. Discrete model predictive control design using Laguerre functions., Journal of Process Control. pp. 14:131–142 (2004).

[8] B. Khan, J.A. Rossiter, G. Valencia-Palomo. Exploiting Kautz functions to improve feasibility in MPC. 18th IFAC World Congress, Milano (Italy), August 28 – September 2, (2011).

[9] A. Dubravić, Z. Šehić. Using Orthonormal Functions in Model Predictive Control. Tehnički vjesnik – Technical Gazzete no. 19, Vol. 3, (2012).

Replacing reward function with user feedback

Zvezdan Lončarević, Rok Pahič, Aleš Ude, Bojan Nemec, Andrej Gams

All authors are with the Department of Automatics, Biocybernetics and Robotics, Jožef Stefan Institute, and with the Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
e-mail: zvezdan.loncarevic@ijs.si

Abstract

Reinforcement learning refers to powerful algorithms for solving goal related problems by maximizing the reward over many time steps. By incorporating them into the dynamic movement primitives (DMPs) which are now widely used parametric representations in robotics, movements obtained from a single human demonstration can be adapted so that a robot learns how to execute different variations of the same task. Reinforcement learning algorithms require carefully designed cost function, which in most cases uses additional sensors to evaluate some environment criteria.

In this paper we explore possibilities of learning robotic actions using only user feedback as a reward function. Two reward functions have been used and their results are presented and compared. These were user feedback and a simplified reward function. Experimental results show that for the simple actions where only the terminal reward is given, these 2 reward functions work almost as good as having a reward function based on the exact measurement.

1 Introduction

As robots become a mass consumer product, besides standard industrial environments, they are expected to be able to perform in unstructured environments, such as households or some other real-life situations. One of the biggest barriers to wider implementation of robots in our home environment is the lack of environment models. In most cases, a complex sensory system is required to estimate relevant environment parameters. That is why standard ways of programming robotic movements, that need an expert programmer for every variation of the task, are not sufficient. With humanoid robots with many degrees of freedom, autonomy is one of the main unresolved issues in contemporary robotics [1]. Imitation and reinforcement learning are the two most common approaches for solving this issue. Because robot actions are mainly recorded as parametric representations with many parameters search space for reinforcement learning algorithms that use gradient methods (finite difference gradient, natural gradient, vanilla gradient, etc.) is large. Only recently, probabilistic algorithms such as PI^2 and PoWER have been developed and able to deal with high dimensional search space [2, 3].



Figure 1: Experimental setup with Mitsubishi PA-10 robot

One of the major problems in reinforcement learning is determining the reward function. It was usually solved by modifying reward function according to the teachers behavior [4, 5, 6].

The main goal of this paper is to show that a simplified, reduced reward function can be used directly for some robot actions. An example of such is throwing a ball, i. e. the robot learns how to perform an accurate throwing action relying only on feedback from naive user or cheap imprecise sensors.

The paper is organized as follows: In the next section, we briefly present dynamical movement primitives that are already a widely used method in programming by demonstration, and a probabilistic algorithm called PoWER. We used it to learn new DMP parameters for the task. In Section III we introduce the reduced reward functions that were used. Next section presents experimental setup and results obtained in simulation and on the real robot. The paper concludes with a short outlook on the obtained results and suggestions for future work.

2 Reinforcement Learning

In this section we provide the DMPs and PoWER basics.

2.1 Dynamic Movement Primitives

The basic idea of Dynamic Movement Primitives (DMPs) [7] is to represent the trajectory with the well known dynamical system described with equations for the mass on spring-damper system. For a single degree of freedom (DOF) denoted by y, in our case one of the joint task-space coordinates, DMP is based on the following second order differential equation:

$$\tau^2 \ddot{y} = \alpha_z(\beta_z(g - y) - \tau\dot{y}) + f(x), \tag{1}$$

where τ is the time constant and it is used for time scaling (time in which the trajectory needs to be reproduced), α_z and β_z are damping constants ($\beta_z = \alpha_z/4$) that make system critically damped and x is the phase variable. The nonlinear term $f(x)$ contains free parameters that enable the robot to follow any smooth point-to-point trajectory from the initial position y_0 to the final configuration g [8]. The phase and kernel functions are given by

$$f(x) = \frac{\sum_{i=1}^{N} \psi(x)\omega_i}{\sum_{i=1}^{N} \psi_i(x)} x, \tag{2}$$

$$\psi_i(x) = \exp\left(-\frac{1}{2\delta_i^2}(x - c_i)^2\right), \tag{3}$$

where c_i are the centers of radial basis functions ($\psi_i(x)$) distributed along the trajectory and $\frac{1}{2\delta_i^2}$ (also labeled as h_i [9]) their widths. Phase x makes the forcing term $f(x)$ disappear when the goal is reached because it exponentially converges to 0. Its dynamics are given by

$$x = \exp\left(-\alpha_x t/\tau\right), \tag{4}$$

where α_x is a positive constant and x starts from 1 and converges to 0 as the goal is reached. Multiple DOFs are realized by maintaining separate sets of (1–3), while a single canonical system given by (4) is used to synchronize them. The weight vector **w**, which is composed of weights w_i, defines the shape of the encoded trajectory. Learning of the weight vector using a batch approach has been described in [10] and [7] and it is based on solving a system of linear equations in a least-squares sense.

2.2 PoWER

Before the robots become more autonomous, they would have to be capable of autonomous learning and adjusting their control policies based on the feedback from the interaction with varying environment. Policy learning is an optimization process by which we want to maximize the state value of cost function:

$$J(\theta) = E\Big[\sum_{k=0}^{H} \alpha_k r_k(\theta)\Big]. \tag{5}$$

with varying policy parameters $\theta \in R^n$ [11]. In the last equation (5), E is the expectation value, r_k is reward given for time step k and it depends of parameters that are chosen (θ). H is the number of time steps in which reward is given and α_k are the time step dependent weighting factors. Although there are numerous methods that can be used to optimize this function, the main problem is high dimensionality of parameters θ (for DMP it is typically 20-50 per joint, we use 25 per joint). Stochastic methods as PI^2 and PoWER were recently introduced. It was proven that PI^2 and PoWER perform identically when only terminal reward for periodic learning was available. With PoWER it is easy to incorporate also other policy parameters [1] such as starting (y_0) and ending point (g) and time of execution (τ) of the trajectory for each DOF and not only DMP weights (**w**) so in this research the parameters that are learned are:

$$\theta = \big[\mathbf{w}, g, y_0\big]. \tag{6}$$

During the learning process this parameters are updated using the rule:

$$\theta_{m+1} = \theta_m + \frac{\sum_{k=1}^{L}(\theta_{ik} - \theta_m)r_k}{\sum_{k=1}^{L} r_k} \tag{7}$$

where θ_{m+1} and θ_m are parameters after and before update, L is number of parameters in importance sampler matrix and they represent L executions with highest rewards (r_k). θ_i is selected using stochastic exploration policy

$$\theta_i^* = \theta_m^* + \epsilon_i, \tag{8}$$

where ϵ_i is Gaussian zero noise. Variance of the noise (σ^2) is the only tuning parameter of this method. Because θ consists of three DMP parameters that cannot be searched with the same variance, it requires three different variances to be chosen. In general, higher σ^2 would lead to faster, and lower σ^2 to more precise convergence.

3 Reward function

As the goal of this paper is to examine the possibility to avoid need for expensive and precise sensors that can be set and calibrated only in laboratories for some actions to be successful, and enable a user to train a robot to perform variation of some action with simple instructions, thus discretized reward function is introduced. Two reward functions are compared between each other and to the exact reward function based on the distance measured with sensors.

In the first reward function (unsigned), rewards are given on a five-star scale where the terms {"one star","two stars","three stars","four stars" and "five stars"} have the corresponding rewards: r={1/5, 2/5, 3/5, 4/5, 1}. This means that the robot did not know in which direction to change its throws. Similar is presented in [12].

In the second reward function (signed), robot converged to its target using the feedback in which reward was formed using five possible rewards:"too short" ($r = -1/3$), "short" ($r = -2/3$), "hit" ($r = 1$), "long" ($r = 2/3$), and "too long" ($r = 1/3$). This allowed us to always put in importance sampler matrix the shots that are on the different sides of the target.

This means that both functions are of the same complexity because they have only five possible rewards. Although the five-star system can discretize better, the second should be able to compensate its lower discretization

with the new way of choosing movements that will be in the importance sampler.

4 Experimental evaluation

4.1 Experimental setup

The experiment was conducted in simulation and on a real-system. People as well as computer-simulated human reward systems were used.

The participants used a GUI to rate (give terminal reward) to the shots until the robot manages to hit the target.

In order to evaluate statistical parameters describing the success of learning, we have also created the program that should simulate human ratings with discretized reward function and with the variance between the reward borders (to simulate human uncertainty). The uncertainty was determined empirically. Simulation and this program allowed us to make much more trials without human participants. This way we tested our algorithm for ten different positions of the target with the diameter of 10 cm. (30 trials for each position) within the possible range of the robot (between 2.4m to 4m providing that the robot base was mounted on a stand of 1 m height).

Finally, the functionality of the algorithm was confirmed in the real-world using the Mitsubishi PA-10 robot. It needed to hit the basket with a 20 cm diameter with a 13 cm diameter ball. The experimental setup with the real robot is shown In Fig. 1.

4.2 User study

In the following, human-robot interaction (HRI) study with volunteer participants that are naive to the learning algorithm is described. Participants were using graphical user interface (GUI) in which they could rate the success of the shot in the five-star rating system (first reward function) and in the GUI where they could choose if the shot was "too short", "short", "hit", "long", or "too long" (second reward function) in a randomized order. They were not informed on how the reward function works, but only that learning was finished after they rate a shot with five stars in the first reward function system, or after they pressed "hit" in the second reward function system. The maximum number of iterations was 120. Tests with users were conducted using a simulation created in Matlab, where the robot model was made using the same measurements and limitations as on the real robot.

4.3 Results

Figure 2 shows throwing error convergence, which tends towards zero for all experiments. The top plot shows the results where human judgment was simulated using Matlab environment and presents statistics of 300 trials. Each trial had update in 120 iterations. Bottom figure represents error convergence from the results with human judgment. Although human judgment criteria differs among participants, the robot was still able to converge to its goal.

In Fig. 3 rewards that were given to the executed shot are shown. The top plot shows rewards (r) that were computer generated and the bottom one shows rewards that were given using human judgment. Statistics for computer generated rewards is taken from 300 trials and for human rewards from 10 participants that did the experiment.

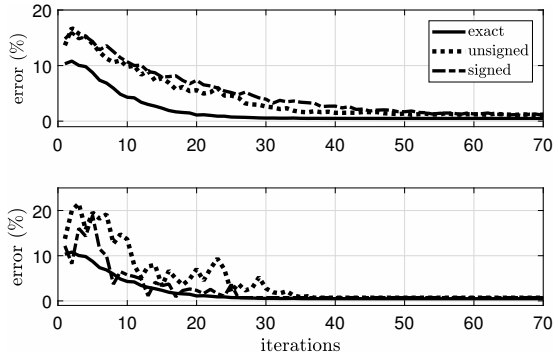


Figure 2: The top picture shows mean error convergence caused by computer simulated human judgment and the bottom one by real human judgment. Solid line denotes the results based on the real measurement reward function, dotted denotes the results for the unsigned reward function and dashed denotes results for the signed reward function.

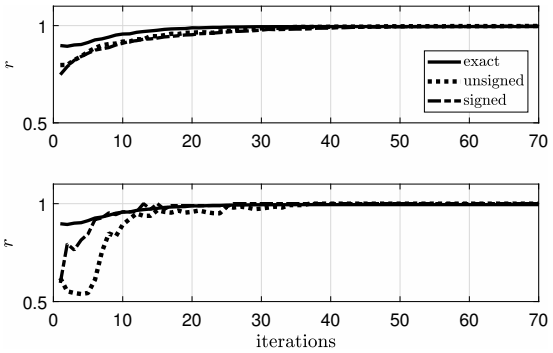


Figure 3: The top picture shows computer simulated reward and the bottom one real human mean reward convergence. Solid line denotes the results based on the real measurement reward function, dotted denotes results for the unsigned reward function and dashed denotes results for the signed reward function.

Figure 5 (top) shows the statistics of average and the last (bottom) iteration in which the first shot happened in computer simulated judgment (shaded bars) and in the experiment with people (white bars). Results are shown for the cases where exact distance was measured, for five-star reward function (unsigned discrete), and for the reward function where people were judging according to the side of the basket where the ball fell (signed discrete). It shows that with computer simulated human judgment (shaded bars), unsigned works only slightly better, but with the signed one, the worst case scenario is better. Surprisingly, with the human judgment, signed reward function drastically outperformed the unsigned. This can be explained by the fact that people tend to rate the shots in comparison to the previous one and that way unintentionally form a gradient. This is a fact similar to what was discussed in [13].

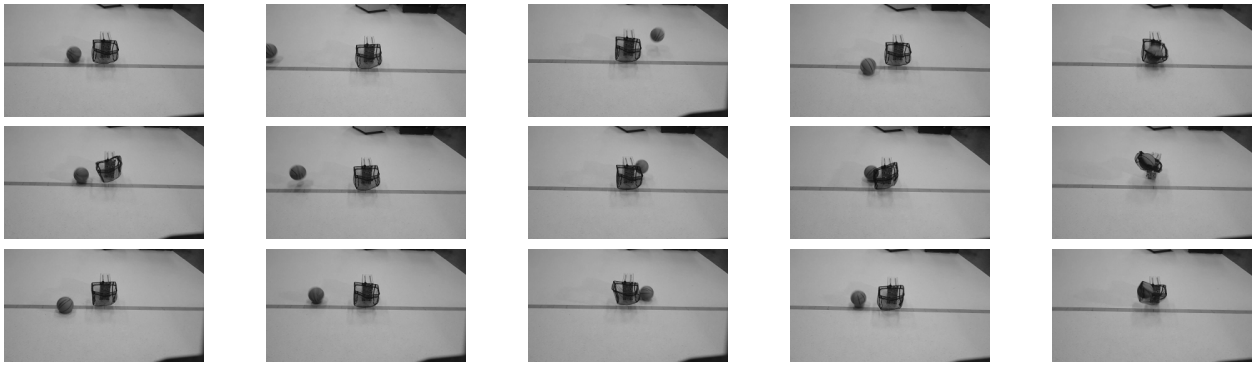


Figure 4: Experiment on the real robot: In the first row, exact distance was measured, in the second row unsigned reward system was used and in the third one signed reward function was used. The robot was throwing the ball from the right.

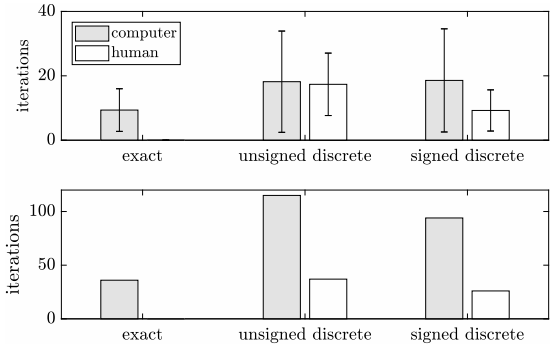


Figure 5: Average first shot and its deviation in Nakagami distribution (top) and worst case first shot (bottom): computer (shaded bars), human (white bars)

5 Conclusion

Results show that reinforcement learning of some simple tasks can be done with the reduced reward function almost equally good as with exact measurement based reward function. Even so, the problems related to reinforcement learning in general stay the same: it is very difficult and time-consuming task to set the appropriate noise variance (especially if more parameters are varied like in this case). Note that excessive noise variance results in jerky robot trajectories, which might even damage the robot itself.

That is why in the future, we will test this using the latent space of the neural network, where we have less different search variances to tune and also a possibility to train neural network on executable shots so that too big variance cannot lead to the trajectories that differ a lot from the original, demonstrated one.

References

[1] B. Nemec, D. Forte, R. Vuga, M. Tamosiunaite, F. Worgotter, and A. Ude, "Applying statistical generalization to determine search direction for reinforcement learning of movement primitives," *IEEE-RAS International Conference on Humanoid Robots*, pp. 65–70, 2012.

[2] E. Theodorou, J. Buchli, and S. Schaal, "A Generalized Path Integral Control Approach to Reinforcement Learning," *Journal of Machine Learning Research*, vol. 11, pp. 3137–3181, 2010.

[3] J. Kober and J. Peters, "Learning motor primitives for robotics," *2009 IEEE International Conference on Robotics and Automation*, pp. 2112–2118, 2009.

[4] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," *Twenty-first international conference on Machine learning - ICML '04*, p. 1, 2004.

[5] W. Knox, C. Breazeal, and P. Stone, "Learning from feedback on actions past and intended," *In Proceedings of 7th ACM/IEEE International Conference on Human-Robot Interaction, Late-Breaking Reports Session (HRI 2012)*, 2012.

[6] S. Griffith, K. Subramanian, and J. Scholz, "Policy Shaping: Integrating Human Feedback with Reinforcement Learning," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1–9, 2013.

[7] A. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 2, no. May, pp. 1398–1403, 2002.

[8] D. Forte, A. Gams, J. Morimoto, and A. Ude, "On-line motion synthesis and adaptation using a trajectory database," *Robotics and Autonomous Systems*, vol. 60, no. 10, pp. 1327–1339, 2012.

[9] A. Gams, A. J. Ijspeert, S. Schaal, and J. Lenarčič, "On-line learning and modulation of periodic movements with nonlinear dynamical systems," *Autonomous Robots*, vol. 27, no. 1, pp. 3–23, 2009.

[10] 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, no. 5, pp. 800–815, 2010.

[11] B. Nemec, A. Gams, and A. Ude, "Velocity adaptation for self-improvement of skills learned from user demonstrations," *IEEE-RAS International Conference on Humanoid Robots*, vol. 2015-February, no. February, pp. 423–428, 2015.

[12] A.-L. Vollmer and N. J. Hemion, "A User Study on Robot Skill Learning Without a Cost Function: Optimization of Dynamic Movement Primitives via Naive User Feedback," *Frontiers in Robotics and AI*, vol. 5, no. July, 2018.

[13] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.

Razvoj vibracijskega rotacijskega mikromotorja v tekočini

BožidarBratina[1], Álvaro Baines García[2], Matjaž Petek[1], Suzana Uran[1], Riko Šafarič[1]

[1]*Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Smetanova ulica 17, 2000 Maribor*
[2]*Public University of Navarre, Campus de Arrosadia, 31006 Pamplona, Navarra, Španija*
E-pošta: bozidar.bratina@um.si

Abstract.
The paper presents a micromotor development based on vibration-induced rotation flow in a fluid, where a rotational velocity of the fluid which rotates the object (rotor) is obtained by external vibrations. Two types of fluidic motors are presented, firstly a micro sized motor as a microstructure consisting of four pillars (stator) and a hexagon ring made of glass micro-spheres (rotor). And secondly a submillimetre sized motor with rotating disk on the surface of the fluid in a pot. The paper explains calculations of the micro-motor design, and use of van der Waals and capillary force methods to build microstructures. The results confirm the vibration-induced rotation flow around pillars, where carefully designed pillar structure enables equilibrium point inside the pillars for micro rotor rotation. Several microstructure designs under different induced vibrations were tested. Both micro sized and submillimetre sized motor designed with different rotors (disk, ring and axis mounted ring) were tested and successfully rotated.

1 Uvod

Sestavni del vsakega robota predstavljajo aktuatorji, pogosto elektromotorji, ki skrbijo za linearno ali rotacijsko premikanje robotskih delov. Trendi s področij biomedicine in nanotehnologij se kažejo v integraciji raznih miniaturnih robotskih sistemov, ki bi omogočali manipulacije na mikronivoju (gradnja mikrostruktur, ciljno doziranja zdravil, itd). Miniaturizacija oblik robotov pa zaradi velikosti sestavnih delov ter principov delovanja pogonskih sklopov (virov energije) ni enostavna. Elektromagnetne sile v mikrosvetu niso nujno prevladujoče ter je posledično potrebno izkoriščati druge sile za delovanje. Današnji mikro in nanopogoni temeljijo na mehanskih ali kemijskih principih. Takšni motorji se že uporabljajo v praksi (Piezo) za translatorna in omejena rotacijska gibanja, razne eksotične izvedbe pa služijo bolj kot dokaz delovanje in so v praksi še neuporabne. [1, 2] Mikromotor, ki bi deloval podobno kot elektromotor (vrtenje z možnostjo vodenja vrtljajev) pa v mikrosvetu ni enostavno zgraditi zaradi miniaturne zgradbe in vplivov različnih sil na objekte.

V prispevku je predstavljen razvoj in izgradnja mikromotorja s funkcijo vrtenja osi s pomočjo zunanjih induciranih vibracij. Takšni mikromotorji so lahko aplikativno uporabljeni kot črpalke na področju mikrofluidike, pogonski stroji, itd. V nadaljevanju je opisan razvoj mikromotorja, kjer vrtenje rotorja v tekočini temelji na vrtečem se pretoku tekočine, ki ga ustvarjajo umetne vibracije okoli ali znotraj pritrjenih mikroobjektov. [3, 4] Opisan je princip delovanja in razvoj mikrostrukture motorja v tekočini, ter opis potrebne vibracijske naprave. V tretjem poglavju sledi prikaz testiranja vrtenja tekočine okrog pritrjenih mikroobjektov, mikroobjektov okrog mikrostruktur in znotraj mikrostrukture, v tekočini zaradi induciranih vibracij. Prispevek zaključujemo s prikazom razvoja motorja v lončku na submilimetrskem nivoju, kjer je kot rotor uporabljen disk, obroč in obroč nataknjen na os, ki se vrti na površini tekočine. Prikazani rezultati potrjujejo osnovni princip delovanja razvitih mikromotorjev z vrtenjem v ali na površini tekočine.

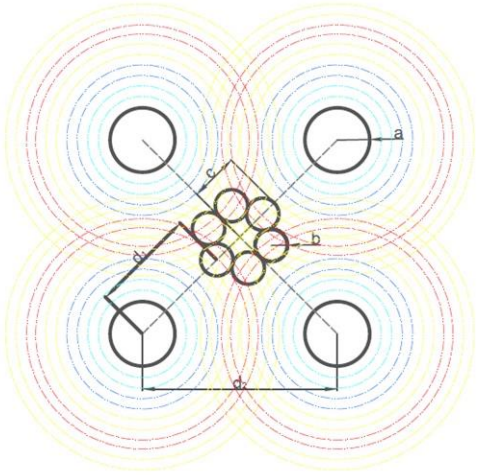
2 Vibracijski rotacijski mikromotor v tekočini

Dandanes je z mikroobjekti že možno zgraditi mikrostrukture. Predstavljene metode z uporabo van der Waalsovih [5] ali kapilarnih sil [6] omogočajo relativno hitro izgradnjo, končno lepljenje ali sintranje pa utrdi mikrostrukture (zobnik, os, ležaj, itd), da so primerne za uporabo v različnih aplikacijah. Te enostavne strojne gradnike je mogoče sestaviti v reduktor ali kakšen bolj kompleksen gibajoč mehanski mikrostroj. Za pogon je seveda potrebno energija in njen prenos, kjer smo se osredotočili na princip vrtenja mikroobjekta (mikromotor) s pomočjo vibracij v ustreznih atmosferskih pogojih. Vibracije namreč vzbudijo pretok tekočine tankega filma na površini mikroobjekta, le-ta pa s kapilarnim principom posledično vrtenje mikroobjekta. Takšno primitivno in nekontrolirano vrtenje steklene krogle, smo omejili in nadgradili z valjasto obliko rotorja, ki se lahko vrti le v smeri vzdolžne osi. Zaradi nezadovoljivega vrtenja zaradi vpliva težnosti, pretankega filma vode in prenosa vibracij na mikrostrukturo, smo bili primorani koncept spremeniti. Glede na različne vire [3, 4], ki opisujejo razvoj mikrofluidnih generatorjev pretoka vzbujanih z mehanskimi vibracijami smo preizkusili tehnologijo za izdelavo mikrorotorjev različnih dimenzij. Vibracije, ki povzročijo gibanje tekočine omogočajo kontroliran transport mikroobjektov po tekočini. Z ustrezno načrtano strukturo je možno doseči vrteč pretok tekočine okoli ali znotraj pritrjenih mikroobjektov. Pri razvoju smo izkoristili predhodno znanje gradenj mikrostruktur, da smo zgradili testni poligon vrtenja mikroobjektov. Na mikro nivoju so strukture običajno ustvarjene z jedkanjem, v našem primeru pa so zgrajene iz mikrodelcev s pomočjo nanorobotskega manipulatorja z enoprstnim prijemalom [5], ki omogoča

prenos mikroobjektov z Van der Waalsovo ali kapilarno silo. Prijemanje/ odlaganje mikroobjektov je kontrolirano s hlajenjem/gretjem v lokalni okolici. [6] Vse mikrostrukture v prispevku so sestavljene iz steklenih kroglic in utrjene z lepljenjem.

2.1 Pretok tekočine na osnovi vibracij

Avtorji [3] so prikazali transport in vrtenje posameznega mikroobjekta (celice velikosti cca. 100-150 μm) s pomočjo pritrjenih stebrov na površino, ki je vzbujana z ustreznimi vibracijami. Rotacijo celice dosežejo z zaprto strukturo treh stebrov, kjer se ustvari ravnovesno stanje sil, ki vrtijo objekt z določeno hitrostjo in smerjo. S spremembo frekvence in/ali amplitude vibracij je mogoče premikanje oz. vrtenje kontrolirati, kar omogoča opazovanje celic v prostoru za namene biomedicinskih aplikacij. Vibracije, pravokotne na površino, okrog pritrjenega mikroobjekta v tekočini povzročijo usmerjen krožni tok tekočine okrog njega. Gibanje površine z vibracijami mora biti usklajeno po X-Y osi s sinusnim/ kosinusnim zbujanjem, da se doseže krožno gibanje. Glede na vir [3] je maksimalna hitrost krožnega toka dosežena pri razdalji cca. 100 μm od površine stebra, seveda pa je odvisno od oblike in zasnove stebra. S postavitvijo več stebrov v zaprto obliko je možno krožni pretok okrog posameznih stebrov usmeriti v krožno gibanje tekočine znotraj strukture, kar izkoriščamo za vrtenje mikroobjektov v centru. S teoretičnimi izračuni glede na število stebrov in dimenzij mikroobjekta (rotor) je na sliki 1 prikazana zasnova poligona za testiranje vrtenja v centru.



Slika 1. Tlorisni prikaz pretoka tekočine (črtkane črte) okrog štirih stebrov in šestkotnega mikrorotorja znotraj stebrov.

Dimenzije gradnikov so sledeče: polmer stebra a=40 μm, polmer steklenih kroglic šestkotnika b=20 μm, polmer šestkotnika c=60 μm. Načrtovanje celotnega poligona, razmik med stebri, in razmik med stebri in rotorjem v centru je izveden po [3]. Po enačbi (1) se določi razmik med rotorjem in stebri d_1=100 μm, a da je zadoščen pogoj v enačbi, je izbran malce večji razmik d_1=120 μm.

$$d_1 > a + c \tag{1}$$

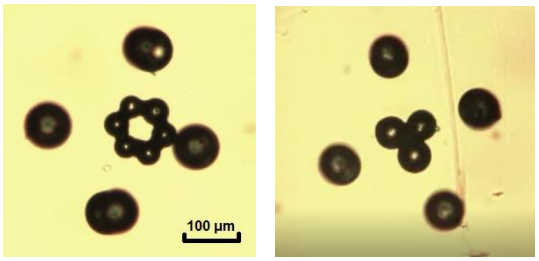
Stranica enakostraničnega trikotnika med dvema stebroma in centrom rotorja je določena z izrazom (2).

$$stranica = d_1 + (c - \frac{b}{2}) \tag{2}$$

Po (3) se lahko izračunamo zadostna razdalja med stebri.

$$d_2 = \sqrt{\left(\left(\left(c - \frac{b}{2}\right) + d_1\right)^2 + \left(\left(c - \frac{b}{2}\right) + d_1\right)^2\right)} \tag{3}$$

Glede na dimenzije je izračunana razdalja d_2=226 μm. Stebri so izdelani iz dveh večjih steklenih kroglic postavljenih ena na drugo ter prilepljeni na površino steklene ravnine z UV lepilom. Manjše steklene kroglice pa so sestavljene v rotor šestkotne ali trikotne oblike, ter zlepljene v čvrsto mikrostrukturo. Slika 2 prikazuje končno obliko poligona štirih stebrov z rotorjem v centru.



Slika 2. Prikaz različnih oblik rotorjev vstavljenih med štiri stebre v tekočini. Merilna skala velja za slike 2-4.

2.2 Vibracijska naprava

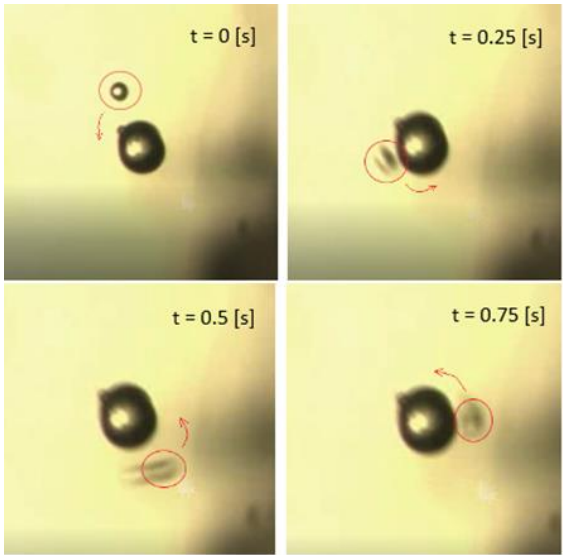
Vibracijska naprava krožno vzbuja ravno stekleno površino (mizico) z nameščenim lončkom ali štirimi stebri, da povzroči krožno gibanje tekočine znotraj lončka ali okoli stebrov. Zgrajena je iz dveh linearnih piezo motorjev pravokotnega preseka 6 x 6 mm, ki sta nameščena medsebojno pravokotno, eden po x-osi, drugi po y-osi. Mizica vibracijske naprave je vpeta med oba linearna piezo motorja. Na obeh oseh vibracijske naprave so uporabljeni linearni piezo motorji proizvajalca Nanofaktur MPO 050015. Ta piezo motor izvaja pomike v območju 0-15 μm s krmilno napetostjo med 0 in 150 V (max. -45V do +180 V). V osnovnem območju napetosti lahko ti linearni motorji proizvajajo silo potiska do 1125 N, silo vleka pa 50 N. Togost linearnega piezo motorja je 75 N/μm. Vsak linearni piezo motor krmilimo s pomočjo eno kanalnega gonilnika EOO-050100 istega proizvajalca. Ta gonilnik omogoča povečan obseg izhodne krmilne napetosti, nastavitev offset napetosti med 0 in 10V, na njegov vhod pa lahko priključimo napetost v območju -3 do 12V. Krožno gibanje mizice vibracijske naprave dosežemo tako, da krmilimo premikanje linearnega motorja na x-osi po sinusni funkciji, medtem ko

krmilimo premikanje linearnega motorja na y-osi po kosinusni funkciji. Sinusni in kosinusni časovni potek napetosti generiramo s pomočjo mikrokrmilnika PIC in pulzno širinske modulacije s frekvenco 50kHz. S filtriranjem pulzno širinskega signala dobimo sinusni in kosinusni signal s frekvenco v območju 0 do 1kHz.

3 Testiranje vrtečega pretoka tekočine okrog in znotraj poligona

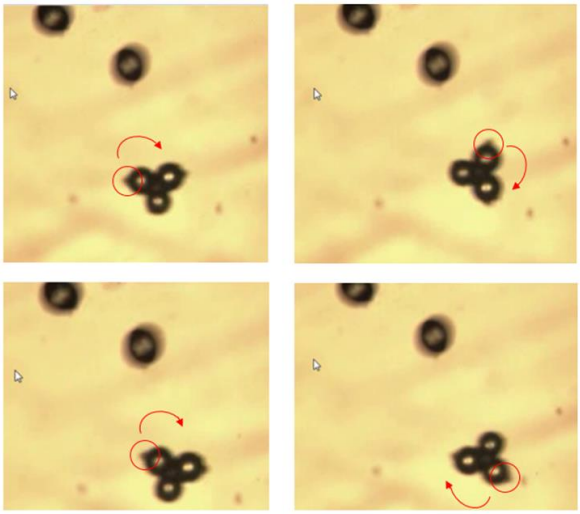
Testiranje smo pričeli s preizkušanjem vrtenja mikroobjekta, kjer je bila celotna ravna površina s poligonom in mikroobjektom potopljena v tekočino (kapljica vode). Potopljen poligon oz. površino steklene ravnine smo pričeli vzbujati z vibracijami ter tako ustvarili pretok tekočine okrog mikroobjektov. Intenziteto vibracij smo nastavljali s spreminjanjem napetosti aktuatorjev (vibrator), kjer je bilo za zadovoljivo krožno vzbujanje potrebno zagotoviti frekvenco cca. 250 Hz in amplitudo 10 µm. Dokler je mikroobjekt (steklena kroglica) potopljen na dnu kapljice oziroma v kontaktu s površino ravnega stekla deluje kot steber, zaradi česar se okrog njega vzpostavi krožni tok. Če je pretok okrog njega dovolj močan se mikroobjekt prične vrteti. Če ga sila vzgona med vrtenjem dvigne od tal krožni tok okrog njega oslabi in izgubi vrtilno hitrost ter potone na dno. Z ustrezno nastavitvijo frekvence in amplitude vibracijske naprave je bilo mogoče doseči zanesljivo vrtenje mikroobjekta, ter krmiliti hitrost glede na intenziteto vzbujanja. Hitrosti krožnega vrtenja je bila med $1 < \omega < 10$ rad/s.

Nadalje smo na stekleno površino pritrdili steber iz dveh kroglic ter posneli obnašanje mikroobjekta v krožnem toku tekočine okrog stebra. Opisan primer je predstavljen s serijo slik (slika 3), ki kažejo na vrtečo kroglico okrog stebra. Slika stebra je zaradi induciranih vibracij nejasna. Nazadnje smo izdelali še mikrostrukturo rotorja v obliki šestkotnika, ki ga lahko dogradimo v enostaven zobnik. Sestavljen in zlepljen je iz steklenih kroglic, ter vstavljen med štiri stebre.



Slika 3. Krožno vrtenje steklene kroglice v tekočini okrog pritrjenega stebra na površino z induciranimi vibracijami.

Izkazalo se je, da je bila lepljena mikrostruktura rotorja problematična saj se je del rotorja z UV lepilom zaradi velike adhezivnosti površine lepila večkrat dotaknil stebrov in rotor prilepil na steber. Zato smo izvedli poenostavljeno verzijo rotorja, ki je bil sestavljen iz le treh kroglic v obliki trikotnika zlepljenega v sredini. Zaradi manjših dimenzij nove oblike trikotnega rotorja le-ta ni pravilno usklajen z načrtovanimi dimenzijami poligona in je pri neustreznem vzbujanju zdrsnil iz centra poligona. Potreben bi bil nov preračun in postavitev poligona, da bi v notranjosti ustvarili ravnovesne pogoje za enakomerno vrtenje trikotnega rotorja. Slika 4 predstavlja zaporedje vrtenja trikotnega rotorja v tekočini med vzbujanjem z vibracijami, kjer je ušel iz centra poligona. Na slikah rotorja je opazna nepravilnost na kroglici, kar služi za indikacijo vrtenja.



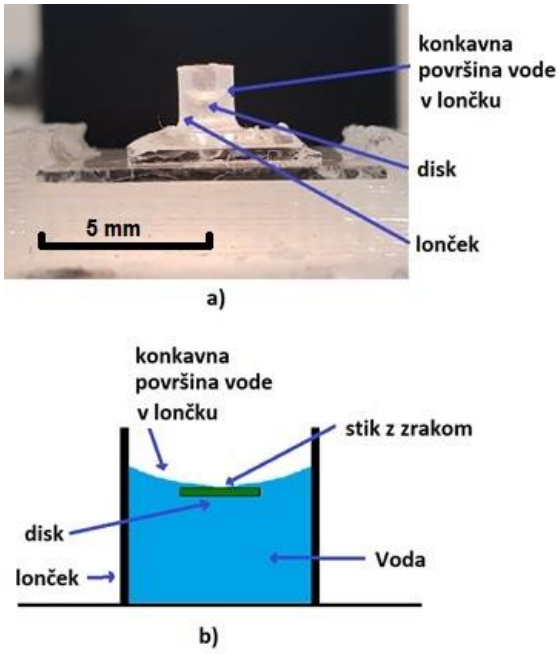
Slika 4. Inducirano vrtenje trikotnega rotorja na ravni površini, izven poligona (dva stebra od štirih sta vidna zgoraj).

S povečanjem števila stebrov se poligon spremeni v zaprto obliko oziroma lonček. Zato smo vzporedno z razvojem fluidnega mikromotorja preverili tudi delovanje malo večje submilimetrske verzije motorja.

4 Submilimetrski vibracijski mikrofluidni rotacijski motor

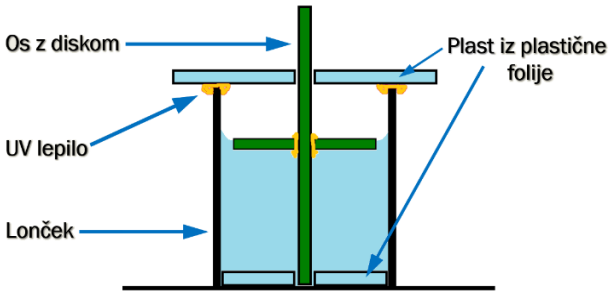
Na zgornji predpostavki smo nadaljevali razvoj submilimetrske verzije, kjer na ravno stekleno površino namestimo lonček višine cca. 1,5 mm in notranjim premerom 0,9 mm v katerem se nahaja voda z nadrobljenimi obruski papirja. Površino krožno vzbujamo s pomočjo vibracijske naprave, s frekvenco 260 Hz in amplitudo 10 µm. Center vrtenja vode je bil točno v središču lončka. Če vibracija ni bila natančno krožno vibriranje, se je center vrtenja premaknil izven središča ali pa so se v lončku ustvarili dva ali tri središča vrtenja. Cilj je doseči krožno gibanje medija z enim središčem v centru krožnice lončka. Z eksperimentom smo dokazali delovanje krožnega toka ob različnih amplitudah in frekvencah vibriranja in dosegli nastavljivo krmiljenje kroženja vode v urini ali protiurini smeri z rotacijsko krožno hitrostjo vode na površini v lončku -15 rd/s $< \omega < +15$ rd/s.

Naslednji eksperiment, ki smo ga opravili je bil rotirajoči se disk z diametrom cca. 500 µm, ki je "lebdel" neposredno pod konkavno površino vode, zaradi površinske napetosti na stiku med vodo in zrakom. Disk, narejen iz plastične folije debeline cca. 200 µm smo vrteli v obeh smereh s krožno hitrostjo cca. -15 rd/s < ω < +15 rd/s. Eksperimentalno napravo vibrirajočega lončka in diska prikazuje slika 5a in njegovo skico slika 5b. Da bi sila zaradi površinske napetosti delovala mora biti vsaj en del diska (rotorja) v stiku z zrakom, čeprav je večji del diska potopljen pod gladino površine vode v lončku. Površinska napetost vode ima še eno dobro lastnost, to je, da disk centrira na dno konkavne ploskve površine vode v lončku, kjer je tudi center rotirajoče se vode, tako, da se disk med vrtenjem ne odmika proti robu lončka. Ta efekt je nujno potreben za delovanje mikrofluidnega vibracijskega rotacijskega motorja z osjo, ki ga prikazuje slika 6.



Slika 5. Eksperiment z rotirajočim diskom pod gladino vode.

Zadnji eksperiment, ki smo ga izvedli je bil delujoči vibracijski mikrofluidni motor, ki je omogočal kotne hitrosti cca. -10 rd/s < ω < +10 rd/s, čeprav se je voda vrtela z večjo hitrostjo, to je cca. -15 rd/s < ω < +15 rd/s. Vzrok za ta pojav vidimo v tem, da se je del energije rotirajoče vode porabil, za premagovanje trenja v preprostih ležajih v spodnjem in zgornjem delu osi motorja, na katerega je prilepljen disk (rotor ali turbina) motorja. Skico opisanega motorja prikazuje slika 6. Os z diskom se je vrtela le, če smo gladino vode postavili natančno na višino diska, kjer je vsaj del zgornje ploskve diska imel stik z zrakom. S tem smo omogočili delovanje sil površinske napetosti na disk, ki je bil s tem centriran na sredino lončka, na dno konkavne ploskve površine vode v lončku, naravnost v sredino krožeče se vode v lončku. To je imelo blagodejen vpliv na zmanjšanje sil trenja v preprostem ležaju v zgornjem in spodnjem delu osi (luknjica v plastični foliji premera cca. 200 µm). Os je bila izdelana iz steklenega vlakna premera 125 µm. Lonček in vsi sestavni elementi motorja so bili pritrjeni ali lepljeni z UV lepilom. Z ustreznim vibracijskim vzbujanjem po drugih oseh bi bilo možno z zaprtim lončkom s tekočino takšen motor uporabljati tudi v horizontalni smeri.



Slika 6. Vibracijski mikrofluidni motor z osjo.

5 Zaključek

V članku je predstavljen razvoj različnih vibracijskih rotacijskih mikromotorjev v tekočini za namene izdelave mikrostrojev. Kontrolirane vibracije površine na kateri so nameščeni mikroobjekti in potopljeni v tekočino, ustvarijo gibanje tekočine, kar s pravilnim načrtovanjem omogoča vrtenje mikrostruktur. Hitrost vrtenja je odvisna od frekvence in amplitude vzbujanja vibracij, smer vrtenja pa od sinusnega/ kosinusnega vzbujanja po X-Y osi. Dosežene hitrosti vrtenja in sile vrtenja so zadostne za izkoriščanje kot pogonsko silo mikromotorjev. Izvedba predstavljenih mikromotorjev je trenutno še v fazi razvoja in testiranja različnih oblik ter še ni primerna za aplikativno rabo.

Literatura

[1] B. Jurado-Sánchez, M. Pacheco, R. Maria-Hormigos, A. Escarpa: Perspectives on Janus micromotors: Materials and applications. Applied Materials Today, str. 407–418, 2017.

[2] M. Guix, C.C. Mayorga-Martinez, and A. Merkoci: Nano/Micromotors in (Bio)chemical Science Applications. Chemical Reviews, str. 6285-6322, 2014.

[3] T. Hayakawa, S. Skuma, F. Arai: On-chip 3D rotation of oocyte based on a vibration-induced local whirling flow. Mycrosystems & Nanoengineering, 2015.

[4] M. Hagiwara, T. Kawahara, F. Arai: Local streamline generation by mechanical oscillation in a microfluidic chip for noncontact cell manipulations. Applied Physics Letters, št. 101, str. 074102, 2012.

[5] R. Šafarič, D. Lukman: One-finger gripper based on the variable van der Waals force used for a single nano/micro-sized object. Journal of micromechanics and microengineering, št. 24, str. 1-13, 2014.

[6] S. Uran, R. Šafarič, B. Bratina: Reliable and accurate release of micro-sized objects with a gripper that uses the capillary-force method. Micromachines, 2017.

Eksperimentalni haptični vmesnik za visoko kvalitetno upodabljanje haptičnih kod

Mitja Golob, Aleš Hace

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, Slovenija
E-pošta: mitja.golob@um.si, ales.hace@um.si

The Experimental Haptic Interface for High Quality Rendering of Haptic Codes

This paper describes the design of a novel, high-performance experimental linear haptic interface. The paper shortly presents the whole system. The main purpose of this paper is to show a comparison between conventional Proportional Integral (PI) force control and Sliding Mode Control (SMC) for the haptic interface. The experimental results confirm the validity of the proposed haptic interface design with the advanced SMC algorithm.

1 Uvod

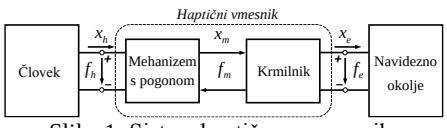
Haptični vmesnik je naprava, ki človeku omogoča interakcijo z navideznimi ali fizično oddaljenimi okolji. V splošnem so lahko haptični vmesniki vodeni kot impedančno ali admitančno. V admitančnem modelu haptične interakcije haptični vmesnik generira premike kot odziv na izmerjeno interakcijsko silo. V impedančnem modelu haptične interakcije haptični vmesnik generira sile kot odziv na vsiljen položaj oz. hitrost. Impedančna regulacijska shema, ki je pogosto uporabljena za vodenje vmesnikov človek-stroj zahteva poznavanje mehanske dinamike vmesnika vključno s trenjem. Slednje je velikokrat zelo težko točno modelirati, čeprav ga je po drugi strani treba za kvalitetno realizacijo želenega obnašanja popolnoma kompenzirati, kar moramo zagotoviti z ustrezno načrtovano regulacijsko shemo. Najbolj preprost pristop za zagotovitev ustreznega odziva haptičnega vmesnika je primer odprtozančnega vodenja. V tem primeru v regulacijsko shemo vodenja ne vključujemo povratne informacije o interakcijski sili. Natančnost vodenja sile je odvisna predvsem od dinamike haptičnega vmesnika, ki mora biti zanemarljiva v primerjavi z dinamiko modeliranega okolja. V tem primeru se trenje, gravitacija, vztrajnostne sile, modelne netočnosti in napake na pogonih neposredno prištevajo k silam, ki jih človek zaznava. Prednost takšnega pristopa je, da ni potreben senzor sile. Drugi pristop za zagotovitev ustrezne povratne sile je primer zaprtozančnega vodenja. V tem primeru v regulacijsko shemo vodenja vodi povratna informacija o sili. Bistvena prednost pri tem je možnost odprave sistemskih motenj, kar pomeni boljšo upodobitev povratne haptične informacije [1].

Poleg vodenja pa ima pomembno vlogo na kvaliteto haptičnega prikaza tudi strojna oprema. Treba je upoštevati, da lastna dinamika haptičnega vmesnika povzroča popačenje, kar občuti tudi uporabnik. To pomeni, da je treba zagotoviti haptični vmesnik z minimalno lastno impedanco, ob tem pa je treba zagotoviti še čim večjo pasovno širino, togost, ločljivost in ne nazadnje tudi sposobnost generiranja sil, ki odločilno vplivajo na kvaliteto haptičnega prikaza [2][3].

V nadaljevanju članka bomo tako predstavili eksperimentalni linearni haptični vmesnik, ki omogoča zelo kakovostno upodabljanje različnih haptičnih kod. Predstavili bomo posamezne gradnike kot celotno arhitekturo sistema vključno z vodenjem sistema. Na koncu bodo še predstavljeni eksperimentalni rezultati.

2 Opis haptičnega vmesnika

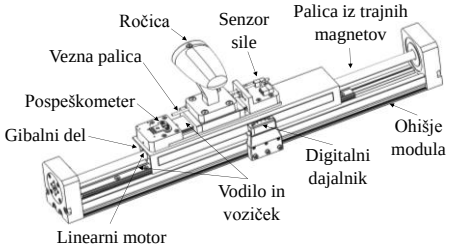
Haptični vmesnik v splošnem omogoča komunikacijo človek-stroj oz. navidezno okolje. Sistem lahko predstavimo s sliko 1, pri čemer pomenijo $x_{(.)}$ in $f_{(.)}$ položaj in silo človeka, haptičnega mehanizma in navideznega okolja. Haptični vmesnik torej z velikostjo in obliko povratne sile pri premikanju ročice uporabniku sporoča različne informacije v obliki t. i. haptičnih kod. Sistem linearnega haptičnega vmesnika sestavljajo: krmilnik, mehanizem s servo pogonom in ročica.



Slika 1: Sistem haptičnega vmesnika.

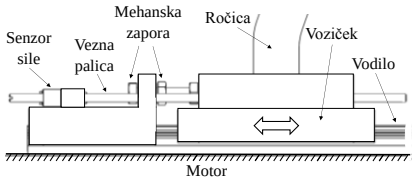
2.1 Haptični mehanizem

Haptični mehanizem z linearnim motorjem je prikazan na sliki 2. Linearni motor je vgrajen v modul, ki se sestoji iz kompaktnega ohišja, palice iz trajnih magnetov, ki je vpeta v ohišje modula, ter iz gibalnega dela linearnega motorja, tj. ohišje z navitjem, ki je povezano na voziček linearnega vodila. Vodilo slednjega pa je povezano na ohišje linearnega pogonskega modula. Na ohišje motorja je pritrjena ročica, s katero lahko uporabnik opravlja premike, pri čemer se generirajo predprogramirane haptične kode.



Slika 2: Zgradba linearnega haptičnega vmesnika.

Na linearnem pogonu haptičnega vmesnika so nameščeni senzorji za položaj (digitalni dajalnik), pospeškometer in senzor sile. Vgradnja senzorja sile predstavlja precejšen izziv, saj je treba poskrbeti za ustrezno mehansko zaščito pred preobremenjevanjem (slika 3). Apliciranih sil uporabnika namreč ni mogoče omejevati.

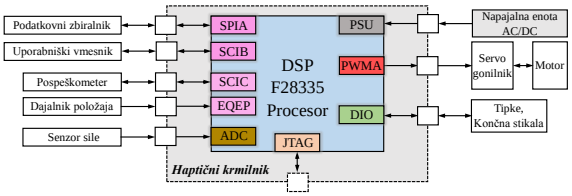


Slika 3: Vgradnja in izvedba mehanske zaščite senzorja sile.

Senzor sile je na eni strani togo povezan na ohišje motorja, medtem ko je na drugi strani z vezno palico, preko katere se prenaša aplicirana sila uporabnika, povezan na ročico. Ročica pa je togo povezana na precizno linearno vodilo oz. voziček. Precizno linearno vodilo je namenjeno zgolj kompenzaciji vpliva tistih sil in navorov, ki ne delujejo v delovni smeri. V delovni smeri je senzor sile zaščiten z mehansko zaporo, ki nudi obojestransko mehansko zaščito senzorja sile pred preobremenitvijo. Mehanska zapora namreč omejuje pomik vozička in s tem tudi odmik senzorja sile (odmik senzorja sile je proporcionalen pomiku vozička). Največji dovoljen odmik senzorja sile znaša 0.16 *mm*.

2.2 Haptični krmilnik

Haptični krmilnik sestavljajo digitalni signalni procesor TMS320F28335 družine Delfino, vhodno-izhodni in komunikacijski vmesniki ter napajalna enota [4]. Haptični krmilnik omogoča priklop digitalnih dajalnikov položaja, analognih in logičnih signalov, ter vključuje komunikacijske vmesnike USB, UART (RS232) in CAN, ki omogočajo povezavo z zunanjimi napravami. Logični signali, vključno z izhodi PWM, dopuščajo široko območje priključenih napetosti. Funkcionalna razširitev krmilnika je možna z notranjimi kom. vmesniki po protokolih I^2C, SPI ali SCI.



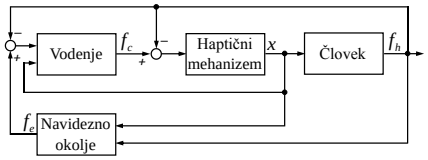
Slika 4: Blokovna shema haptičnega krmilnika s priključenimi zunanjimi enotami.

Blokovna shema na sliki 4 predstavlja haptični krmilnik s priključenimi zunanjimi enotami. Strojno opremo za zajem podatkov sestavljajo digitalni dajalnik položaja, digitalni pospeškometer in senzor sile z uporovnimi merilnimi lističi, medtem ko za sprotni prikaz oz. shranjevanje podatkov služita uporabniški vmesnik oz. podatkovni zbiralnik. Na PWM izhodno enoto je priključen servo gonilnik z linearnim motorjem.

3 Vodenje in zajem podatkov

3.1 Impedančni vmesnik

Slika 5 prikazuje shemo vodenja haptičnega vmesnika. V našem primeru enoprostostni linearni haptični vmesnik načrtujemo tako, da izkazuje impedančno obnašanje. To pomeni, da mora zagotavljati želeno impedanco, ki jo človek zaznava, ko premika ročico haptičnega vmesnika.



Slika 5: Blokovna shema vodenja haptičnega vmesnika.

V primeru, ko človek izvaja premik haptičnega vmesnika, na ta način posledično tudi aplicira določeno silo na ročico haptičnega vmesnika.

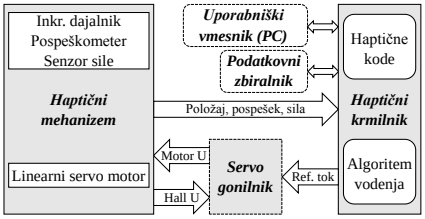
Tipični vhodi v impedančna navidezna okolja so položaj x, hitrost $\dot{x}$ in pospešek $\ddot{x}$ haptičnega vmesnika ter izmerjena sila f_h. Sila f_e, kot izhod navideznega okolja predstavlja referenčno silo za izbran algoritem vodenja, ki mora zagotoviti realizacijo želene impedance haptičnega vmesnika.

V tem članku bomo podali primerjavo dveh zaprtozančnih sistemov vodenja, in sicer: *a.)* klasična PI regulacije sile [5], *b.)* regulacijska shema izpeljana po postopku vodenja v drsnem režimu (SMC) [6]. Za takšno (robustno) vodenje je značilna visoka robustnost na motnje sistema (kompenzira nemodelirano dinamiko, kot je trenje, modelne netočnosti, dinamike višjega reda itd.). Pri uporabi klasične PI regulacije sile pogrešek sil f_e in f_h predstavlja vhod v regulator sile. V primeru uporabe robustne sheme vodenja pa je za estimacijo in kompenzacijo motnje potrebna še informacija o dejanski hitrosti $\dot{x}$. Izhod iz algoritma vodenja pa je regulacijska sila f_c, ki jo mora realizirati linearni pogon.

3.2 Zajem podatkov

Blokovna shema na sliki 6 predstavlja strukturo krmilja linearnega haptičnega vmesnika. Pri tem strojno opremo predstavljajo haptični mehanizem, haptični krmilnik, servo gonilnik, uporabniški vmesnik (osebni računalnik) ter podatkovni zbiralnik (ang. *Data Logger*).

Algoritem vodenja je v celoti implementiran na haptičnem krmilniku z DSP kontrolno enoto. Sprotno spremljanje in nadzor delovanja haptičnega vmesnika pa poteka v celoti preko uporabniškega vmesnika, ki se izvaja na osebnem računalniku (PC). Zaradi dokaj počasne komunikacije z uporabniškim vmesnikom (100 *ms*) in omejene količine pomnilnika RAM (68 *kB*) na haptičnem krmilniku oz. DSP kontrolni enoti smo dodali še enoto za zbiranje podatkov (Nexys 2 Digilent s FPGA vezjem). Slednja ima na razpolago 16 *MB* pomnilnika RAM, v katerega lahko shranjujemo podatke. Te podatke lahko nato na zahtevo prenesemo na osebni računalnik preko vmesnika USB.

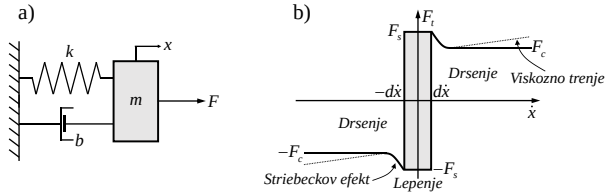


Slika 6: Struktura krmilja linearnega haptičnega vmesnika.

Merilni sistem se sestoji iz naslednjih sklopov:
- *Digitalni inkrementalni dajalnik* omogoča natančno meritev položaja ročice haptičnega vmesnika. Optični dajalnik omogoča ločljivost 0.1 μm. Za estimacijo hitrosti je uporabljena MT metoda, ki omogoča natančno merjenje tudi nizkih hitrosti v razredu nižjem, kot je en pulz na tipalni interval.
- *Digitalni MEMS 3-osni senzor pospeška* omogoča meritev pospeška ročice haptičnega vmesnika. Merimo pospeške v območju ±2 g z 10-bitno natančnostjo in frekvenco zajemanja 1600 *Hz*.
- *Senzor sile* z uporovnimi merilnimi lističi omogoča meritev tako tlačnih kot nateznih sil v območju ±55 *N*. Izhodni napetostni signal je ojačan in nato pretvorjen v digitalno obliko z 12-bitnim A/D pretvornikom.

4 Haptično prikazovanje

V predhodnih poglavjih smo pokazali tako posamezne gradnike kot celotno arhitekturo sistema haptičnega vmesnika. Pomemben preizkus haptičnega vmesnika in algoritmov vodenja pa predstavlja, kako realistično lahko prikažemo različne haptične kode. Treba vedeti, da pasovna širina, togost, ločljivost in generirana sila skupaj odločilno vplivajo na kvaliteto haptičnega prikaza. V nadaljevanju bomo zato, predhodno predstavljen sistem ovrednotili s pomočjo namensko izbranih modelov haptičnih kod: *masa-dušilnik-vzmet* in *trenje*.



Slika 7: Modeli haptičnih kod: a) model masa-dušilnik-vzmet in b) Karnoppov model trenja.

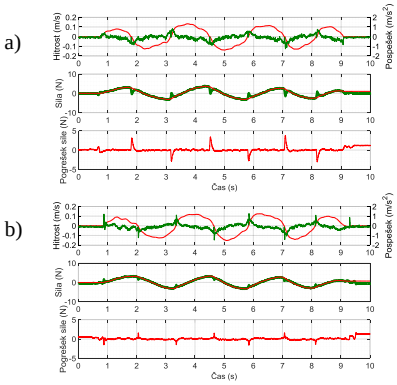
- *Model masa-dušilnik-vzmet* na sliki 7a predstavlja pogost način preizkusa kvalitete haptičnega prikaza. Model je podan s $F = m\ddot{x} + b\dot{x} + kx$, pri čemer so *m, b, k, x* masa, koef. viskoznega trenja in togosti ter položaj [3].
- *Model trenja.* Različni modeli trenja se razlikujejo po načinu opisa, kompleksnosti, natančnosti, itd. V našem primeru smo uporabili Karnoppov model trenja, ki je prikazan na sliki 7b in vključuje tako statično trenje F_s kot dinamično trenje F_c, odvisno od hitrosti $\dot{x}$. Prednost uporabe Karnoppovega modela trenja je v vpeljavi umetno vsiljenega intervala drsne hitrosti $\dot{x} = 0$ pri absolutni vrednosti drsne hitrosti manjše od $d\dot{x}$ (na ta način obidemo težavo nezveznosti in singularnosti). V praksi pa je potrebno v zahtevi, da trenje izkazuje impedančno kavzalnost, slednje zasnovati tako, da je sila trenja aktivna zgolj v fazi premika haptičnega vmesnika [7]. Slednje lahko dosežemo z uporabo navidezne vzmeti v modelu trenja. V primeru majhnih pomikov od območja lepenja x_l, se generira sila trenja $F_t \approx -k_t\,(x-x_l)$, ki je proporcionalna dejanskemu odmiku *x*.

5 Eksperimentalni sistem

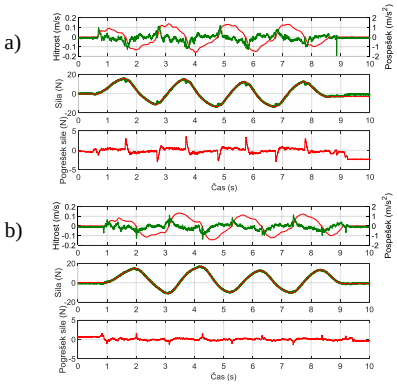
Linearni motor, ki je vgrajen v haptični vmesnik omogoča visoko dinamiko in nima samodržne sile (ang. *Cogging Force*). Poganja ga zanj prirejen servo gonilnik, ki deluje v tokovnem načinu. Programska oprema haptičnega krmilnika s frekvenco izvajanja regulacijske zanke 10 *kHz* je bila zasnovana v MATLAB/Simulinku.

5.1 Eksperimentalni rezultati

Izvedeni eksperimenti so razdeljeni v dva sklopa. Prvi sklop rezultatov predstavlja emulacijo navideznega dušenja in togosti (slika 8 in 9) ter emulacijo navidezne mase (slika 10 in 11). Drugi sklop rezultatov pa predstavlja emulacijo trenja (slika 12 in 13). Pri vsakem eksperimentu so podani rezultati za primer uporabe tako klasičnega PI vodenja (slika 8-13a), kot za primer uporabe naprednega robustnega vodenja (slika 8-13b). Na slikah 8-13 so prikazane: dejanska hitrost (rdeča krivulja) in dejanski pospešek (zelena krivulja) na zgornjem grafu, referenčna sila (rdeča krivulja) in dejanska sila (zelena krivulja) na sredinskem grafu ter razlika med referenčno in dejansko silo – pogrešek sile (rdeča krivulja) na spodnjem grafu.

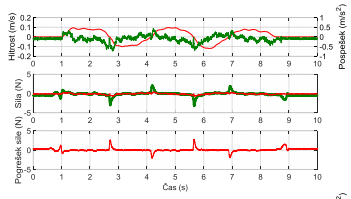


Slika 8: Emulacija *dušenja in togosti* – primer 1 (haptični parametri: *m = 0 kg, b = 10 Ns/m, k = 50 N/m*).

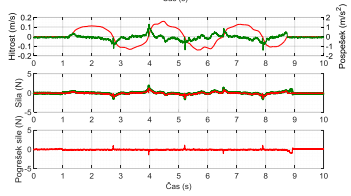


Slika 9: Emulacija *dušenja in togosti* – primer 2 (haptični parametri: *m = 0 kg, b = 10 Ns/m, k = 300 N/m*).

a)

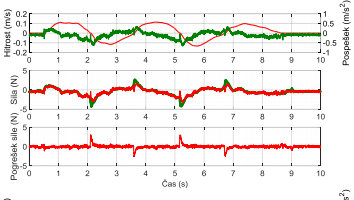


b)

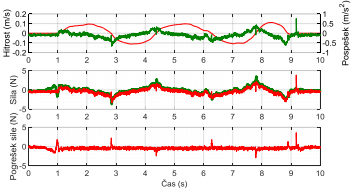


Slika 10: Emulacija *mase* – primer 1 (haptični parametri: *m = 1 kg*, *b = 0 Ns/m*, *k = 0 N/m*).

a)

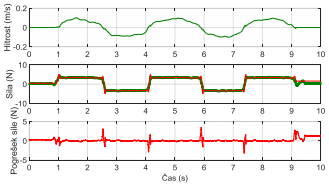


b)

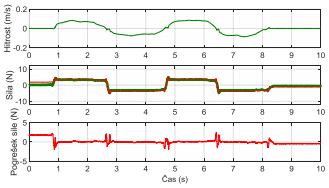


Slika 11: Emulacija *mase* – primer 2 (haptični parametri: *m* = 5 *kg*, *b = 0 Ns/m*, *k = 0 N/m*).

a)

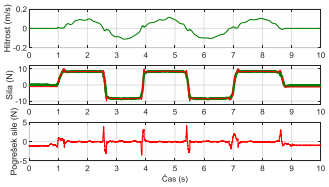


b)

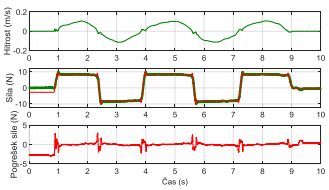


Slika 12: Emulacija *trenja* – primer 1 (haptični parametri: $F_s = 5$ *N*, $F_c = 3$ *N*, *B = 5 Ns/m*, $d\dot{x} = 5$ *mm/s*).

a)



b)



Slika 13: Emulacija *trenja* – primer 2 (haptični parametri: $F_s = 10$ *N*, $F_c = 8$ *N*, *B = 5 Ns/m*, $d\dot{x} = 5$ *mm/s*).

Pri emulaciji dušenja in togosti lahko vidimo, da je pogrešek sile pri obeh strategijah vodenja v primeru enakomernega pomika majhen. Večji pogrešek sile pa se pojavi pri PI algoritmu vodenja (*slika 8a in 9a*) v primeru spremembe smeri gibanja, takrat namreč dejanska sila (zelena krivulja) ne uspe slediti referenčni sili (rdeča krivulja). V primeru izbranega robustnega algoritma vodenja (*slika 8b in 9b*) pa slednji zelo dobro kompenzira vpliv trenja, zato je tudi pogrešek sile manjši. Podobno je tudi v primeru emulacije mase, kjer PI algoritem vodenja (*slika 10a in 11a*) v primeru spremembe smeri gibanja ne zmore dovolj dobro kompenzirati vpliva trenja, zato se pojavi določen pogrešek sile v trenutku spremembe smeri gibanja. V primeru izbranega robustnega algoritma vodenja (*slika 10b in 11b*) pa je pogrešek sile manjši. V primeru emulacije trenja je razvidno, da sta tako PI algoritem vodenja (*slika 12a in 13a*), kot robustni algoritem vodenja (*slika 12b in 13b*) zagotavljala dobro sledenje referenčni sili v primeru aktivnega dinamičnega trenja, medtem ko v trenutku nastopa statičnega trenja boljše sledenje ref. sili zagotavlja robustno vodenje.

6 Zaključek

V članku smo predstavili eksperimentalni sistem linearnega haptičnega vmesnika. Pokazali smo tako posamezne gradnike kot celotno arhitekturo sistema haptičnega vmesnika. Na kratko je bilo predstavljeno tudi vodenje in zajem vseh potrebnih podatkov. Eksperimentalni rezultati so pokazali, da je možno z ustrezno strojno implementacijo na eni strani in na drugi strani z vključitvijo robustnega vodenja doseči bolj kvalitetno upodobitev različnih haptičnih kod.

7 Literatura

[1] Ueberle, M.W.: Design, Control, and Evaluation of a Family of Kinesthetic Haptic Interfaces. PhD thesis, Technische Universität München (2006).

[2] T.A. Kern (ed.): Engineering Haptic Devices, Springer - Verlag Berlin Heidelberg (2009).

[3] M. Mihelj, J. Podobnik: Haptics for Virtual Reality and Teleoperation. Intelligent Systems, Control and Automation: Science and Engineering, vol. 64, Springer (2012).

[4] M. Čurkovič, R. Pučko, M. Golob, A. Hace: Dvoosni krmilnik za hitro eksperimentiranje v robotskih aplikacijah, UM-FERI, AIG'17, Maribor, Slovenija (6. – 7. april 2017).

[5] T. Sirithanapipat: Haptic Interface Control Design for Performance and Stability Robustness. PhD thesis, Nashville, Tennessee (May, 2006).

[6] A. Hace, M. Golob: Chattering-Free Sliding Mode Control Algorithm for a Haptic Throttle Lever, IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy (23-26 Oct. 2016).

[7] C. Richard: On the Identification and Haptic Display of Friction. PhD thesis (August 2000).

Generalizirani robotski algoritmi za specifične funkcije pri kuhanju - rezanje

Jan Vodopivec[1], Matej Amon[2], Sebastjan Šlajpah[1], Marko Munih[1], Matjaž Mihelj[1]

[1]Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška c. 25, 1000 Ljubljana
[2] BSH Hišni aparati d.o.o. Nazarje, Savinjska cesta 30, 3331 Nazarje
E-pošta: jv8760@student.uni-lj.si

Generalized robotic algorithms in cooking – food slicing

Abstract. *In this paper we present one of the possibilities of using robots in kitchen. In collaboration with BSH Hišni aparati we focused on slicing different types of food with a robot. Motion capture system was used to track human motion and JR3 sensors to measure forces and torques while slicing different types of food. Captured trajectories were processed and implemented on a Motoman MH5 robot. Minimum-jerk-based interpolation was used in combination with impedance control. The robot successfully repeated slicing of food.*

1 Uvod

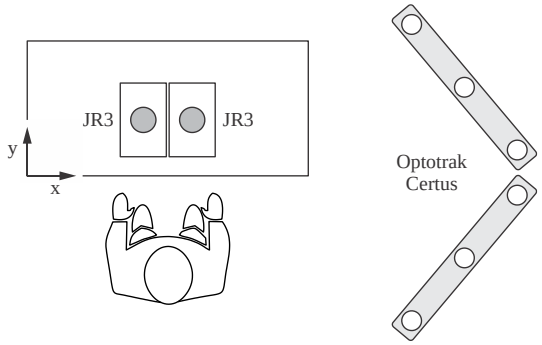
Družba zadnja leta stremi k avtomatizaciji in robotizaciji na praktično vseh področjih. V trgovinah uporabljamo samopostrežne blagajne, avtomobilska podjetja razvijajo samovozeča vozila, pri rehabilitaciji pomagajo robotski mehanizmi, travo lahko kosi robotska kosilnica, kmalu pa bodo verjetno v uporabo prišli tudi roboti, ki bodo pomagali pri raznih hišnih opravilih in v kuhinji. Podjetja, kot je Moley Robotics, že razvijajo kuhinjske pomočnike, ki naj bi nas ob kuhanju opazovali, nato pa znali sami poustvariti te jedi [1].

V sodelovanju s podjetjem BSH Hišni aparati smo na primeru rezanje hrane preverili možnost uporabe robotov pri kuhinjskih opravilih. Kuharski strokovnjak naj bi predstavil pravilne tehnike rezanja, pridobljene podatke pa bi uporabili za vodenje robota, da bi ta lahko samostojno narezal različne vrste živil. Poudarek je torej predvsem na zajemu in analizi trajektorij pri človeškem rezanju različnih vrst hrane ter implementaciji naloge na robotu.

2 Metodologija

2.1 Merjenje aktivnosti človeka

Meritve človekovih aktivnosti so bile osnova za načrtovanje trajektorij in analizo sil, ki nastopajo pri rezanju. Večina raziskav se osredotoča predvsem na sekanje hrane, vendar s tem poškodujemo hrano, rez pa ni čist. Zato ljudje uporabljamo tudi strig, torej vlečenje noža pri rezanju, na ta način pa se tudi snov lažje reže [2]. Pri pravilnem rezanju nastopata zato dve komponenti sile, normalna in tangencialna na površino rezane snovi.
Za meritve sta bila uporabljena dva senzorja sil in navorov JR3 50M31A-I25, sistem za merjenje gibanja Optotrak Certus in sedem inercialnih merilnih enot. Podatki so bili beleženi s frekvenco 80 Hz.



Slika 1. Postavitev merilnega mesta z označenim koordinatnim sistemom, pogled od zgoraj.

Optotrak Certus kameri sta bili postavljeni na desni strani delovnega mesta, saj je merjenec nož med rezanjem držal v desni roki. Koordinatni sistem je bil poravnan z robovi mize, os *z* je bila usmerjena pravokotno iz mize (Slika 1).



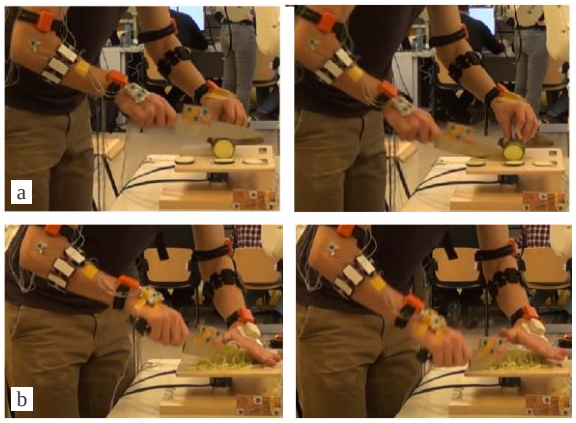
Slika 2. Postavitev markerjev in merilnih inercialnih enot. Levo postavitev na merjencu, desno zgoraj postavitev skupka markerjev na deski, desno spodaj postavitev skupka na nožu ter prikaz obeh desk, pod katerimi sta JR3 senzorja.

Na merjenca je bilo nameščenih 14 optičnih markerjev. Iz skupkov treh markerjev določimo pozicijo in orientacijo ploskve v prostoru, na kateri so nameščeni. Skupki so bili nameščeni na podstavku za desko, na nožu, dlani in na hrbtu, po en marker pa na komolcu in rami (Slika 2). Inercialne merilne enote so bile postavljene na dlani, podlakti in nadlakti obeh rok, še ena pa je bila na prsnici. Senzorja sil in navorov sta bila postavljena pod dvema rezalnima deskama (Slika 2), s ciljem ločitve sil, s katero pritiskamo ob držanju hrane in s katero potiskamo nož in režemo. Na levi deski smo hrano držali, na desni pa rezali.

2.2 Analiza trajektorij

Rezali smo več različnih živil (Slika 3): bučke, jabolko, por, korenje in banano. Živila smo rezali na kolobarje in rezine, por pa smo tudi sekljali.

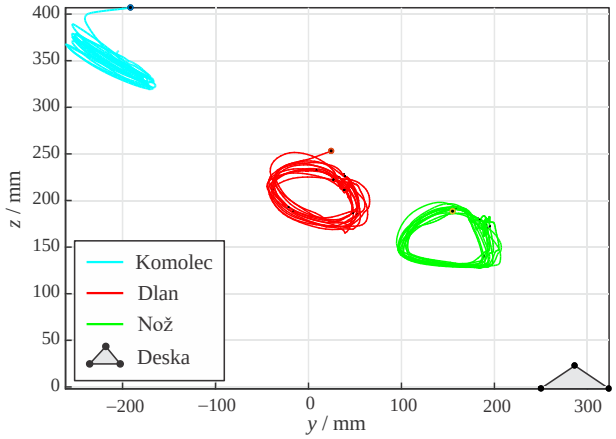
2.2.1 Rezanje bučke na kolobarje



Slika 3. Primeri rezanja: a) rezanje bučke na kolobarje in b) sekljanje pora.

Trajektorije pri rezanju so na Sliki 4 prikazane v ravnini $y-z$, za lepši prikaz je iz skupkov treh markerjev na dlani in nožu prikazan samo po eden. Pri rezanju se je komolec premikal približno premo naprej in nazaj, z dlanjo smo med posameznimi rezi delali krožne gibe, nož pa je pol reza opravil skupaj z dlanjo in naredil polkrog, drugo polovico reza pa se je premikal premo, saj se je z delom rezila dotikal deske. Položaj markerja na rami je v prostoru miroval, zato na grafu ni prikazan.

Potrebno je poudariti, da navkljub trudu, da bi rezali tako, kot je pokazal šolani kuhar, tehnika rezanja še vedno ni bila povsem pravilna. V našem primeru smo rezali navpično navzdol in izvajali večjo silo v smeri osi z, nato pa naredili poteg v smeri y, namesto da bi cel rez izvedli s potegom noža v smeri y.

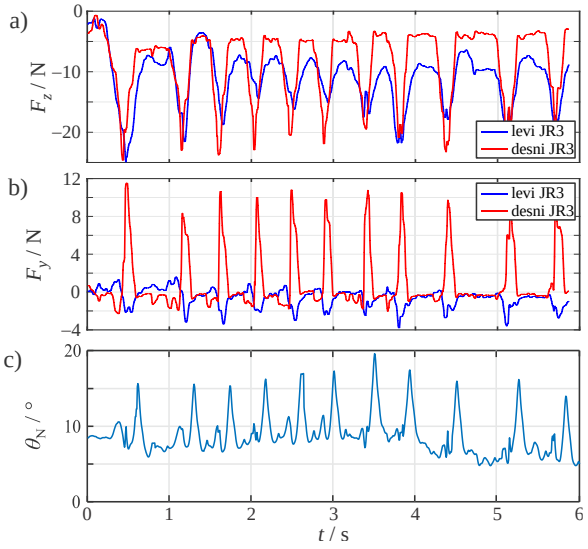


Slika 4. Trajektorije pri rezanju bučke na kolobarje v ravnini $y-z$.

Bučko smo držali na levi deski, z nožem pa rezali na desni deski. Sila F_z z levega senzorja je obarvana z modro, sila desnega pa z rdečo (Slika 5a in 5b). Sile se preko bučke prenašajo z ene deske na drugo. Tako že pri samem držanju bučke silo zazna tudi desni senzor, medtem ko se pri rezanju sila prenese tudi na levi senzor. Za optimalno meritev bo potrebno nož opremiti s senzorjem sil in navorov.

Normalna sila je delovala na ploščo v nasprotni smeri osi z, zato je bila njena vrednost negativna. V nadaljevanju bomo govorili o večjih in manjših silah, pri čemer bo govora o absolutnih vrednostih sile v posamezni smeri. Sila F_z v normalni smeri med posameznim rezom narašča do maksimalne vrednosti. Takrat je bila bučka prerezana in po dvigu noža se sila zmanjšuje (Slika 5a). Vsota sil obeh senzorjev je enaka teži živila in sili, s katero smo pritiskali na desko med držanjem živila.

Med rezanjem smo opazovali tudi tangencialno silo F_y, ki deluje v smeri potega noža (Slika 5b). Zaradi prej opisane napačne tehnike rezanja se tangencialna sila pojavi šele proti koncu reza, saj smo šele takrat naredili poteg z nožem, nato narašča do konca potega, ko pa nož dvigujemo z deske, zopet pade na prvotno vrednost. Ob pravilnem rezanju naj bi se sicer normalna in tangencialna sila pojavili skupaj.



Slika 5. a) Sili F_z, b) sili F_y in c) kot med nožem in mizo θ_N pri rezanju bučke na kolobarje.

Graf na Sliki 5c prikazuje kot noža glede na mizo. Med rezanjem živila je kot približno konstanten, ob koncu reza pa smo nož privzdignili in s tem za nekaj stopinj povečali kot. Pri pripravi na nov rez smo nož zopet postavili v začetni položaj, kot pa je bil približno konstanten do novega dviga.

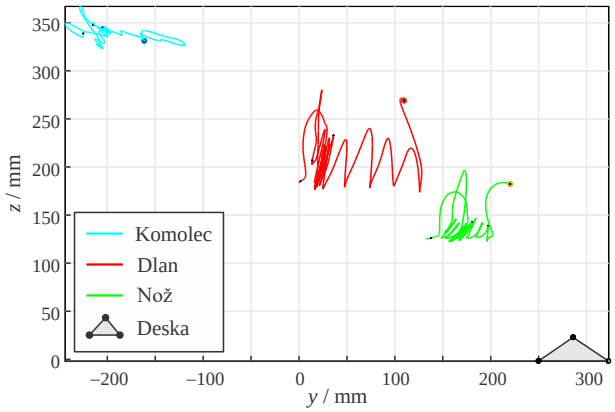
Rezali smo tudi jabolko na rezine, rezultati pa so podobni kot pri rezanju bučke.

2.2.2 Sekljanje pora

Pri sekljanju pora (Slika 3b) smo uporabljali samo desno desko, saj smo eno roko položili na zgornji rob rezila, z drugo pa držali za ročaj.

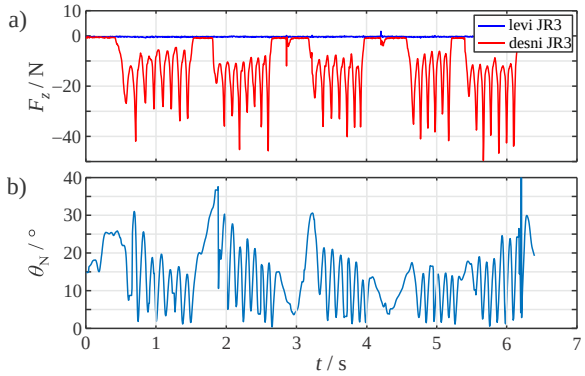
Prikaz trajektorij na Sliki 6 je v tem primeru omejen na 13 ponovitev giba. Med sekljanjem smo enkrat

dvignili nož in ga ponovno postavili na desko. Opazimo, da imamo povsem drugačne trajektorije kot pri rezanju bučke; dlan in nož premikamo po manjšem območju.



Slika 6. Trajektorije pri sekljanju pora v ravnini $y - z$.

Kot se relativno glede na silo F_z zopet spreminja podobno kot pri prejšnjem primeru (Slika 7). Ko nož pritisnemo ob desko in s tem odsekamo del pora, sila naraste na maksimalno vrednost, kot pa se v tistem trenutku približa 0°, saj je nož skoraj poravnan z desko. Ko nož dvigamo, se sila zmanjša, kot pa se poveča.



Slika 7. a) Sila F_z in b) kot med nožem in mizo θ_N pri sekljanju pora.

2.3 Načrtovanje trajektorij za robota

Ker naj bi robot premikal nož podobno kot človek pri rezanju oziroma sekljanju, smo se osredotočili na trajektorije noža. Preizkusili smo rezanje na kolobarje, kjer so trajektorije noža podobne elipsam.

Izmerjeno trajektorijo smo interpolirali, ker so bili markerji občasno zakriti. Nato smo trajektorijo (Slika 4, zelena trajektorija), razdelili na posamezne reze in pogledali, kateri rez je bil najpočasnejši, torej je imel največ vzorcev. Vsem ostalim rezom smo nato s kubično interpolacijo z zlepki povečali število vzorcev, da se je ujemalo z najpočasnejšim. Iz posameznih rezov smo izračunali povprečno trajektorijo, na enak način pa hkrati tudi pripadajoče kote pri rezih. Kot rezultat postopka smo pridobili povprečno trajektorijo s pripadajočim povprečnim kotom noža v danem trenutku glede na mizo. Temu smo dodali še gib v stran po vsakem posameznem rezu, s tem smo dobili gibanje še vzdolž živila, rezultat rezanja pa so na ta način kolobarji.

Število točk za zapis trajektorije je lahko majhno, pomembno je le zajeti spremembe smeri trajektorije. Med posameznimi točkami robotski krmilnik izvede interpolacijo. Uporabljena je zaprtozančna oblika metode minimalnega časovnega odvoda pospeška (»*minimum-jerk trajectory*«) [3]. Enačba (1) opisuje omenjeno metodo v prostoru stanj:

$$\begin{bmatrix} \dot{x}_r \\ \ddot{x}_r \\ \dddot{x}_r \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\frac{60}{D^3} & -\frac{36}{D^2} & -\frac{9}{D} \end{bmatrix} \begin{bmatrix} x_r \\ \dot{x}_r \\ \ddot{x}_r \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{60}{D^3} \end{bmatrix} x_{f,} \qquad (1)$$

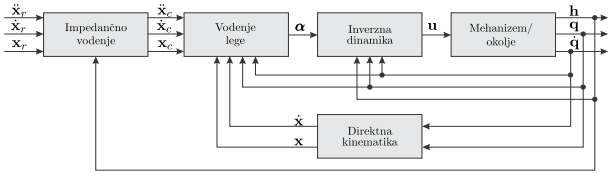
kjer sta z x_r označeni točka na interpolirali trajektoriji in x_f želena lega robota na osnovi meritev človeka, $D = t_f - t$ pa je razlika med časom, ko naj bi robot dosegel točko x_f in trenutnim časom. Interpolacija zagotavlja gladko trajektorijo, s spreminjanjem časa t_f pa je mogoče vplivati na hitrost izvajanja naloge.

2.4 Vodenje robota

Rezanje je bilo izvedeno z robotom Motoman MH5 s šestimi prostostnimi stopnjami, nosilnostjo 5 kg in ponovljivostjo ±0,02 mm (Slika 8). Izdelali smo prijemalo za nož, s katerim smo nož pritrdili na senzor, ki je bil na vrhu robota. Implementirano je bilo impedančno vodenje (Sliki 9), trajektorija pa je bila podana kot referenca v zunanjih koordinatah. Želeni položaj in orientacija x_r skupaj z merjenima silo in navorom dotika h so vhodne vrednosti v impedančno enačbo, ki generira želeno lego podajnosti x_c, uporabljeno kot reference za vodenje gibanja robota. Lega podajnosti je enaka želeni legi x_r, če sta sila in navor dotika enaka nič, sicer pa se trajektorija prilagodi glede na izmerjene sile in navore ter impedančne parametre. Prednost impedančnega vodenja je ravno vnos podajnosti pri sledenju lege. V primeru, da želimo povsem podajen mehanizem, lahko nastavimo parameter togosti K_p na 0. Z večanjem togosti K_p povečujemo togost robota in posledično natančnost sledenja želeni trajektoriji tudi pod vplivom zunanjih sil.



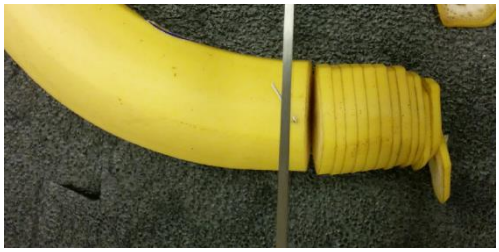
Slika 8. Robot Motoman MH5 s pritrjenim nožem.



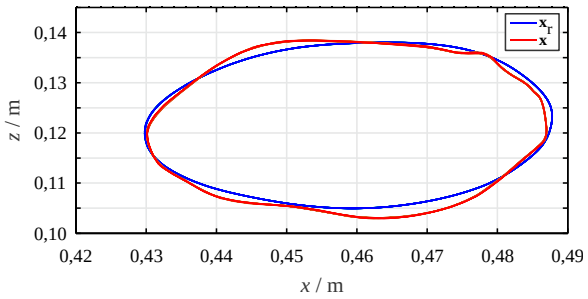
Slika 9. Vodenje na osnovi aktivne impedance [4].

3 Rezultati rezanja z robotom

Z robotom smo preizkusili rezati banano, kot referenčno trajektorijo pa smo uporabili trajektorijo rezanja bučke na kolobarje. Banana je bila uspešno narezana na želene 3 mm debele kolobarje (Slika 10), kar dokazuje, da lahko robot uspešno izvede nalogo.



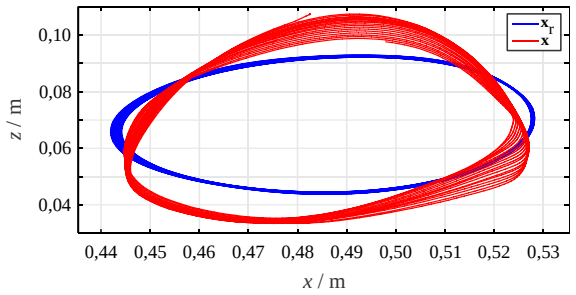
Slika 10. Banana, rezana z robotom.



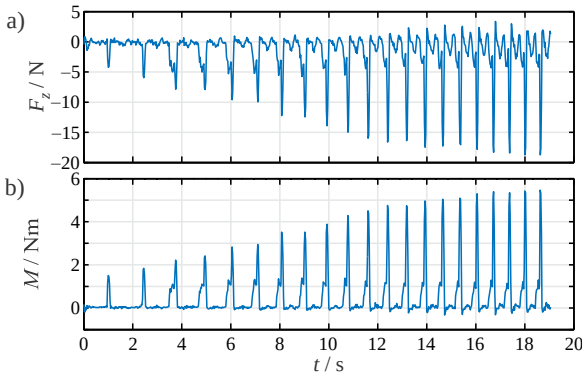
Slika 11. Robotsko sledenje trajektoriji: z modro je označena referenca, z rdečo pozicija robota.

Na Sliki 11 sta prikazani referenčna in dejanska trajektorija pri rezanju z robotom. Sledenje ni povsem natančno, kar je posledica podajnosti impedančnega vodenja. Zaradi tega lahko z nožem zadenemo tudi v rezalno desko, sistem pa je še vedno stabilen.

Preizkusili smo tudi rezanje, pri katerem je robot povečeval svojo hitrost in rezal vse hitreje. Posledica impedančnega vodenja je v tem primeru odstopanje dejanske trajektorije od referenčne (Slika 12). S pospeševanjem sta se večala tudi sila F_z (Slika 13a) in navor (Slika 13b), ki sta sicer s parametrom K_p do neke mere nastavljiva.



Slika 12. Rezanje z večanjem hitrosti robota.



Slika 13. a) Sila F_z in b) navor M med rezanjem z večanjem hitrosti robota.

4 Zaključek

Z meritvami aktivnosti človeka smo pridobili uporabne podatke, ki smo jih analizirali in implementirali na robotu. Robot reže z zadovoljivo natančnostjo in zanesljivo.

Projekt je mogoče nadaljevati, saj ostaja še ogromno nerešenih izzivov, zaradi katerih roboti še niso primerni kuhinjski pomočniki. Optimizirali bi lahko merjenje človeškega rezanja, predvsem z zmanjšanjem števila merilnih naprav. Koristno bi bilo tudi merjenje sile preko noža, ter s tem pridobitev dejanske sile, s katerimi nož reže. Po optimizaciji merjenja bi bilo potrebno meritve ponoviti s kuharskim mojstrom, ki bi rezal na različne načine. Kot primer, rezanje surovega mesa je povsem drugačno od rezanja na kolobarje, kar bi bilo potrebno upoštevati in najti ustrezno rešitev za implementacijo na robotu. Izboljšali bi lahko tudi metode pridobivanja trajektorij in izvajanja na robotu, na primer z dinamičnimi primitivi gibanja DMP (ang. Dynamic movement primitives) [5], s čimer bi imeli trajektorije opisane z enakim številom parametrov, ne glede na čas izvajanja.

Literatura

[1] Moley - The world's first robotic kitchen [Online]. Dosegljivo: http://www.moley.com/. [Dostopano 11. 1. 2018].

[2] E. Reyssat et al., “Slicing softly with shear,” *Physical review letters*, vol. 109, št. 24, str. 244301, 2012.

[3] S. Šlajpah. »Vodenje štirih robotov v arhitekturi dvoročnega teleoperacijskega sistema.« Univerzitetno diplomsko delo, Fakulteta za Elektrotehniko, Univerza v Ljubljani, Ljubljana. 2011.

[4] M. Mihelj, T. Bajd, M. Munih, *Vodenje robotov*, Ljubljana: Založba FE in FRI, 2011.

[5] Stefan Schaal, "Dynamic Movement Primitives - A Framework for Motor Control in Humans and Humanoid Robotics" *Adaptive Motion of Animals and Machines*, str. 261-280, 2006.

Haptično vstavljanje igle z uporabo teleoperacije

**Žiga Bizjak, Tim Kambič, Robi Ravnikar, Damjan Panić,
Janez Podobnik, Matjaž Mihelj, Marko Munih, Sebastjan Šlajpah**

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

E-mail: zb3444@student.uni-lj.si, tk2777@student.uni-lj.si, robi.ravnikar@gmail.com, dp0678@student.uni-lj.si, janez.podobnik@fe.uni-lj.si, matjaz.mihelj@fe.uni-lj.si, marko.munih@fe.uni-lj.si, sebastjan.slajpah@fe.uni-lj.si

Teleoperated needle insertion

Haptic needle insertion is one of the key components in haptic teleoperation surgery. In this paper we present the use of one Omega.7 (master device) and three Phantom robots connected with a triangle platform (slave device) to insert a needle into lower arm. Pseudoadmittance control was implemented to control the pose of the slave device and also to provide feedback to master device. In addition, an augmented reality was used to further enhance the performance of the operator.

The application was divided into six phases: initialisation, moving slave device from the workspace, selecting working point, moving slave robot to that working point, hold robots, and teleoperation. The performance of the presented system was tested on a mockup (artificial arm with integrated tube representing vein) by successfully extracting fluid from artificial vein.

1 Uvod

Uporaba teleoperacijsko vodenih robotov v medicini je zelo pomembna, saj nam omogoča izvajanje preciznih gibov, ki presegajo zmožnosti kirurga. Z robotom pa lahko uporabnikove gibe ustrezno zmanjšamo, omejimo delovno območje, zmanjšamo nevarnost postoperativnih zapletov ter posledično vplivamo na samo trajanje hospitalizacije po operaciji. Vstavljanje igle je naloga, ki prikazuje uporabnost robotskega sistema v medicini. Na temo robotov v medicini je objavljenih kar nekaj del [1, 2, 3].

Teleoperacija predstavlja *delo na daljavo*, kjer tako *delo* kot *razdalja* nista natančno določena. Razdalja se lahko se nanaša na fizično razdaljo, kjer je operater ločen od robota z veliko razdaljo, lahko pa pomeni tudi razliko v merilu, kot na primer pri kirurških mikromanipulatorjih za operacije na mikroskopski ravni. Teleoperacijski sistem je sestavljen iz treh delov: upravljalne naprave, izvršne naprave ter regulatorja. Primer teleoperacijskega sistema v medicini je kirurški sistem DaVinci [4].

Uporaba teleoperacijskih sistemov je smotrna v primeru, kjer je manipulacija nevarna za človeška življenja ali pa fizične zahteve za manipulacijo presegajo človeške zmožnosti. Teleoperacija se uporablja v vesoljskih in vojaških aplikacijah, varnostnih sistemih, podvodnih vozilih in v medicini [5].

Robotski sistemi in slikovne tehnologije so revolucionirale medicinsko področje. Kljub naprednku pa algoritmi umetne inteligence zaenkrat še niso sposobni človeške razsodnosti in intuicije, zato je pomembno sodelovanje kirurga z robotom. Tako izvedeni posegi so manj invazivni ter pozitivno vplivajo na postoperativne procese.

V tem delu bomo predstavili aplikacijo za teleoperacijsko robotsko vstavljanje igle v žilo. Operater preko haptične povratne zanke dobi informacijo o kontaktu igle in žile, kot pomoč operaterju pa je v aplikacijo vključena tudi obogatena resničnost.

2 Metodologija

Teleoperacijski sistem sestavljata upravljalna in izvršna naprava. Upravljalno napravo predstavlja haptični robot Omega.7 (Force Dimension, Švica). Robot ima 7 prostostnih stopenj (tri pozicije, tri orientacije ter še dodatno prijemalo na vrhu). Robot ima aktivno vodeno pozicijo in prijemalo, medtem ko je orientacija vrha robota zgolj pasivna. Robot lahko generira haptično povratno informacijo samo v obliki sil v smeri x, y in z.

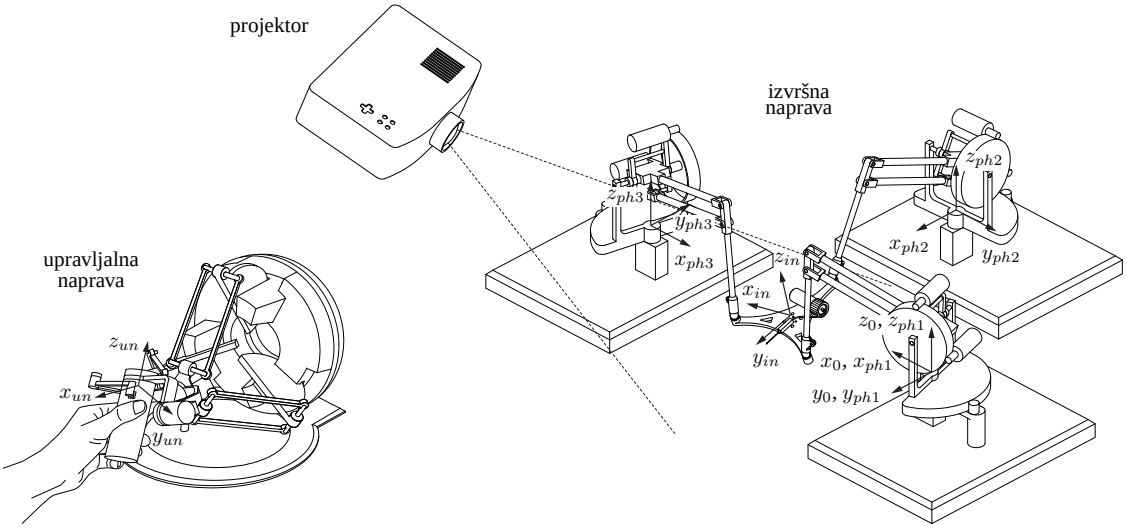
Izvršna naprava je sestavljena iz treh robotov Phantom, ki so preko trikotne plošče povezani v paralelni mehanizem (glej sliko 1). Vsak robot ima tri prostostne stopnje, medtem ko ima plošča šest prostostnih stopenj. Roboti so zaradi varnosti na ploščo pritrjeni preko magnetnih sferičnih sklepov. Dodaten nivo varnosti pa predstavlja varnostno stikalo, ki mora biti pritisnjeno celoten čas izvajanja gibanja.

V sistem je vključen še projektor NEC VT570 z ločljivostjo 1024×768 slikovnih pik. Služi kot prikazovalnik za obogateno resničnost. Izvršna naprava je opremljena s kirurško kanilo, za simulacijo pacientove roke pa je uporabljena plastična roka opremljena s cevko, ki predstavlja žilo.

Celotno vodenje je narejeno v programskem paketu Matlab/Simulink, vizualizacija pa je realizirana s programsko knjižnico OpenCV. Vzorčna frekvenca robotskega sistema znaša 5 kHz.

2.1 Direktna in inverzna kinematika izvršne naprave

Globalni koordinatni sistem x_0–y_0–z_0 izvršne naprave (plošče) sovpada s koordinatnim sistemom robota Phantom 1. Homogene transformacijske matrike $\mathbf{H}_1$, $\mathbf{H}_2$ in $\mathbf{H}_3$, ki



Slika 1: Shema sistema (upravljalna in izvršna naprava) z označenimi koordinatnimi sistemi

povezujejo globalni koordinatni sistem plošče z globalnimi koordinatnimi sistemi posameznih Phantom robotov (enačba (1)), smo določili na podlagi znane konfiguracije sistema.

$$\left[\begin{array}{cc} \mathbf{p}_{ph}^{\mathrm{T}} & 1 \end{array}\right]^{\mathrm{T}} = \mathbf{H} \left[\begin{array}{cc} {}^{ph}\mathbf{p}_{ph}^{\mathrm{T}} & 1 \end{array}\right]^{\mathrm{T}} \tag{1}$$

Phantom roboti, pritrjeni na ploščo, tvorijo enakostranični trikotnik. Določanje direktne kinematike plošče temelji tako na računanju geometrije trikotnika. Vrh izvršne naprave je definiran v središču plošče. Pozicija vrha $\mathbf{p}_{in}$ je določena z enačbo (2).

$$\mathbf{p}_{in} = \frac{\mathbf{p}_{ph1} + \mathbf{p}_{ph2} + \mathbf{p}_{ph3}}{3} \tag{2}$$

Vektorji $\mathbf{p}_{ph1}$, $\mathbf{p}_{ph2}$ in $\mathbf{p}_{ph3}$ so pozicije vrhov posameznih robotov Phantom izražene v globalnem koordinatnem sistemu (slika 1). Orientacija vrha izvršne naprave $\mathbf{R}_{in}$ je določena s sistemom enačb (3)–(6).

$$\mathbf{x}_R = \frac{\mathbf{p}_{ph2} - \mathbf{p}_{ph1}}{|\mathbf{p}_{ph2} - \mathbf{p}_{ph1}|} \tag{3}$$

$$\mathbf{z}_R = \mathbf{x}_R \times \frac{\mathbf{p}_{ph1} - \mathbf{p}_{ph3}}{|\mathbf{p}_{ph1} - \mathbf{p}_{ph3}|} \tag{4}$$

$$\mathbf{y}_R = \mathbf{z}_R \times \mathbf{x}_R \tag{5}$$

$$\mathbf{R}_{in} = \left[\begin{array}{ccc} \mathbf{x}_R & \mathbf{y}_R & \mathbf{z}_R \end{array}\right] \tag{6}$$

Inverzna kinematika izvršne naprave ni direktno vezana na sklepne spremenljivke robotov Phantom, ampak se prevede na določitev pozicije vrha posameznega robota Phantom. Glede na geometrijo (plošča je enakostranični trikotnik s stranico $a = 150$ mm) se jih določi kot

$$\mathbf{p}_{ph1} = \mathbf{p}_{in} + \frac{a\sqrt{3}}{2}\mathbf{R}_{in}\left[\begin{array}{ccc} -\frac{1}{2} & \frac{1}{3} & 0 \end{array}\right]^{\mathrm{T}} \tag{7}$$

$$\mathbf{p}_{ph2} = \mathbf{p}_{in} + \frac{a\sqrt{3}}{2}\mathbf{R}_{in}\left[\begin{array}{ccc} \frac{1}{2} & \frac{1}{3} & 0 \end{array}\right]^{\mathrm{T}} \tag{8}$$

$$\mathbf{p}_{ph3} = \mathbf{p}_{in} + \frac{a\sqrt{3}}{2}\mathbf{R}_{in}\left[\begin{array}{ccc} 0 & \frac{2}{3} & 0 \end{array}\right]^{\mathrm{T}} . \tag{9}$$

Dobljene pozicije vrhov robotov Phantom smo z uporabo matrik $\mathbf{H}_1$, $\mathbf{H}_2$ in $\mathbf{H}_3$ pretvorili iz globalnega koordinatnega sistema izvršne naprave v globalne koordinatne sisteme posameznih robotov Phantom (enačba (10)).

$$\left[\begin{array}{cc} {}^{ph}\mathbf{p}_{ph}^{\mathrm{T}} & 1 \end{array}\right]^{\mathrm{T}} = \mathbf{H}^{-1}\left[\begin{array}{cc} \mathbf{p}_{ph}^{\mathrm{T}} & 1 \end{array}\right]^{\mathrm{T}} \tag{10}$$

2.2 Regulacija lege izvršne naprave

Za znano lego izvršne naprave lahko določimo željene položaje robotov Phantom (${}^{ph1}\mathbf{p}_{r,ph1}$, ${}^{ph2}\mathbf{p}_{r,ph2}$, ${}^{ph3}\mathbf{p}_{r,ph3}$, enačbe (7)–(9)). Za regulacijo položaja vrha Phantom robotov smo implementirali proporcionalno-diferencirni (PD) regulator v zunanjih koordinatah na osnovi transponirane Jacobijeve matrike $\mathbf{J}_{ph}^{\mathrm{T}}$. PD regulator lahko zapišemo z

$$\mathbf{F}_{ph} = \mathbf{K}_p\left({}^{ph}\mathbf{p}_{r,ph} - {}^{ph}\mathbf{p}_{ph}\right) - \mathbf{K}_d{}^{ph}\dot{\mathbf{p}}_{ph} \tag{11}$$

$$\boldsymbol{\tau}_{ph} = \mathbf{J}_{ph}^{\mathrm{T}}\mathbf{F}_{ph} , \tag{12}$$

kjer sta $\mathbf{K}_p$ in $\mathbf{K}_d$ proporcionalno in diferencirno ojačenje regulatorja (določena eksperimentalno), ${}^{ph}\dot{\mathbf{p}}_{ph}$ hitrost vrha robota Phantom, $\boldsymbol{\tau}_{ph}$ pa določeni navori za posamezne motorje.

2.3 Psevdoadmitančno vodenje

Teleoperacija na podlagi psevdoadmitance [6] temelji na uporabi virtualnega posrednika. Pri tem izvršna naprava sledi virtualnemu posredniku, sama izvršna naprava pa na upravljalno enoto vpliva preko haptične povratne povezave. Virtualni posrednik se giblje kot funkcija razlike med pozicijama upravljalne ($\mathbf{p}_{un}$) in izvršne ($\mathbf{p}_{in}$) naprave.

Za vodenje izvršne naprave določimo silo virtualnega posrednika $\mathbf{F}_p$ z uporabo PD regulatorja

$$\mathbf{F}_p = \mathbf{K}_{pp}(\mathbf{p}_{un} - \mathbf{p}_{in}) + \mathbf{K}_{dp}\dot{\mathbf{p}}_{un} . \tag{13}$$

Z uporabo admitančnega modela zapisanega v Laplaceovem prostoru

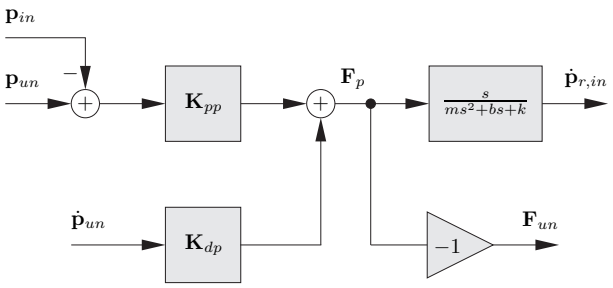
$$\frac{\dot{\mathbf{p}}}{\mathbf{F}} = \frac{s}{ms^2 + bs + k} \tag{14}$$

lahko določimo hitrost virtualnega posrednika $\dot{\mathbf{p}}_p$

$$\dot{\mathbf{p}}_p = \frac{s}{ms^2 + bs + k}\mathbf{F}_p\,. \qquad (15)$$

Pri tem veličine m, b in k opisujejo dinamične parametre sistema (maso, dušenje in togost). Hitrost virtualnega posrednika $\dot{\mathbf{p}}_p$ predstavlja tudi referenčno hitrost izvršne naprave $\dot{\mathbf{p}}_{r,in} = \dot{\mathbf{p}}_p$, po integraciji pa tudi željeno lego $\mathbf{p}_{r,in} = \int \dot{\mathbf{p}}_{r,in}\mathrm{d}t$.

Sila, ki deluje na virtualnega posrednika $\mathbf{F}_p$, nam predstavlja tudi nasprotno povratno silo, s katero izvršna naprava vpliva na upravljalno enoto $\mathbf{F}_{un} = -\mathbf{F}_p$. Celoten model vodenja je prikazan na sliki 2.



Slika 2: Model teleoperacijskega sistema na podlagi psevdoadmitance s haptično povratno zanko

2.4 Gibanje upravljalne naprave

Pri zagonu aplikacije kot tudi med samo aplikacijo je potrebno upravljalno napravo postaviti v želeno lego. Ker ima upravljalna naprava aktivne samo pozicije, rotacije pa so pasivne, lahko vodimo samo pozicijo vrha $\mathbf{p}_{un}$. Trajektorijo gibanja generiramo po metodi minimalnega časovnega odvoda pospeška [7]. Spremembo pospeška upravljalne naprave $\dddot{\mathbf{p}}_{un}$ lahko zapišemo z enačbo (16)

$$\dddot{\mathbf{p}}_{un} = -\frac{9}{D}\ddot{\mathbf{p}}_{un} - \frac{36}{D^2}\dot{\mathbf{p}}_{un} + \frac{60}{D^3}(\mathbf{p}_{f,un} - \mathbf{p}_{un})\,, \qquad (16)$$

kjer je $\mathbf{p}_{f,un}$ končna pozicija upravljalne naprave, D pa razlika med določenim časom izvedbe naloge t_f in trenutnim že pretečenim časom t ($D = t_f - t$). Željeno lego upravljalne naprave določimo s trojno integracijo odvoda pospeška $\mathbf{p}_{r,un} = \iiint \dddot{\mathbf{p}}_{un}\mathrm{d}t\,\mathrm{d}t\,\mathrm{d}t$.

Z implementacijo PD regulatorja izračunamo potrebno silo $\mathbf{F}_{un}$, da premaknemo upravljalno napravo v želeno lego

$$\mathbf{F}_{un} = \mathbf{K}_{pun}(\mathbf{p}_{r,un} - \mathbf{p}_{un}) - \mathbf{K}_{dun}\dot{\mathbf{p}}_{un}\,. \qquad (17)$$

To silo pošljemo na krmilnik robota Omega, ki poskrbi za ustrezen premik robota.

3 Aplikacija

Aplikacija robotskega vstavljanja igle v žilo je razdeljena na šest stanj. Diagram prehajanja stanj je prikazan na sliki 3.

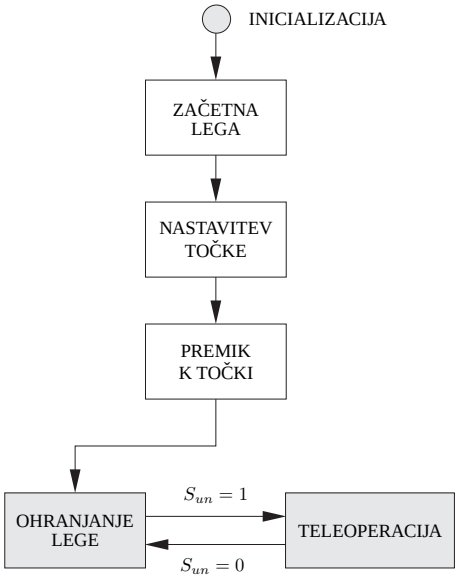
INICIALIZACIJA: Na začetku aplikacije se tako upravljalna kot izvršna naprava postavita v svoji začetni legi. Ko manipulatorja dosežeta začetni legi, se sistem avtomatsko premakne v stanje *začetna lega*.

ZAČETNA LEGA: V tem stanju se izvršna naprava odmakne na rob delovnega prostora. S tem omogočimo, da lahko bolnik varno namesti roko v za to predvideno delovno območje. Upravljalna naprava v tem primeru sledi gibanju izvšne naprave. Končno lego ter hitrost izvršne naprave se lahko prilagodi glede na potrebe aplikacije. Ko izvršna naprava doseže željeno lego, se aktivira stanje *nastavitev točke*.

NASTAVITEV TOČKE: To stanje vključuje obogateno resničnost. Z upravljalno napravo premikamo virtualno točko (rdečo piko), ki se jo s projektorjem projecira v delovni prostor. Z postavitvijo pike določimo grobo pozicijo delovne točke oziroma žile. Ker je premikanje točke zgolj ravninski problem, so premiki upravljalne naprave možni zgolj v ravnini (premiki v vertikalni smeri so onemogočeni). Naslednje stanje se aktivira, ko uporabnik zapre prijemalo na vrhu upravljalne naprave.

PREMIK K TOČKI: V tem stanju se vrh izvršne naprave (igla) premakne nad prej določeno delovno točko. Izvršna naprava se premika zgolj po horizontalni ravnini. Ker smo v prejšnjem stanju upravljalno napravo uporabili za nastavitev delovne točke, je potrebno po končanem gibanju izvršne naprave upravljalno napravo postaviti v začetno lego, legi obeh naprav pa ustrezno resetirati in poravnati. Ko vrh izvršne naprave doseže delovno točke, se aktivira stanje *ohranjanje lege*.

OHRANJANJE LEGE: To stanje je pogojeno z uporabo prijemala na vrhu upravljalne naprave. Le-to je uporabljeno kot stikalo: ko je prijemalo odprto, je signal $S_{un} = 0$, ko pa je prijemalo zaprto, je signal $S_{un} = 1$.



Slika 3: Diagram prehanja stanj pri izvajanju naloge. Med stanjema *ohranjanje lege* in *teleoperacija* preklapljamo s prijemalom na vrhu upravljalne naprave (odprto prijemalo $S_{un} = 0$ oziroma zaprto prijemalo $S_{un} = 1$).

To stanje je aktivno ob odprtem prijemalu. V tem stanju tako upravljalna kot izvršna naprava držita trenutni legi. Uporabi se ga takrat, ko je premikanje izvršne naprave nezaželjeno, npr. ko je igla vstavljena v žilo.

TELEOPERACIJA: Ob preklopu prijemala izvršne naprave v $S_{un} = 1$ se aktivira teleoperacijsko vodenje. V tem primeru je lega vrha upravljalne naprave posredna referenca za gibanje izvršne naprave, kot je to opisano v poglavju 2.3. Vodimo lahko tako pozicijo kot orientacijo izvršne naprave, preko haptične povratne informacije pa uporabnik tudi čuti sile interakcije izvršne naprave. To stanje se uporablja za izvedbo ustrezne manipulacije igle (vstavljanje igle v žilo, odstranitev igle). Z uporabo prijemala se lahko preklaplja med tem stanjem in stanjem *ohranjanje lege*.

Za namen aplikacije smo pripravili tudi preprost uporabniški grafični umestnik, ki omogoča zagon in ustavitev aplikacije.

4 Razprava

Aplikacijo smo testirali s pomočjo umetne roke in tekočine, ki je predstavljala umetno kri. Prvi operater je skrbel za zagon aplikacije in za nadzor varnostnega stikala na robotih Phantom, drugi operater pa je preko robota Omega vodil izvajanje teleoperacije. Na sliki 4 je prikazana uspešna izvedba aplikacije vstavitve igle v roko in pretok tekočine v posodo, nameščeno na platformi.

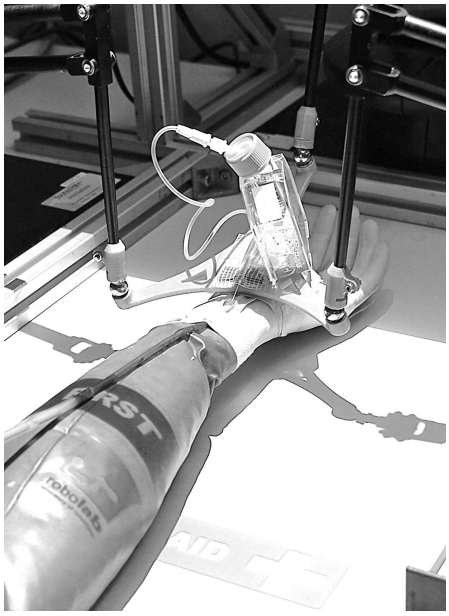
Prednost arhitekture teleoperacije je, da lahko operater spremeni merilo premikov, kar pomaga pri natančnosti vstavljanja igle. V aplikaciji lahko nastavljamo merilo premikov tako za rotacijske kot za translacijske premike.

V testnih pogojih je celoten proces vstavljanja igle trajal okoli 30 sekund. Samo izvajanje bi sicer lahko pohitrili, ampak smo zaradi varnostnih razlogov pustili daljše čase izvajanja določenih stanj.

Medicinski roboti verjetno ne bodo nikoli prevzeli nalog zdravstvenega osebja, vendar pa so jim lahko v veliko pomoč. S pravo uporabo aplikacije lahko rešimo veliko problemov v moderni medicini. Robotsko vstavljanje igle bi lahko uporabili v kontaminiranih področjih, v področjih kjer medicinsko osebje fizično ne more dostopati do bolnika ali pa bi lahko aplikacijo razšili na uporabo pri operacijah. Primer uporabe bi lahko bil vstavljanje infuzije kesonskim delavcem, ki delajo globoko pod morjem na velikih pritiskih, kjer zdravstveno osebje nima dostopa. Primer uporabe je tudi bolnik v karanteni, kjer ne želimo dodatno izpostavljati zdravstvenih delavcev. V tem primeru bi lahko vbrizgali zdravila ali vzeli kri z robotom.

Zaradi uporabe umetne roke in žile, bi za realno uporabo morali nekatere parametre (predvsem za togost sistema) na novo oceniti in preizkusti. Stanja smo določili z ozirom na varnost bolnika, kar posledično pomeni, da je večina gibov izvedenih z majhno hitrostjo. Grafični vsmesnik za operaterja je zelo enostaven in omogoča le vklop in izkop sistema, kar pomeni, da uporabnik ne potrebuje dodatnega izobraževanja za uporabo aplikacije.

Za to aplikacijo je bila uporabljena generična oblika



Slika 4: Slika uspešnega testa na umetni roki. Na sliki vidimo realizacijo postavitve igle ter ciljne posode za tekočino. Na roki je vidna tudi umetna žila, s pomočjo katere smo tekočino pripeljali do mesta vboda.

platforme. Za realno aplikacijo bi bilo potrebno realizirati namenskega robota za izvajanje medicinskih operacij. Prav tako bi bilo potrebno prilagoditi tudi ostale sklope sistema. Čeprav je ta aplikacija zgolj testiranje principa delovanja, predstavlja osnovo za prihodnje aplikacije namenjene v medicinske namene in aplikacije, kjer je teleoperacije potrebna za uspešno izvedbo.

Literatura

[1] F. T. Cullen, *Technology and Terrorism*. Routledge, 2017.

[2] P. F. Hokayem and M. W. Spong, "Bilateral teleoperation: An historical survey," *Automatica*, vol. 42, no. 12, pp. 2035–2057, 2006.

[3] N. Abolhassani, R. Patel, and M. Moallem, "Needle insertion into soft tissue: A survey," *Medical Engineering and Physics*, vol. 29, no. 4, pp. 413–431, 2007.

[4] G. H. Ballantyne and F. Moll, "The da Vinci telerobotic surgical system: the virtual operative field and telepresence surgery," *Surgical Clinics of North America*, vol. 83, no. 6, pp. 1293–1304, 2003.

[5] S. Lichiardopol, "A survey on teleoperation," *Technische Universitat Eindhoven, DCT report*, 2007.

[6] J. J. Abbott and A. M. Okamura, "Pseudo-admittance bilateral telemanipulation with guidance virtual fixtures," *The International Journal of Robotics Research*, vol. 26, no. 8, pp. 865–884, 2007.

[7] A. Piazzi and A. Visioli, "Global minimum-jerk trajectory planning of robot manipulators," *IEEE Transactions on Industrial Electronics*, vol. 47, no. 1, pp. 140–149, 2000.

Design and preliminary testing of a pneumatic exoskeleton for walking assistance

Marko Jamšek[1,2], **Jan Babič**[1]

[1]*Laboratory for Neuromechanics and Biorobotics, Department of Automation, Biocybernetics and Robotics*
Jožef Stefan Institute, Ljubljana, Slovenia
[2]*Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia*
E-mail: marko.jamsek@ijs.si

Abstract

Tripping is one of the major causes of falls in elderly people. Injuries that originate from falls usually have severe consequences on their quality of life. The best solution for this problem is prevention of such traumatic events which proves to be a challenge in the current state of elderly care. This is why we are in an urgent need of developing new solutions that could help in the prevention of falls and fall related injuries in this ever increasing part of our entire population. One possible preventive solution could be the use of assistive devices such as exoskeletons. In this work we present a design of a pneumatic exoskeleton for walking assistance primarily designed as a platform to study active control solutions for preventing falls caused by tripping. We also present a preliminary evaluation of the exoskeleton capabilities on affecting the gait characteristics of a subject during walking.

1 Introduction

Population ageing is one of the many global issues we are facing. In 2012, 11.5 % of the global population was older than 60 years and by 2050 this percentage is projected to increase to 21.8 % [1]. This is a problem because elderly people usually require more assistance for even simple everyday tasks and they are more prone to injuries. One of the big causes of injuries in elderly people are falls which usually have severe consequences on their quality of life [2]. And one of the main causes of falls is tripping [3, 4, 5]. The mechanisms of tripping have been widely explored in the past [6, 7]. In order to prevent a fall caused by a trip the support limbs have to counteract the angular momentum gained by the body during impact with the obstacle. This can be achieved with the adaptation of different recovery strategies. In the work of Eng et al. [8] two such strategies were described which depended on the part of the swing cycle the perturbation occurred. If the perturbation occurred in late swing, a lowering strategy of the swing leg was adopted followed by a step of the contralateral limb to overtake the obstacle. When perturbations occurred in early swing phase (at approximately 20 % of the swing phase) an elevating strategy was adopted by the subjects. This comprised of a flexion of the swing limb and extension of the stance limb. This strategy was further investigated in the works of Pijnappels et. al. [9, 10]. Here more focus was put towards the support limb and its contribution in the control of angular momentum after tripping. It was shown that the support limb provided subjects with more time and clearance for proper positioning of the recovery limb. Additionally the support limb also restrained or reduced the angular momentum ob the body acquired when tripping. Further work with elderly people showed that some individuals cannot properly reduce the angular momentum after a trip [11, 12]. This was due to a lower rate of change of moment generation in all support limb joints. One possible solution for this could be the use of an assistive exoskeleton to provide additional torque to counteract the uncompensated forward momentum. Only one similar device was successfully presented in the work of Monaco et. al. [13] where an assistive exoskeleton was used to help regain stability of a human after a slip. In this work we present the design and preliminary testing of an exoskeleton that will provide a platform to explore possibilities of assistance in tripping scenarios. Additionally we present an early investigation in the effects of the exoskeleton on the kinematics of human gait.

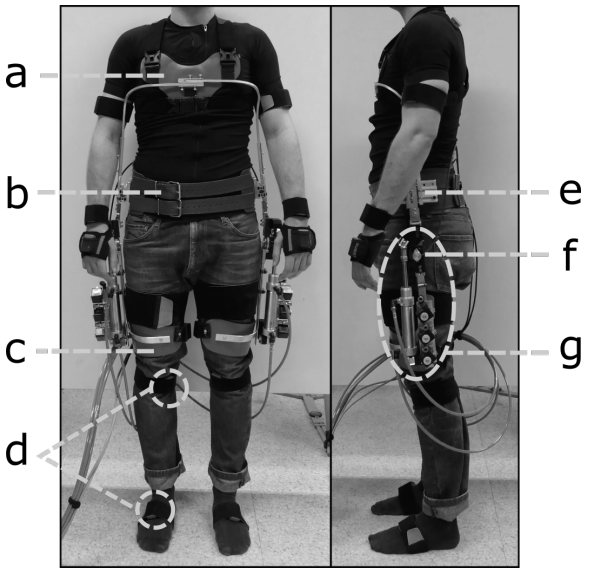


Figure 1: Subject wearing the exoskeleton. Main components: a) chest strap, b) hip strap, c) leg strap, d) Xsens sensors, e) adjustable hip mount, f) encoder, g) pneumatic actuator.

2 Design and experiment

We developed a bilateral powered pneumatic exoskeleton. The main mechanical components are presented in Figure 1. They include a belt, chest strap, leg strap, adjustable hip mount and an actuation mechanism on each side. The actuation mechanism consists of a bidirectional pneumatic cylinder and a rotational joint (This is more clearly visible in the schematic diagram in Figure 3). The Xsens sensors (Xsens, Enschede, Netherlands) are not an integral part of the exoskeleton but were used in this work to evaluate the effects of the exoskeleton on the human kinematics.

2.1 Actuation mechanism

The applied pressure in the cylinder chambers produces torque at the hip to aid in the flexion or extension of the legs. Due to the constraints of the mechanism's kinematic chain the torque in the hip joints can be calculated based on the angle of the rotational joint and the applied pressure. The angle is measured by rotational encoders mounted in these joints. Whereas the pressure in the cylinders is measured with the feedback loop provided by the valves used for controlling the cylinders.

2.1.1 Calculations of insertion points

One of the most important design aspects of such an actuation mechanism are the insertion points of the cylinder. These define the maximum torque produced at the rotational joint and the range of motion of the mechanism. An analytical formulation difficult to achieve for such a problem. This is why a numerical approach was used to calculate the optimal solution. At first, the requirements for the range of motion were set to $-20°$ to $110°$. Then different cylinders were taken into account from a list of possible configurations of length and cylinder width. Cylinders of same lengths with different diameters provide more force, but are also heavier. Our optimization procedure included weight as a parameter for finding the best possible solution. After the mathematical calculations we designed a 3D model of the actuation mechanism to check for mechanical constraints that would limit the construction of the mechanism (thickness of the rotational joint, encoder mount, etc.). The model is presented in Figure 2. After minor modifications of the insertion points placement in the modelling environment, the parts for the fixation of the cylinders were 3d printed. This enabled us to achieve the desired mechanism characteristics.

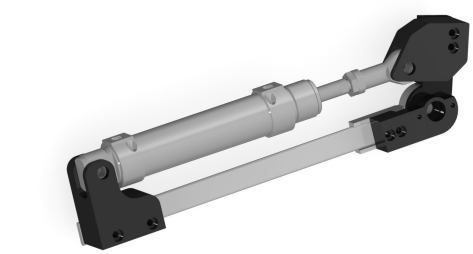


Figure 2: 3D model of the actuation mechanism.

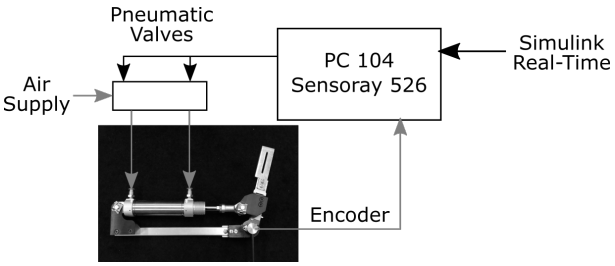


Figure 3: Schematic diagram for the exoskeleton control.

2.2 Control

The schematic diagram for the exoskeleton control is presented in Figure 3. Each actuation mechanism is controlled with two pneumatic valves (Norgren VP50-10SBJ1-11H00, Norgren Ltd, Blenheim Way, Fradley Park, Lichfield) which enable precise control of the pressure in both cylinder chambers. These valves are controlled via analog signals (0-10 V) and provide an analog output (0-10 V) for feedback of the current pressure. The controller was developed in Simulink Real-Time (Mathworks, Natick, MA, United States) and runs on a PC 104 with added connectivity of a Sensoray 526 card (SENSORAY, 7313 SW Tech Center Dr., Tigard, OR 97223).

2.2.1 Modes of operation

Our exoskeleton actuation mechanism and controller enables us to use different modes of operation. The exoskeleton can be operated in a torque control mode or position control mode. In torque mode the pressure in the cylinders is controlled so that for the given position of the mechanism the torque produced by the mechanism remains constant. A position control scheme was also tested where the pressure in the cylinders was controlled to position the mechanism at a given angle. However, this mode of operation was not used in this work.

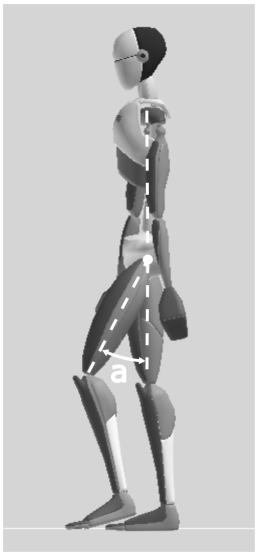


Figure 4: Human model in MVN Analyze software. a) hip angle of left leg

2.3 Experiment

For preliminary testing of the exoskeleton capabilities we tested the system on one subject. During normal walking with a constant speed on a level floor the left actuator was activated to aid in the flexion of the left leg (4 bar input pressure in the bottom chamber). The activation occurred in mid swing phase at a 10 degree flexion of the left leg (measured with the encoder) and was randomly enabled or disabled. The duration of the activation was set to 200 ms. During the experiment, kinematic data of the subject were recorded using the Xsens Awinda motion capture system (Xsens, Enschede, Netherlands). Figure 4 shows the human model during the experiment.

Ad-hoc post-processing of the acquired data was performed in Matlab (Mathworks, Natick, MA, United States). Data from the controller (encoder angles, pressure in cylinder chambers) and data exported from MVN Analyze (human joint angles) was synchronized and combined.

3 Results

One of the most critical aspects of control is timing. In our set-up we analysed the weakest link, to asses the capabilities of the exoskeleton. One of the most important characteristics for the actuation mechanism is the delay of the pneumatic valves after an input signal. There are two key temporal characteristics that are of interest in our system. The delay time (the time it takes for the output to reach 10 % of the end value) and the response time (the time it takes for the output to reach 90 % of the end value) of the pneumatic valves. We measured this based on the mean and standard deviation of 20 responses to the 4 bar step input signal used in our experiment. For the pneumatic actuation mechanism used in this experiment we measured a delay time t_d of 35 ms and a response time t_r of 100 ms. The graph is presented in Figure 6.

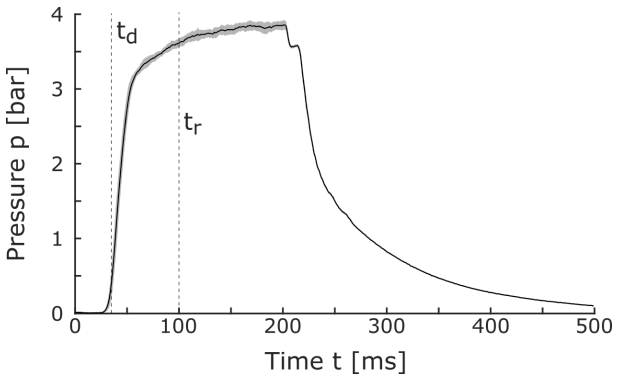


Figure 6: Mean and standard deviation of 20 responses to the 4 bar step signal. t_d marks the delay and t_r marks the response time.

Data from the Xsens MVN Analyze software was used to evaluate the effect of the exoskeleton on gait kinemat-

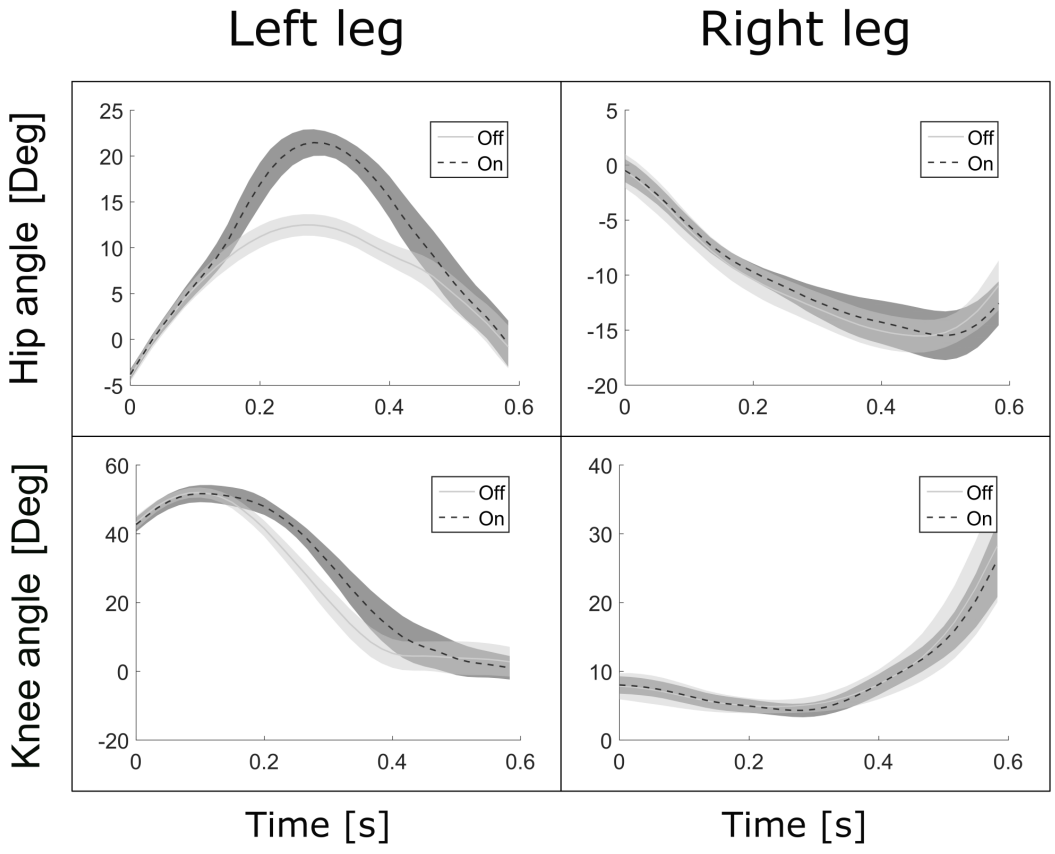


Figure 5: Joint angles for flexion/extension at the hip and knee joint.

ics. To achieve this we first had to synchronize various steps in time space. We did this by selecting points in the gait cycle when the value of the left leg hip angle crossed the threshold of -5 degrees with a positive gradient. This threshold corresponds to the activation of the exoskeleton but can be used to compare steps when the exoskeleton was disabled as it marks the same period of the gait cycle. This points established the start of a 600 ms time window for each step. We randomly selected 20 of this time windows for when the exoskeleton was activated or deactivated to compare the effects of the exoskeleton. For these two sets of steps we calculated the mean and standard deviation of the hip and knee angles for flexion/extension of both legs. We present this data graphically in Figure 5. Time zero represents the time when the exoskeleton was activated. As mentioned before, the activation occurred only for the left leg when it was in the swing phase. It can be clearly seen that there is a significant change in the hip joint angle for the left leg. A smaller change is also noticeable for the left knee angle. If we compare the delay from Figure 6 to the delay in Figure 5 the latter appears longer. This is due to the mass inertia of the legs as the applied torque in the joint needs to accelerate the leg in order for a change in the hip angle to be visible. On the right hand side graphs we observe that there was no significant effect on the hip and knee angles of the right leg. This was expected as the right exoskeleton actuator remained passive throughout the experiment. However this information is useful as it shows, that we can compare two separate sets of steps in the experiment.

4 Conclusion

The presented exoskeleton prototype is capable of inducing significant changes in the gait cycle in a controlled manner. This will enable further experiments and validation of various control algorithms for walking assistance or other tasks such as forward bending, squatting etc. A more precise measuring of the system parameters would be needed (delay and response times at different pressures and for different volumes) in order to conclude if the exoskeleton is viable for tripping prevention i.e. if it could react fast enough. This will be addressed in future work, where the exoskeleton design and experimental setup will be further optimized.

Acknowledgement

This work was supported by the European Union's Horizon 2020 through the SPEXOR project (contract no. 687-662); AnDy project (contract nr. 731540); and by the Slovenian Research Agency (research core funding no. P2-0076). The authors would like to thank Rok Goljat for developing the pneumatics control setup.

5 References

[1] United Nations Population Fund (UNFPA) and and HelpAge Internationa, *Ageing in the Twenty-First Century: A Celebration and A Challenge*, 2012. [Online]. Available: https://www.unfpa.org/sites/default/files/pub-pdf/Ageingreport.pdf

[2] M. C. Nevitt, S. R. Cummings, and E. S. Hudes, "Risk Factors for Injurious Falls: A Prospective Study 1991, Vol.46. No. 5. M164-I70," *Journal of Gerontology: MEDICAL SCIENCES*, vol. 46, no. 5, pp. M164–M170, 1991.

[3] P. W. Overstall, A. N. Exton-Smith, F. J. Imms, and A. L. Johnson, "Falls in the elderly related to postural imbalance." *BMJ*, vol. 1, no. 6056, pp. 261–264, jan 1977. [Online]. Available: http://www.bmj.com/cgi/doi/10.1136/bmj.1.6056.261

[4] W. P. Berg, H. M. Alessio, E. M. Mills, and C. Tong, "Circumstances and consequences of falls in independent community-dwelling older adults," *Age and Ageing*, vol. 26, no. 4, pp. 261–268, 1997.

[5] B. S. Roudsari, B. E. Ebel, P. S. Corso, N. A. M. Molinari, and T. D. Koepsell, "The acute medical care costs of fall-related injuries among the U.S. older adults," *Injury*, vol. 36, no. 11, pp. 1316–1322, 2005.

[6] a. M. Schillings, B. M. van Wezel, T. Mulder, and J. Duysens, "Muscular responses and movement strategies during stumbling over obstacles." *Journal of neurophysiology*, vol. 83, no. 4, pp. 2093–2102, 2000.

[7] A. M. Schillings, B. M. Van Wezel, and J. Duysens, "Mechanically induced stumbling during human treadmill walking," *Journal of Neuroscience Methods*, vol. 67, no. 1, pp. 11–17, 1996.

[8] J. J. Eng, D. A. Winter, and A. E. Patla, "Strategies for recovery from a trip in early and late swing during human walking," *Experimental Brain Research*, vol. 102, no. 2, pp. 339–349, 1994.

[9] M. Pijnappels, M. F. Bobbert, and J. H. Van Dieën, "Contribution of the support limb in control of angular momentum after tripping," *Journal of Biomechanics*, vol. 37, no. 12, pp. 1811–1818, 2004.

[10] M. Pijnappels, M. F. Bobbert, and J. H. Van Dieën, "How early reactions in the support limb contribute to balance recovery after tripping," *Journal of Biomechanics*, vol. 38, no. 3, pp. 627–634, 2005.

[11] M. Pijnappels, M. F. Bobbert, and J. H. Van Dieën, "Push-off reactions in recovery after tripping discriminate young subjects, older non-fallers and older fallers," *Gait and Posture*, vol. 21, no. 4, pp. 388–394, 2005.

[12] M. Pijnappels, M. F. Bobbert, and J. H. Van Dieën, "Control of support limb muscles in recovery after tripping in young and older subjects," *Experimental Brain Research*, vol. 160, no. 3, pp. 326–333, 2005.

[13] V. Monaco, P. Tropea, F. Aprigliano, D. Martelli, A. Parri, M. Cortese, R. Molino-Lova, N. Vitiello, and S. Micera, "An ecologically-controlled exoskeleton can improve balance recovery after slippage," *Scientific Reports*, vol. 7, no. May, pp. 1–10, 2017. [Online]. Available: http://dx.doi.org/10.1038/srep46721

Kalibracija inercijskih merilnih enot brez dodatne merilne opreme

Jernej Puc, Janez Podobnik, Marko Munih

Laboratorij za robotiko
Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
E-pošta: jp4745@student.uni-lj.si, {janezp, marko}@robo.fe.uni-lj.si

Abstract

This paper covers a full calibration process of all sensors incorporated in an inertial measurement unit (IMU): the gyroscope, accelerometer and magnetometer. Sensor characteristics are discussed and different methods presented with respect to the specifics of each sensor. The presented methods assume no aid of additional measuring equipment and inevitably differ in complexity; particular focus is given to the scalar calibration of the magnetometer, as it is the most complex. Method effectiveness is compared to two approaches from literature.

1 Uvod

Inercijske merilne enote (ang. *inertial measurement unit - IMU*) predstavljajo pomemben merilni sistem na področju sledenja gibanja, kjer se na podlagi meritev različnih količin, kot so kotna hitrost, pospešek in gostota magnetnega polja, ter metodami za senzorno fuzijo določa orientacijo objekta glede na globalni koordinatni sistem. Navadno vsebujejo triosni žiroskop in pospeškometer ter pogosto magnetometer. Poleg širše uporabnosti je razvoj majhnih in lahkih naprav (ang. *microelectromechanical systems - MEMS*) omogočil različne aplikacije v povezavi z določanjem gibanja pri človeku in stroju [1, 2].

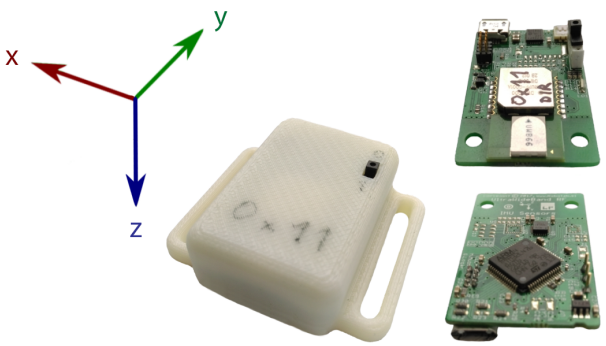
Tovrstni problemi, kot je npr. določanje orientacije z integriranjem kotne hitrosti, so občutljivi na nekompenzirane napake, zato je kalibracija senzorjev ključnega pomena; kot bo vidno pri rezultatih, imajo lahko meritve z nekalibriranimi senzorji velika odstopanja, ki lahko vodijo do napačnega končnega rezultata.

Kalibracija IMU naprav pojmuje določitev parametrov, s katerimi poskušamo zmanjšati vpliv sistemskih pogreškov vseh senzorjev, ki napravo sestavljajo (žiroskop, pospeškometer, magnetometer). Želimo jo izvesti v doglednem času na način, ki je preprost in v praktičnem smislu ponovljiv. Potreba je torej po kalibraciji brez dodatne merilne opreme, kjer se zahteva zadostno zanesljivost uporabljenega analitičnega pristopa. Zaradi pomembnih razlik med značilnostmi senzorjev se do kalibracije posameznega senzorja v splošnem pristopa različno.

Cilj prispevka je podati praktično osnovo za samostojno kalibracijo IMU naprav v celoti. Predstavljenim modelom senzornih napak sledijo opisi pristopov kalibracije in ovrednotenje učinkovitosti glede na primerljive metode iz literature.

2 Metodologija

Slika 1 prikazuje primer IMU naprave uporabljene v tem prispevku. Osi lokalnega koordinatnega sistema so skladne z osmi posameznega senzorja in prikazane glede na zaščitno ohišje.



Slika 1: Realen IMU s pripadajočimi osmi.

V idealnem primeru so izmerjeni odzivi žiroskopa ($\boldsymbol{\omega}_m$), pospeškometra ($\mathbf{a}_m$) in magnetometra ($\mathbf{h}_m$) odvisni le od merjenih količin (kotna hitrost $\boldsymbol{\omega}$, pospešek $\mathbf{a}$ in gostota magnetnega polja, v literaturi pojmovana kot magnetno polje $\mathbf{h}$), kjer vsi simboli predstavljajo tridimenzionalne vektorje. Zaradi raznih lastnosti uporabljenih materialov in neizogibnih napak pri izdelavi navedena razmerja odstopajo, kar se lahko glede na vsak senzor ponazori z ustreznim matematičnim modelom.

2.1 Viri senzornih napak

Poleg visokofrekvenčnega šuma, ki se pojavlja pri vseh senzorjih, so glavni viri napak še skalirni faktor (2.1.1), neortogonalnost (2.1.2) in ničelni odmik (2.1.3).

2.1.1 Skalirni faktor

V splošnem občutljivosti senzorjev niso enake za vse osi. Za skaliranje po oseh zato namesto navadne sorazmerne konstante uvedemo diagonalno matriko $\mathbf{S}$ s skalirnimi koeficienti K:

$$\mathbf{S} = diag(K_x,\ K_y,\ K_z) \tag{1}$$

V primeru magnetometra vpliv prisotnih feromagnetnih materialov popači zunanje magnetno polje, kar se odraža v odzivu senzorja (ang. *soft iron error*) [3]. $\mathbf{S}_h$ je tako polna kvadratna matrika velikosti 3×3.

2.1.2 Neortogonalnost

Odstopanja v poravnavi osi senzorjev z lokalnimi osmi enote privedejo do neortogonalnosti; prava vrednost za eno os tako ni odvisna le od odziva tiste osi, temveč vsebuje tudi manjšinska prispevka preostalih. Vsak stolpec v matriki neortogonalnosti $\mathbf{N}$ je vektor treh vrednosti, ki za pripadajočo os določa opisano razmerje glede na lokalni koordinatni sistem:

$$\mathbf{N} = \begin{bmatrix} \mathbf{n}_x & \mathbf{n}_y & \mathbf{n}_z \end{bmatrix} \tag{2}$$

2.1.3 Ničelni odmik

Ničelni odziv zaradi specifične narave senzorja izkazuje odmik od ničelne vrednosti, npr. zaradi vpliva magnetiziranih feromagnetnih materialov (ang. *hard iron error*). Ničelni odmik predstavimo z vektorjem $\mathbf{b}$, ki načeloma ni povsem konstanten; poleg naključnega nihanja ima zlasti odmik žiroskopa znatno temperaturno odvisnost, ki ni predstavljena v opisanem modelu, zaradi katere prihaja do časovnega lezenja.

2.2 Modeli senzornih odzivov

Glede na opisano so odzivi za posamezni senzor predstavljeni sledeče:

$$\boldsymbol{\omega}_m = \mathbf{S}_\omega \mathbf{N}_\omega \, \boldsymbol{\omega} + \mathbf{b}_\omega + \boldsymbol{\epsilon}_\omega \tag{3}$$

$$\mathbf{a}_m = \mathbf{S}_a \mathbf{N}_a \, \mathbf{a} + \mathbf{b}_a + \boldsymbol{\epsilon}_a \tag{4}$$

$$\mathbf{h}_m = \mathbf{S}_h \mathbf{N}_h \, \mathbf{h} + \mathbf{b}_h + \boldsymbol{\epsilon}_h \tag{5}$$

Matriki $\mathbf{S}$ in $\mathbf{N}$ združimo v transformacijsko matriko $\mathbf{A}^{-1}$ in zanemarimo šum $\boldsymbol{\epsilon}$. S preoblikovanjem enačb (3), (4) in (5) izpostavimo približke pravih vrednosti, ki predstavljajo kalibrirani senzorni odziv:

$$\widehat{\boldsymbol{\omega}} = \mathbf{A}_\omega \, (\boldsymbol{\omega}_m - \mathbf{b}_\omega) \tag{6}$$

$$\widehat{\mathbf{a}} = \mathbf{A}_a \, (\mathbf{a}_m - \mathbf{b}_a) \tag{7}$$

$$\widehat{\mathbf{h}} = \mathbf{A}_h \, (\mathbf{h}_m - \mathbf{b}_h) \tag{8}$$

Problem kalibracije tako privede do iskanja ustreznih transformacijskih matrik in ničelnih odmikov za posamezni senzor na enoti.

2.3 Kalibracija žiroskopa

Za zanesljivo določitev iskane transformacijske matrike $\mathbf{A}_\omega$ bi potrebovali stabilen vir referenčne kotne hitrosti, kar pomeni uporabo drage merilne opreme in daljših meritvenih postopkov. Zavoljo preprostosti smo primorani privzeti idealno sorazmerje, ponazorjeno s konstanto K_ω. Preostane nam povprečenje meritev v mirovanju za kompenzacijo ničelnega odmika:

$$\mathbf{b}_\omega = \overline{\boldsymbol{\omega}} \tag{9}$$

2.4 Kalibracija pospeškometra

Pri določanju odziva na pospešek imamo zanesljivo referenco vedno na voljo; gravitacijsko polje Zemlje je stabilno in za omejeno geografsko območje praktično nespremenljivo. Navadno je vektor polja $\mathbf{g}$ normaliziran in definiran z enoto $g \doteq 9.80665 \frac{m}{s^2}$:

$$\mathbf{g} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \, g \tag{10}$$

Glede na definicijo (10) so prave izhodne vrednosti v lokalnem sistemu pri vsaki orientaciji enote določljive s preprostimi trigonometričnimi zvezami. Za kalibracijo običajno zadostujejo meritve v vseh 6 medsebojno pravokotnih položajih; tako je pričakovani odziv ene osi vedno ± 1, preostalih dveh pa 0. Pri tem je ključno mirovanje za vsako os, sicer se v meritve vnašajo motnje translacijskega pospeška.

V nadaljevanju izhajamo iz enačbe (7) in jo s substitucijo $\mathbf{b}_a^* = \mathbf{A}_a \, \mathbf{b}_a$ preoblikujemo v linearno vsoto, da za pričakovano vrednost n-tega vzorca $\widehat{\mathbf{a}}_n$ velja:

$$\mathbf{A}_a \, \mathbf{a}_{m,n} - \mathbf{b}_a^* = \widehat{\mathbf{a}}_n \tag{11}$$

To nam omogoča, da sestavimo matrični sistem oblike $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$:

$$\begin{bmatrix} a_m^{x_1} & a_m^{y_1} & a_m^{z_1} & -1 \\ a_m^{x_2} & a_m^{y_2} & a_m^{z_2} & -1 \\ \vdots & \vdots & \vdots & \vdots \\ a_m^{x_n} & a_m^{y_n} & a_m^{z_n} & -1 \end{bmatrix} \begin{bmatrix} \mathbf{A}_a^T \\ \mathbf{b}_a^{*T} \end{bmatrix} = \begin{bmatrix} \widehat{a}_{x_1} & \widehat{a}_{y_1} & \widehat{a}_{z_1} \\ \widehat{a}_{x_2} & \widehat{a}_{y_2} & \widehat{a}_{z_2} \\ \vdots & \vdots & \vdots \\ \widehat{a}_{x_n} & \widehat{a}_{y_n} & \widehat{a}_{z_n} \end{bmatrix} \tag{12}$$

Sistem je predoločen in zato nima enolične rešitve. Rešimo ga po linearni metodi najmanjših kvadratov, tako da ga preoblikujemo na minimizacijski problem, kjer za $\boldsymbol{\beta} = \begin{bmatrix} \mathbf{A}_a^T \\ \mathbf{b}_a^{*T} \end{bmatrix}$ iščemo približek $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 \tag{13}$$

Numerično stabilne rešitve problema najmanjših kvadratov vključujejo ortogonalne razcepe. Rešitev poiščemo npr. s singularnim razcepom (ang. *singular value decomposition - SVD*) [4]. Iz enačb (11), (12) in rešitve $\widehat{\boldsymbol{\beta}}$, podane po vrsticah in stolpcih, sledi:

$$\mathbf{A}_a = \widehat{\boldsymbol{\beta}}_{1:3,1:3}^T \tag{14}$$

$$\mathbf{b}_a = \mathbf{A}_a^{-1} \, \widehat{\boldsymbol{\beta}}_{1:3,4}^T \tag{15}$$

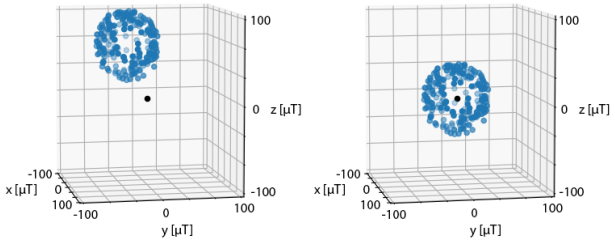
2.5 Kalibracija magnetometra

Z nadzorovano referenco homogenega magnetnega polja bi zveze med primerjanimi vektorji lahko zopet iskali neposredno po metodi najmanjših kvadratov, a se brez merilne opreme zanašamo na Zemljino magnetno polje. Slednje je zaradi relativne šibkosti (red 10^{-5} T) zlahka preglašeno z elektromagnetnimi motnjami. Poleg tega ima horizontalno in vertikalno komponento, kar dodatno otežuje pristop. Pri kalibraciji tako izhajamo iz privzete konstantne magnitude Zemljinega magnetnega polja H_{mag} in stremimo k njeni ohranitvi, zato gre v tem smislu za skalarno kalibracijo.

Idealno je norma oz. skupna velikost senzornega odziva $\mathbf{h}$ ne glede na orientacijo magnetometra konstantna:

$$\|\mathbf{h}\| = konst. = H_{mag} \tag{16}$$

Posledično bi morali vzorci magnetometra, vzeti v različnih orientacijah, opisovati sfero s polmerom, ki je enak podani magnitudi ($cca.\ 47,95\ \mu T$), in središčem v lokalnem koordinatnem izhodišču.



Slika 2: Sfera glede na koordinatno izhodišče, kot jo opisujejo vzorci magnetometra pred in po kalibraciji.

Kot je razvidno iz slike 2, je v skrajnem primeru oblika sfere popačena, njeno središče pa od izhodišča znatno odstopa, zato je smiselna umestitev na površino elipsoida s premaknjenim središčem. V primeru znanih polosi se lahko te po velikosti izenači in ustrezno skalira, središče elipsoida pa prestavi v lokalno izhodišče, s čimer dobimo pričakovano sfero.

Pri določanju elipsoida izhajamo iz enačbe drugega reda v prostoru treh spremenljivk, ki v splošnem opisuje različne kvadrične površine:

$$\begin{aligned} & ax^2 \ + \ by^2 \ + \ cz^2 \\ & + \ 2fyz \ + \ 2gxz \ + \ 2hxy \\ & + \ 2px \ + \ 2qy \ + \ 2rz \ + \ d = 0 \end{aligned} \tag{17}$$

Koeficiente poiščemo po splošnem algoritmu za prileganje vzorcev na elipsoid (Li [5]). Ta se prevede na reševanje matričnih sistemov po metodi najmanjših kvadratov; rezultat končnega sistema sta lastna vektorja $\mathbf{v}_1$ in $\mathbf{v}_2$, ki skupaj tvorita vektor koeficientov kvadrične enačbe $\mathbf{v}$, na podlagi katerih je elipsoid v celoti določen:

$$\mathbf{v} = [a\ b\ c\ f\ g\ h\ p\ q\ r\ d]^T \tag{18}$$

Preoblikovanje izraza (17) v matrični zapis s substitucijami $h_x = x$, $h_y = y$ in $h_z = z$ privede do oblike:

$$\mathbf{h}^T\mathbf{M}\mathbf{h} \ + \ \mathbf{h}^T\mathbf{n} \ + \ d = 0, \tag{19}$$

kjer so $\mathbf{h}$, $\mathbf{M}$ in $\mathbf{n}$ definirani kot:

$$\mathbf{h} = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix}, \ \mathbf{M} = \begin{bmatrix} a & h & g \\ h & b & f \\ g & f & c \end{bmatrix}, \ \mathbf{n} = \begin{bmatrix} p \\ q \\ r \end{bmatrix} \tag{20}$$

Izraz (19) dopolnimo do popolnega kvadrata, da izrazimo središče (iskani odmik b_h):

$$\begin{aligned} & \mathbf{h}^T\mathbf{M}\mathbf{h} \ + \ \mathbf{h}^T\mathbf{n} \ + \ d = \\ & (\mathbf{h} - \mathbf{b}_h)^T \ \mathbf{M} \ (\mathbf{h} - \mathbf{b}_h) \ - \ C \ , \end{aligned} \tag{21}$$

kjer velja:

$$\mathbf{h}_h = -\mathbf{M}^{-1} \ \mathbf{n} \tag{22}$$

$$C = \mathbf{b}_h^T \ \mathbf{M}^{-1} \ \mathbf{b}_h - d \tag{23}$$

V nadaljevanju določimo skalirne vrednosti, vsebovane v iskani transformacijski matriki $\mathbf{A}$. Velja, da lastni vektorji matrike $\mathbf{M}$ ($\mathbf{h}_\lambda$) predstavljajo polosi danega elipsoida (ang. *principal axis theorem* [6]):

$$\begin{aligned} & (\mathbf{h}_{\lambda_{a,b,c}} - \mathbf{b}_h)^T \ \mathbf{M} \ (\mathbf{h}_{\lambda_{a,b,c}} - \mathbf{b}_h) \\ & = \lambda \ (\mathbf{h}_{\lambda_{a,b,c}} - \mathbf{b}_h)^T \ (\mathbf{h}_{\lambda_{a,b,c}} - \mathbf{b}_h) \\ & = \lambda \ \|(\mathbf{h}_{\lambda_{a,b,c}} - \mathbf{b}_h)\|^2 = C \end{aligned} \tag{24}$$

Polosi so tako določene z izrazom (25):

$$s_{a,b,c} = \|(\mathbf{h}_{\lambda_{a,b,c}} - \mathbf{b}_h)\| = \sqrt{\frac{C}{\lambda_{a,b,c}}} \tag{25}$$

Dolžine polosi izenačimo glede na poljubno polos, npr. s_a. Skalirna matrika $\mathbf{S}$ je tako:

$$\begin{aligned} \mathbf{S} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{s_a}{s_b} & 0 \\ 0 & 0 & \frac{s_a}{s_c} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{\frac{\lambda_b}{\lambda_a}} & 0 \\ 0 & 0 & \sqrt{\frac{\lambda_c}{\lambda_a}} \end{bmatrix} \\ &= \sqrt{\frac{1}{\lambda_a}} \begin{bmatrix} \sqrt{\lambda_a} & 0 & 0 \\ 0 & \sqrt{\lambda_b} & 0 \\ 0 & 0 & \sqrt{\lambda_c} \end{bmatrix} = \sqrt{\frac{1}{\lambda_a}} \ \mathbf{\Lambda}^{\frac{1}{2}} \end{aligned} \tag{26}$$

Za pravilno skaliranje polosi poravnamo z osmi koordinatnega sistema; spremenimo orientacijo elipsoida in ga po skaliranju vrnemo v prvotno lego. Orientacijo elipsoida določajo orientacije pripadajočih polosi oz. normalizirani lastni vektorji matrike $\mathbf{M}$; slednji tvorijo stolpce rotacijske matrike $\mathbf{V}$, ki opisuje omenjeno orientacijo. Pravilno skaliranje tako dosežemo z zaporedjem operacij $\mathbf{V}\mathbf{S}\mathbf{V}^T$. Za kvadratno matriko $\mathbf{M}$ z ortogonalnimi lastnimi vektorji, ki tvorijo stolpce matrike $\mathbf{V}$ velja [7]:

$$\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \tag{27}$$

Z upoštevanjem izrazov (26) in (27) ter lastnosti rotacijskih matrik $\mathbf{V}^{-1} = \mathbf{V}^T$ lahko zapišemo:

$$\mathbf{V}\mathbf{S}\mathbf{V}^T = \sqrt{\frac{1}{\lambda_a}} \ \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^{-1} = \sqrt{\frac{1}{\lambda_a}} \ \mathbf{M}^{\frac{1}{2}} \tag{28}$$

V zadnjem koraku upoštevamo zvezo (25) in skaliramo polmer sfere na želeno dolžino, npr. H_{mag} ali 1. Poleg izraza (22) sta tako določena oba kalibracijska parametra:

$$\mathbf{A}_h = \frac{H_{mag}}{\sqrt{C}} \ \mathbf{M}^{\frac{1}{2}} \tag{29}$$

3 Rezultati

Sledi ovrednotnje metod za kalibracijo pospeškometra in magnetometra na podlagi primerjave z dvema metodama iz literature (Merayo [8], Frosio [9]). Prva temelji na kalibraciji s prileganjem vzorcev na elipsoid, medtem ko je druga iterativna in sloni na Gauss-Newtonovi optimizaciji.

3.1 Ovrednotenje kalibracije pospeškometra

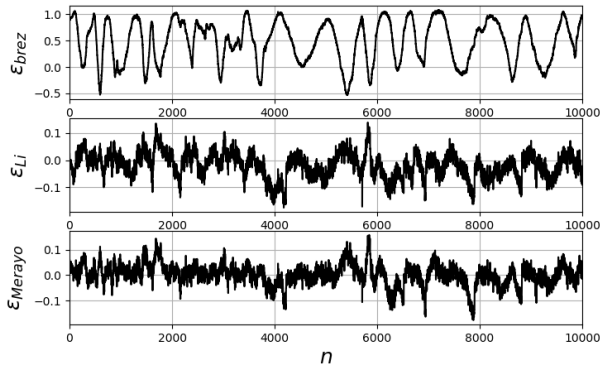
Opaznejša odstopanja se pojavljajo pri RMS vrednostih napake (ang. *Root-Mean-Square*). Vzrok odstopanj leži v izračunanih ničelnih odmikih, ki RMS vrednostim dodaja izrazito enosmerno komponento; primerjanim metodam pritiče večje odstopanje kot pri navadnem skaliranju, kalibracija po metodi najmanjših kvadratov pa daje boljše rezultate.

Tabela 1: RMS napake odstopanj kalibriranih meritev od pričakovanih vrednosti; skupno in po oseh. $LSTSQ$ označuje opisani postopek po metodi najmanjših kvadratov, $Merayo$ [8] in $Frosio$ [9] pa sta primerjani metodi. Razdelek $brez$ predstavlja originalne rezultate brez kalibracije.

RMS	$brez$	$LSTSQ$	$Merayo$	$Frosio$
x-os	0.06343	0.00640	0.08447	0.04743
y-os	0.03357	0.00620	0.03284	0.08435
z-os	0.04977	0.02745	0.03312	0.03339
vse osi	0.08734	0.02886	0.09649	0.10237

3.2 Ovrednotenje kalibracije magnetometra

Pri ovrednotenju smo obravnavali magnetometer z velikim ničelnim odmikom kot najslabši možni primer. Netipično veliki odmiki so se pri iterativni metodi ([9]) izkazali za problematične; konvergenca ni bila nikoli dosežena, zato je metoda na sliki 3 izpuščena.



Slika 3: Napaka norme vzorcev $\epsilon_h = 1 - \|\mathbf{h}\|$ v odvisnosti od n-tega vzorca $\mathbf{h}_n$. ϵ_{Li} [5] označuje opisani postopek, ϵ_{Merayo} [8] primerjano metodo in ϵ_{brez} originalni rezultat na podlagi nekalibriranih vzorcev.

Pomanjkljivost iterativne primerjane metode se je odpravilo tako, da se je skupini vzorcev predhodno odštela njihova povprečna vrednost; končni odmik je tako določala vsota povprečja in novo izračunani odmik. Za ustrezno primerjavo je bil enak pristop uporabljen tudi pri drugih metodah.

Iz tabele 2 je razvidno, da je predhodno odštevanje povprečja povzročilo splošno izboljšanje; napake so pri vseh metodah manjše oz. vsaj manj razpršene. Po izboljšanju opisana metoda le še minimalno odstopa od primerjanih.

Tabela 2: RMS odstopanja norme od pričakovanega radija konstantne dolžine 1 ($\epsilon_h = 1 - \|\mathbf{h}\|$) za vse pristope pred in po predhodnem odštevanju povprečja. Li [5] se nanaša na opisano metodo, $Merayo$ [8] in $Frosio$ [9] sta primerjani metodi, razdelek $brez$ pa se navezuje na originalne rezultate brez kalibracije.

RMS	$brez$	Li	$Merayo$	$Frosio$
Pred	0.62576	0.05010	0.04220	-
Po	0.24736	0.04122	0.03997	0.04372

4 Zaključek

V prispevku je bil predstavljen hiter in praktičen postopek za celotno kalibracijo inercijskih merilnih enot. Podani so bili modeli senzornih napak in glede na specifične lastnosti vsakega senzorja opisani pristopi za njihovo kalibracijo. Obe metodi, ki sta bili primerjani s pristopoma iz literature, sta dosegli zadovoljive rezultate in predstavljata smiselno alternativo na področju enostavne kalibracije. Od teh se je dokaj neposredna uporaba metode najmanjših kvadratov pri kalibraciji pospeškometra kljub očitni preprostosti izkazala za posebej natančno in robustno. Na podlagi rezultatov je bil utemeljen tudi sklep, da z majhnim posegom pred kalibracijo v obliki odštevanja povprečne vrednosti od skupine vzorcev dosežemo vidne izboljšave.

Literatura

[1] O. J. Woodman. An introduction to inertial navigation. University of Cambridge, Computer Laboratory, Technical Report UCAM-CL-TR-696. Avgust, 2007.

[2] J. Podobnik, M. Munih. Določanje transformacije med triado optičnih markerjev in inercijsko merilno enoto. ERK'2014, Portorož, A:103-106.

[3] V. Renaudin, M. H. Afzal, G. Lachapelle. Complete Triaxis Magnetometer Calibration in the Magnetic Domain. Hindawi Publishing Corporation, Journal of Sensors. Vol. 2010.

[4] B. Plestenjak. Razširjen uvod v numerične metode. DMFA - založništvo (2015), ISBN 978-961-212-264-5.

[5] Q. Li, J. G. Griffiths. Least Squares Ellipsoid Specific Fitting. IEEE Computer Society, Proceedings of the Geometric Modeling and Processing 2004.

[6] J. N. Franklin. Matrix theory. Dover Publications (2012), sec. 4.6, 95. ISBN 978-048-641-179-8.

[7] H. Abdi. Eigen-decomposition: eigenvalues and eigenvectors. SAGE Publications, Encyclopedia of Measurement and Statistics (2007), 304–308.

[8] J. M. G. Merayo, P. Brauer, F. Primdahl, J. R. Petersen, O. V. Nielsen. Scalar calibration of vector magnetometers. IOP Publishing, Meas. Sci. Technol. 11 (2000), 120-132.

[9] I. Frosio, F. Pedersini, N. A. Borghese. Autocalibration of MEMS Accelerometers. IEEE Transactions on Instrumentation and Measurement, Vol. 58, No. 6, June 2009.

Implementation of autonomous SLAM on a robotic rover using ROS

Guillaume Peter[1], Tadej Petrič[2] and Andrej Gams[2]

[1]*IUT of Cachan, Université Paris-Sud, 9 Avenue de la Division Leclerc, 94234 Cachan, France*
[2]*Department of Automatics, Biocybernetics and Robotics, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*
E-mail: guillaume.peter@u-psud.fr, andrej.gams@ijs.si

Abstract

In this paper we present an implementation of autonomous rover-robot navigation. This robot makes mapping using a SLAM system implemented with ROS. The robot can make a map of the environment and navigate in this environment without the intervention of a human. Initial implementation results are presented.

1 Introduction

Autonomous and mobile robots are the future of robotics in our world. To operate in the world, robots need to navigate in unknown environments. In order to locate themselves, robots typically use SLAM (Simultaneous Localization And Mapping) algorithms. Robots equipped with such technology are able to explore and understand a complex environment.

Programing of these robots is very complicated, as many things have to be taken into consideration. Building on previously developed and published tools makes it easier. This is the reason why we are using ROS (Robot Operating System) [1, 2]. This powerful and professional tool permits to make a link between simulation, programming and 3D visualization.

This paper presents the implementation of autonomous rover navigatoin behavior using a SLAM system within ROS. In Section 2 the meta-operating system called ROS is presented with its advantages and weaknesses. In Section 3 we discuss autonomous mapping with different approaches. Finally, Section 4 presents initial implementation results in the Piborg rover.

2 ROS

ROS is an open source software used to easily develop robotic applications. It was created in 2007 and since then the ROS community is getting bigger every year. It is being used by independent programmers and companies because of its compatibility. When you program with ROS you have the choice between C++ and Python and for the OS it's compatible with Linux, Mac OS and Windows.

When a project combines numerous micro controllers with a lot of different actions, their communication and integration quickly become overly complex. ROS simplifies all the communications between micro controllers and computers. Like this the calculation part is made by the computer and the micro controller makes only basic calculations. This environment is a multi-task environment with a very precise architecture. First, you need to run a master program which will setup the environment. Then we can run tasks in the environment. A task is called a node and it is used to communicate with another node. It can be used for different actions like read a sensor, control motors or make calculations. All the nodes can subscribe and publish topics. A topic is a named bus over which nodes exchange information. For example, a node which reads a temperature sensor will publish a topic with the value of the temperature. The big advantage of ROS is all the nodes can access to the topics, so data communications are easier.

But ROS doesn't have only advantages. This software need specific version of the OS. You can't adapt the version of ROS to your OS, you have to adapt your OS to the version of ROS. It's a very big forward jump concerning robotic interfaces but it takes some times to understand it clearly.

3 Autonomous Mapping

3.1 Hardware

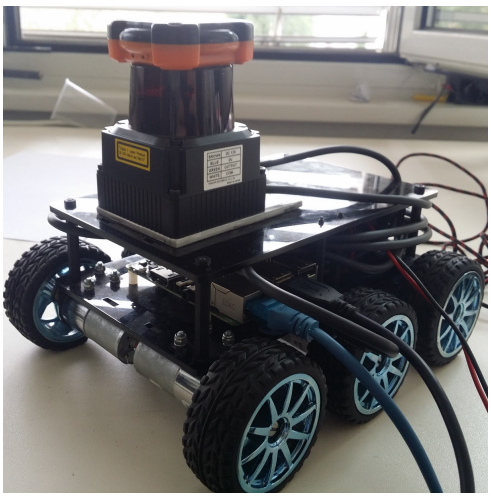


Figure 1: The small robot Piborg used in the project.

Figure 1 shows the robot we used for the project. To

implement a SLAM algorithm we need a mobile robot. For the chassis of the robot we used a simple 6 wheels rover which turns by inverting the rotation of the wheels. Like this the robot can turn on the spot. Each wheels is equipped with a motor in order to have total control of the robot. To control the motors we use a Piborg card communicating in I2C with a Raspberry 2B, hence the name of the robot is Piborg [4]. This controller is a 2 channels controller. You can choose the way of rotation and the speed of rotation with it. In order to make the mapping we use a rotating range finder lidar sensor. It is situated at the top of the robot. This sensor is a distance sensor that measures distances on a plane around it. From the position of the sensor we can create a map with the collected data. We are using the Hokuyo UTM-30LX [3] which has a distance range of 30 meters and a resolution of 0.25 degree on 270 scanning degrees. It communicates by a USB cable with the Raspberry and the Raspberry is linked to the computer through a local network (for now ethernet cable, but we will make it wireless in the next implementation). We can summarize the connections with the schematic presented in Fig. 2. It shows the schematic of the connection between all the materials

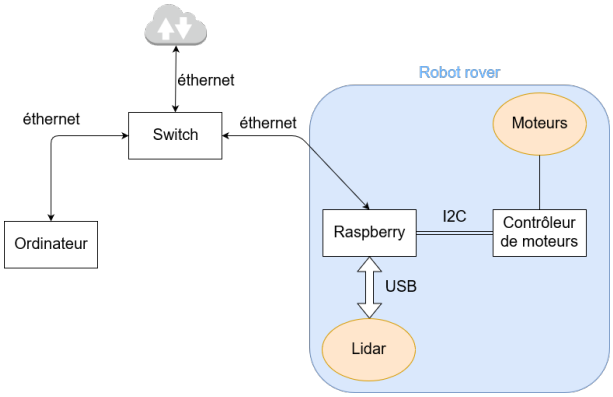


Figure 2: Schematic of the connections.

In Fig. 2, Ordinateur stands for computer, Contrôleur de moteurs is the Motor controller, and Moteurs are Motors. The cloud symbolizes Internet, because you need to be connected to Internet.

3.2 Software

Concerning ROS we are using ROS Kinetic which is known as a very stable version. On the computer we are using Ubuntu 16.04 and in the Raspberry we are using Ubuntu Mate. At the beginning of the project we first tried with Raspian for the Raspberry but quickly we concluded it was not appropriate for the application. In order to visualize the output of the sensor for mapping we use the vizualization software package Rviz.

3.3 SLAM

A robot equipped with a SLAM system is an autonomous robot. The objective is not limitied only to avoiding a wall. The robot must also understand its environment and how it can navigate in this environment. In the next section we give a brief overview about 2D navigation. Here we do not take in the consideration the z axis, since the robot navigates on a flat surface.

The first step is to create a map with the robot. The robot needs to know everything about its environment. It will explore and scan it with the lidar. Result of a scan is shown in Fig. 3. The software regroups all the collected data with the scans and builds a map in 2D.

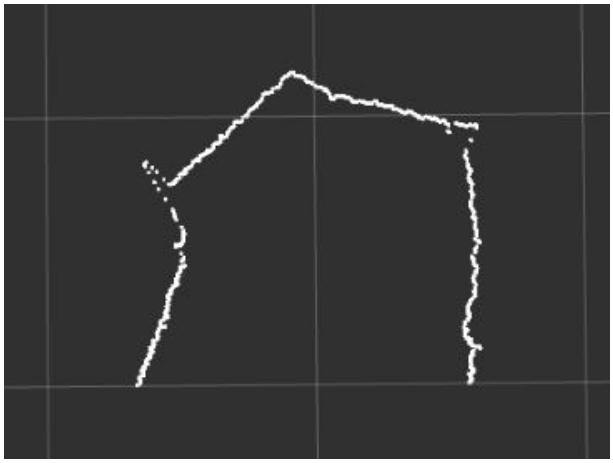


Figure 3: Result of a scan in Rviz

The robot needs numerous scans because like this it can confirm clearly the position of a wall or of an object. The robot needs to create a closed space, on the map there is no open door. To improve the quality of the mapping it's recommended to put encoders on the wheels.

Once it's made, the robot knows its environment but it needs to know its own position as well. It will navigate in a closed space with walls and objects, so the size of the robot is very important. A free space is not all the time a space where the robot can go. When the robot will move in the environment the lidar will scan all around it. Then the software will compare the obtained data with the map and find the position of the robot. This technique works better in an irregular space. In a square room with nothing on the ground, everything is too similar. Each scan has to indicate a unique space.

The SLAM method with ROS can be used easily because it can exploit the main feature of ROS - re-usability of robotic software. We just need to change some parameters. The tool named "hector slam" [6, 5] permits to create maps and we can visualize it with Rviz. Odometry is taken in consideration by the software so it's very complete. But without odometry, if you don't have a referential point in a square room the localization doesn't work.

4 Results

In this part we discuss more about the project. What we have made, how we have made and how we want to exploit the project?

As you can see on the schematic in Fig. 4, our implementation relies on 3 nodes and 2 topics. We decided to separate the tasks to be more efficient.

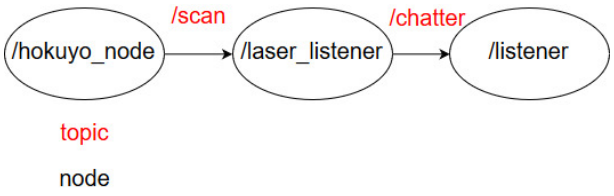


Figure 4: Schematic of architecture of the program.

The hokuyo sensor node is a task taken from the libraries of ROS. It permits to exploit correctly the lidar connected by the USB cable. This node publishes some topics, like the topic named scan. So the first node creates a link between the sensor and the Raspberry. The node named laser listener is a subscriber of the topic scan. It receives different data concerning the scans made by the sensor. Every 0.25 degrees made of the rotating sensor, the distance to objects around the robot is measured. Every revolution of the lidar the topic scan contains an array, each case corresponding to a distance taken at a precise angle. It is in the node laser listener we will detect objects in front of the robot. The robot needs to create a map and to not touch walls or objects. So if we see something in front of the robot it will turn in another direction.

But it is not the only task made by the laser listener node. The two nodes we have seen are implemented in the Raspberry. As said, the raspberry cannot make all the calculations to make the mapping. It would take too much time and we need a fixed screen where we can see the robot moving in the map. So we need to transmit all the data from the raspberry to the computer. For this we use an ethernet cable. Like this all the data are received on the computer and we can create and visualize the map in real time. The software Rviz is able to locate the topic scan as if it was running on the computer. It is one of the reason why ROS is very helpful. Once you are connected to the device you can visualize data as if everything is one system. The topic chatter permits to read the data of the sensor and the node listener permits to receive them.

Concerning the project, we are creating the map from the collected data as you can see in the Fig. 5

The final target of this is to put the robot in a room. Then with the SLAM system it makes a map of the room and we save it. But we want an autonomous navigation from the robot. Our goal is to put the robot in the room and on the software we select a point to go on the map. The robot should autonomously navigate to that position using the collected data from the sensor. Note that no odometry data will be present. The final goal is to build up the knowledge to implement similar autonomous behavior on a more advanced mobile robotic platform.

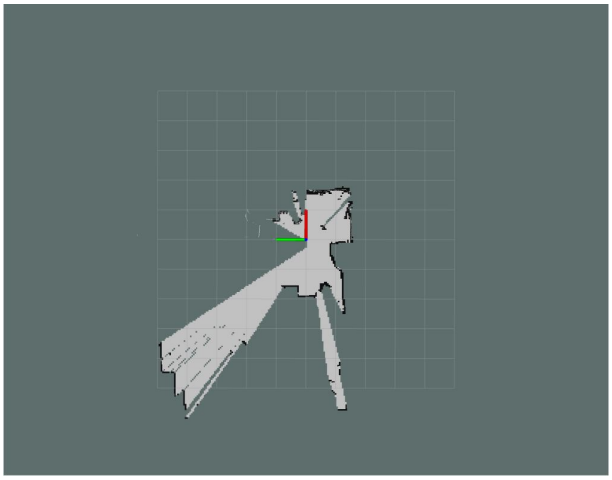


Figure 5: Visualization of a map on Rviz

References

[1] ROS, http://www.ros.org/

[2] ROS,http://www.willowgarage.com/papers/ros-open-source-robot-operating-system

[3] Hokuyo, https://www.hokuyo-aut.jp/

[4] Piborg, https://www.piborg.org/

[5] SLAM, https://husarion.com/tutorials/ros-tutorials/6-slam-navigation/

[6] SLAM, https://www.mrpt.org/

[7] Ubuntu MATE, https://ubuntu-mate.org/raspberry-pi/

Razvoj strojne in programske infrastrukture za module rekonfigurabilne robotske celice

Jakob Purg[1], **Timotej Gašpar**[2]

[1]*Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana,*
[2]*Institut "Jožef Stefan", Jamova cesta 39, 1000 Ljubljana*
E-pošta: jakobpurg@gmail.com

Developement of hardware and software infrastructure for reconfigurable robot cell modules

Frequent product launches and changes in demand require rapid adaptations in industrial manufacturing. Small and medium sized enterprises (SMEs) are feeling the arising pressure the most as they tend to fine tune their production in a "just-in-time" manner. To tackle these problems there is an increase of research in the field of reconfigurable robotic cells.

Those robot workcells can have different peripheral equipment. In this paper we present peripheral modules that can be quickly plugged or unplugged to the cell with the custom made "plug-and-produce" (P&P) connector. The connector allows the pass-through of pressured air, electrical power and communication via Ethernet. This allows the modules to receive all the necessary for their operation and at the same time they are connected to the cell's software network. To ease the access to the functionalities of these modules, we built a web Server that runs on a micro-computer on board of the module.

The web server was built using NodeJS and JavaScript programming language so that it can be easily transferred to different modules and set working with as little modifications as possible.

1 Uvod

Globalni trg je zelo kompetitiven in pritiski po zagotavljanju dovoljšne proizvodnje ter zagotavljanje kakovosti so za proizvodno usmerjena podjetja visoki. Te pritiske še bolj občutijo majhna oz. srednje velika podjetja z maloserijskimi proizvodnjami. Prav v tovrstnih podjetjih pa je avtomatizacija redka saj je vložek sredstev vanjo večinokrat previsok. Visoki stroški pa ne izhajajo le iz visokih cen strojev ampak tudi iz porabljenega časa za postavitev avtomatizirane proizvodne linije [1]. Da bi približali avtomatizacijo in robotiko tudi takim podjetjim je v porastu trend razvoja prilagodljivih oz. rekonfigurabilnih robotskih celic. To so celice, ki jih je mogoče relativno hitro postaviti, je njihovo obliko enostavno prilagoditi in so v principu modularne.

Primere implementacije modularnih celic lahko vidimo tudi kot trend h kateremu se giblje razvoj rekonfigurabilnih celic [2, 3, 4]. Dober primer take implementacije predstavi Chen v svojem delu, ki se osredotoča na iskanje optimalnih konfiguracij modulov celice za izvajanje specifične naloge [5].

Robotska celica, ki je predmet tega dela, je bila razvita s sledenjem smernicam za rekonfigurabilne robotske celice v sklopu projekta "ReconCell" [6, 7]. Celica je zgrajena iz ogrodja, ki omogoča hitro postavitev ter rekonfiguracijo. Ima dva robota UR-10m, oba opremljena s sistemom izmenjevalcev orodja. Celica je seveda tudi modularna, kar omogočajo posebni konektorji, ki so bili razviti v sklopu projekta.

Informacijski sistem celice je bil zgrajen na podlagi sistema "Robot Operating System – ROS". Gre za odprtokodno delovno okolje (*ang. framework*), ki ponuja velik nabor knjižnic in orodij za hiter razvoj robotske programske opreme [8].

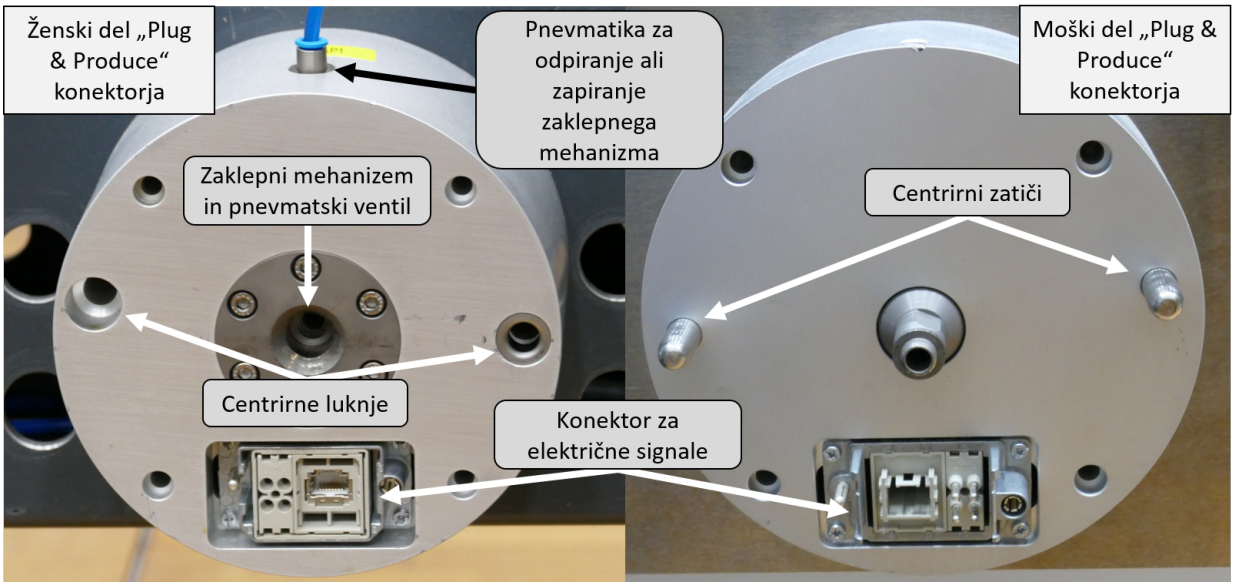
V tem delu bomo predstavili opremo ter orodja, ki smo jih razvili za podporo uporabe modulov celice. Cilj je bil module opremiti z mikroračunalnikom, ki omogoča povezovanje posameznih modulov v ROS omrežje ter njihovo krmiljenje. Hkrati pa smo želeli, da je na mikroračunalnikom postavljen strežnik, ki omogoča upravljanje z nekaterimi funkcionalnostimi modula preko spletnega brskalnika. Razvito strojno opremo bomo opisali v poglavju 2, programsko opremo pa v poglavju 3.

2 Moduli v rekonfigurabilni celici

Da bi povečali stopnjo rekonfigurabilnosti robotske celice, smo določene komponente celice razvili kot module. Namen je bil omogočiti hitro spremembo namembnosti celice s tem, da se lahko module v celico poljubno priklopi ali od nje odklopi. Zahteva po hitrih priklopih in odklopih pa je privedla do razvoja dodatne strojne kot tudi programske opreme. Tekom projekta je bil razvit poseben konektor za priklop modulov v celico ter programska oprema za podporo hitrih priklopov in odklopov modulov.

2.1 Konektorji za hiter priklop modulov

Moduli morajo biti zasnovani na takšen način, da dopuščajo hiter priklop in odklop. Hkrati pa mora ta priklop zagotavljati togo in ponovljivo sklopitev s celico, saj v nasprotnem primeru robot ne more natančno uporabljati modulov. Dodatna lastnost, ki jo mora konektor zagotavljati je, da modul dobi vse potrebne priklope za obratovanje,



Slika 1: “Priklopi in Proizvajaj” konektor.

t. j. napajanje, komprimiran zrak ter komunikacijski kanal. Tem zahtevam zadošča konektor, ki smo ga tekom projekta razvili in sicer t. i. “Priklopi in Proizvajaj” (*ang. Plug and produce – P&P*) konektor. Ženski del konektorja je togo pričvrščen na ogrodje celice, moški pa na perifernih modulih. Konektor omogoča poleg toge sklopitve tudi pretok komprimiranega zraka, omrežno napetost ter Ethernet povezavo. Tako lahko zagotovimo, da periferni moduli dobijo vse potrebno za delovanje ob sklopu na konektor.

Konektor je okrogle oblike in ima na sredini zaklepni mehanizem, ki hkrati služi kot ventil za dovod zraka v module. Na spodnjem delu ima priključke za električne signale, kar pomeni ethernet povezavo ter omrežno napetost. Ethernet povezava se sklene z RJ45 konektorjem, omrežna napetost pa preko namenskih priključkov. Ob straneh ima centrirne luknje (ženski del) oz. zatiče (moški del), za natančno orientacijo ter togo sklopitev (slika 1). Ventil na sredini je zgrajen tako, da je zaprt, torej zraka ne prepušča, ob odsotnosti modula. Konektorjev zaklepni mehanizem se odklene s pomočjo pnevmatike, ki je dovedena na vrhu ženskega konektorja. To pomeni, da je potrebna le takrat, ko želimo modul od celice odklopiti, kadar pa želimo modul v celico priklopiti, pa je dovolj, da dva konektorja sklenemo skupaj s potiskom.

2.2 Obstoječi moduli

Razvito robotsko celico smo do sedaj preizkusili na treh različnih proizvodnih procesih v laboratorjiskih pogojih, kar pomeni, da se nismo osredotočali na hitre cikle ampak na izvedljivost. Prvi proces je bil sestavljanje dveh različnih avtomobilskih luči, drugi se je osredotočal na sestavljanje pogonskega sklopa motoriziranega pohištva, tretji pa je predvideval sestavljanje elektromagnetne sklopke.

Da bi lahko vse te proizvodne procese izvedli v eni celici, je bilo razvitih nekaj različnih modulov. S tem smo dosegli želeno fleksibilnost celice in zadostili zahtevi po hitri rekonfiguraciji. Za naštete primere smo razvili sledeče module:

- pilagodljiv vpenjalni modul za avtomobilske luči,
- vpenjalni modul za pogonski sklop motoriziranega pohištva (slika 2),
- modul s prešo za sestavljanje elektromagnetne sklopke,
- modul z naborom robotskih orodij, ki jih lahko robota po potrebi uporabita,
- modul s katerega robota pobirata kose za sestavljanje.

Trenutno so vsi razviti moduli na kolesih in jih je za premikanje potrebno potiskati. Namen je, da se v bodoče razvije tudi avtonomne module, ki bi se lahko v celico priklopili in odklopili sami. Hkrati, pa je želja tudi robote v celico dodati kot modul, saj bi na ta način lahko dosegli večjo fleksibilnost pri sprotni rekonfiguraciji celice.

3 Programska oprema

Kot že prej omenjeno, je informacijska infrastruktura robotske celice zgrajena na podlagi sistema ROS (slika 3). Vsi programski gradniki med seboj komunicirajo tako, kot predvidevajo konvencije v ROS-u. To nam omogoča relativno enostaven razvoj novih gradnikov oz. funkcionalnosti celice. Da smo lahko module vključili v ROS, jih je bilo potrebno opremiti z neke vrste računalnikom, na katerem lahko tečejo ROS programi. V našem primeru je bil to mikroračunalnik *Raspberry Pi 3 model B*, ki je zelo popularen v ROS-ovi mednarodni skupnosti, ima zadostno število vhodno/izhodnih vrat ter je primerno majhen.



Slika 2: Vpenjalni modul za pogonski sklop motoriziranega pohištva.

3.1 ROS infrastruktura

Centralni del sistema ROS sestavlja t. i. ROS jedro (*ang. roscore*). To je program, ki mora teči na enem od gradnikov celice, do katerega imajo vsi ostali gradniki dostop. Jedro nosi informacije o vseh ostalih programih oz. vozliščih (*ang. node*) ter o podatkih, ki jih ta vozlišča od sebe dajajo oz. prejemajo. To je narejeno po sistemu "prijavi" in "objavi" (*ang. subscribe and publish*). Ti podatki so nato v omrežju dostopni vsak na svoji temi (*ang. topic*). Spisek vseh objavljenih tem je v jedru, tako da vozlišča ne potrebujejo informacije o tem kdo podatke obljavlja, ampak le informacijo na kateri temi podatke najti.

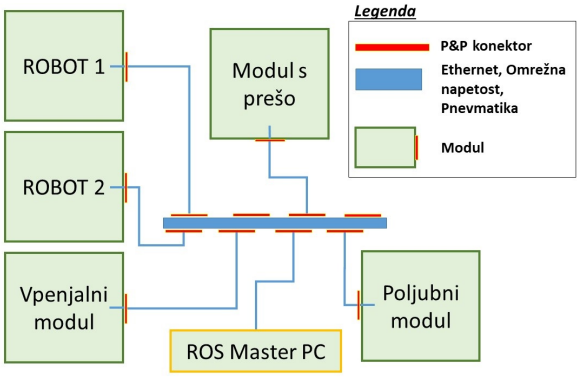
Krmiljenje modulov smo zastavili tako, da na mikroračunalniku teče program, ki ob ukazu spremeni stanje digitalnih izhodov. Poslužili smo se ROS mehanizma storitev (*ang. service*). Ta mehanizem omogoča, da vozlišče v sistemu objavi katere storitve lahko izvede in na kakšen način se jih proži. Storitev, ki smo jo razvili omogoča spreminjanje stanja digitalnih izhodov mikroračunalnika. Ti so nato povezani na elektromagnetne pnevmatske ventile. Slednji lahko nato odpirajo ter zapirajo vpenjala ali pa sprostijo zavore prilagodljivih vpenjal [9].

Razvoj vozlišča na modulovem mikroračunalniku sicer omogoča dostop do funkcionalnosti modula znotraj ROS omrežja. Težava tega pristopa je, da mora računalnik, iz katerega želimo dostopati do teh funkcionalnosti, imeti naložene knjižnice in programe, ki omogočajo komunikacijo s sistemom ROS. Da bi povečali dostopnost do teh funkcionalnosti, smo razvili strežnik, ki uporabnikom, s poznavanjem IP naslova modula, preko internetnega brskalnika ponuja nekatere od teh funkcionalnosti.

3.2 Strežnik

Strežnik je bil razvit v programskem jeziku *JavaScript*, poganja pa ga *Node.js* [10]. Za dostop do nekaterih ROS funkcionalnost, znotraj programskega okolja *JavaScript*, smo uporabili knjižnico *rosnode*. Tako samo lahko razvili strežnik, ki teče na modulu in streže *HTML* stran komurkoli, ki se nanj poveže. Dostopnost do strani je omogočena le, če je odjemalec v omrežju celice.

Strežnik najprej iz datoteke, ki vsebuje celoten zaganjalni paket - *roslaunch*, prebere imena ter tipe storitev, ki so na voljo na posameznem modulu, se prijavi na storitve, ki so pod istim imenom že v ROS omrežju, ter prebere trenutna stanja in tipe posameznih funkcij oz. funkcionalnosti. To informacijo nato preko JSON-a (*ang. JavaScript Object Notation*) pošlje do odjemalca (*ang. Client*). Odjemalec samodejno zgenerira primeren tip ter stanje funkcije oz. funkcionalnosti na grafičnem vmesniku v HTML datoteki, glede na prejete podatke. To stori s pomočo knjižnice jQuerry. Funkcije na strežniku krmilimo preko JSON POST formata, kjer je za delovanje potrebna tudi primerena prošnja (*ang. Request*) na strežnik (glej sliko 5). Do teh funkcij lahko nato s poznavanjem URL-ja ter prošnje dostopamo tudi preko drugih storitev, v kolikor smo v istem ROS omrežju.



Slika 3: Komunikacijska shema robotske celice.

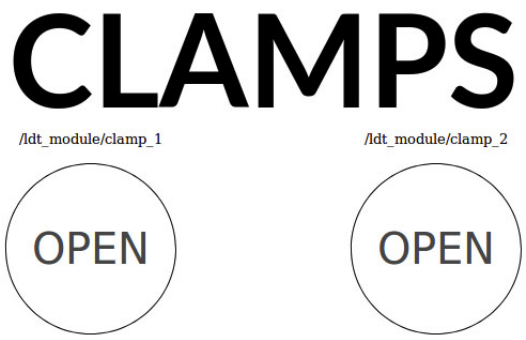
3.3 Spletni vmesnik

Stran je tako uporabniški vmesnik za modul. Primer vmesnika, ki je bil razvit za vpenjalni modul je viden na sliki 4.

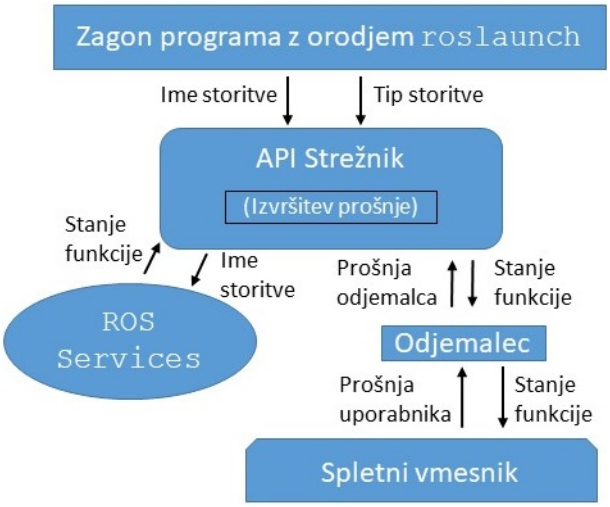
Prednost takšnega pristopa je, da uporabnik ne potrebuje ROS orodij za dostop do nekaterih funkcionalnosti celice. Dostop do funkcionalnosti vpenjalnega modula preko vmesnika med proizvodnim procesom seveda ni potreba, saj takrat celica deluje popolnoma samostojno oz. avtonomno. Taki vmesniki so predvsem uporabni med načrtovanjem ter programiranjem proizvodnega procesa.

4 Zaključek in nadalnje delo

V pričujočem članku smo predstavili podporno strojno in programsko opremo za module v rekonfigurabilni robotski celici. Pokazali smo, kako lahko s posebnimi "Priklopi in proizvajaj" konektorji dosežemo hitro izmenjavo



Slika 4: Uporabniški vmesnik za uporabljanje vpenjalnega modula.



Slika 5: Diagram poteka izmenjave informacije.

modulov celice in s tem tudi hitro spremembo njene namembnosti. Da bi zagotovili enostavnost uporabe modulov smo razvili *HTTP* strežnik, ki uporabnikom v spletni brskalnik streže uporabniški vmesnik. S tem smo pripomogli k dodatnemu lajšanju postopka programiranja dotične robotske celice.

Nadalnje delo vključuje vlaganje v razvoj avtonomnih modulov, ki bi se lahko v celico, preko *P&P* konektorjev sami priključili. Taki moduli bi lahko v industrijskih okoljih dodatno pripomogli k povečanju avtonomije rekonfiguracije robotske celice.

Literatura

[1] T. Dietz, U. Schneider, M. Barho, S. Oberer-Treitz, M. Drust, R. Hollmann, and M. Hägele, "Programming system for efficient use of industrial robots for deburring in SME environments," in 7th German Conference on Robotics; Proceedings of ROBOTIK 2012, pp. 1–6, VDE.

[2] R. M. Setchi and N. Lagos, "Reconfigurability and reconfigurable manufacturing systems: state-of-the-art review," in IEEE International Conference on Industrial Informatics (INDIN), pp. 529–535, 2004.

[3] Z. M. Bi, S. Y. T. Lang, M. Verner, and P. Orban, "Development of reconfigurable machines," The International Journal of Advanced Manufacturing Technology, vol. 39, pp. 1227–1251, Dec. 2008.

[4] M. Fulea, S. Popescu, E. Brad, B. Mocan, and M. Murar, "A literature survey on reconfigurable industrial robotic work cells," Applied Mechanics and Materials, vol. 762, p. 233, 2015.

[5] I.-M. Chen, "Rapid response manufacturing through a rapidly reconfigurable robotic workcell," Robotics and Computer-Integrated Manufacturing, vol. 17, no. 3, pp. 199–213, 2001.

[6] T. Gašpar, B. Ridge, R. Bevec, M. Bem, I. Kovač, A. Ude, and v. Gosar, "Rapid hardware and software reconfiguration in a robotic workcell," in 18th International Conference on Advanced Robotics (ICAR), pp. 229–236, IEEE.

[7] "ReconCell." `http://www.reconcell.eu/`. Accessed: 2018-07-10.

[8] "ROS - The Robot Operating System." `http://www.ros.org/`. Accessed: 2018-07-10.

[9] M. Bem, M. Deniša, T. Gašpar, J. Jereb, R. Bevec, I. Kovač, and A. Ude, "Reconfigurable fixture evaluation for use in automotive light assembly," in 2017 18th International Conference on Advanced Robotics (ICAR), pp. 61–67, July 2017.

[10] "NodeJS - JavaScript Node library." `https://nodejs.org/en/`. Accessed: 2018-07-11.

Simulacije in modeliranje

Simulations and Modeling

Approaching Industry 4.0 Simulation Modelling Paradigm with General Purpose Tools

Blaž Rodič[1]

[1]*Faculty of Information Studies in Novo mesto, Ljubljanska cesta 31A, Novo mesto, Slovenia*
E-mail: blaz.rodic@fis.unm.si

Abstract. *We aim to present the evolution of the simulation modelling paradigm as influenced by the Industry 4.0 paradigm and two real-life examples of the application of the new simulation modelling paradigm. Currently the development of a Digital Twin requires considerable resources and expertise, limiting the accessibility to SMEs. However, we demonstrate that data driven automated process modelling and integration of manufacturing information systems and simulation models is achievable without specialized tools or large budgets, and is thus within reach for research and development projects for SMEs.*

1 Introduction

The aim of this contribution is to present the evolution of the simulation modelling paradigm in connection with the Industry 4.0 paradigm and two real-life cases of the application of the new simulation modelling paradigm using off-the-shelf simulation modelling tools. The presented cases introduce methodologies and solutions which enable the automation and integration of general purpose simulation modelling tools by using data exchange standards such as XML, and the development of automation solutions using the Digital Twin concept with widely available sensor technologies.

The »Industry 4.0« term was coined by the German federal government in the context of its High-tech strategy in 2011. It describes the integration of all value-adding business divisions and of the entire value added chain with the aid of digitalisation. In the "factory of the future", the machines, sensors and other automation technologies are integrated via information and communication technology (ICT) into one system, the cyber-physical production system (CPPS). The CPPS also integrates the non-producing subsystems such as R&D as well as sales partners, suppliers, original equipment manufacturers (OEMs) and customers' information systems.

While the large industrial are concerned with the development of standardized methodologies and architectures that would allow integration within their R&D processes and existing ERP and MES solutions [1], and the purchase or development of automation technologies is not presented as problem, the SMEs have to consider using economical, off-the-shelf simulation modelling tools and commercially available sensors to build proprietary automation solutions, which would allow them to implement selected Industry 4.0 concepts, in order to remain a competitive supplier to their (larger) business partners.

Even though Industry 4.0 standards are still developed, the manufacturing and automation technology developers already market many technologies and solutions as "Industry 4.0". Currently the development of a Digital Twin requires considerable resources and expertise, limiting the accessibility of many solutions to bigger companies to the disadvantage of SMEs. [2] But even without specialized vendors there are many of the building blocks necessary for implementation of Industry 4.0 ideas already available – such as digital and networkable sensors and control elements (actuators), cloud computing, (industrial) communication networks, and general purpose simulation modelling tools as presented in this paper. [1]

2 The evolution of the simulation paradigm

Today, the use of simulation modelling in science and engineering is well established. In engineering, simulation modelling helps reduce costs, shorten development cycles, increase the quality of products and greatly facilitates knowledge management.

Several methods have been developed for mathematical modelling of real systems. Each of them was motivated by the problem itself and the researcher in that field. System dynamics (SD), which is based on cybernetics and system theory ([3] and [4]) and discrete event simulation (DES) were developed several decades ago and are today well-established. Agent based modelling (ABM) [5], [6] is a more recent methodology, and has gained attention outside social sciences due to its flexible abstraction level and applicability to modelling of complex systems [7]. In a simulation modelling project the methods are selected depending on the complexity of the modelled system and required abstraction level.

In the traditional simulation paradigm, the connectivity of a simulation model typically involves integration with a static database with business variables, a user friendly front-end and additional decision support tools such as expert systems (ES) or group decision support systems (GSS) [8]. The schematic of such a system is shown in Figure 1.

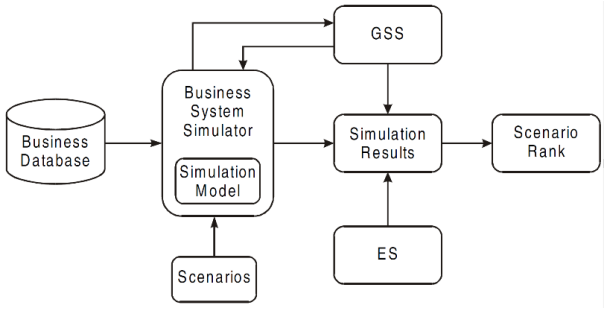


Figure 1: Schematic of a typical simulation modelling based decision support system [8]

2.1 New simulation modelling paradigm in Industry 4.0

Increasing product variants and products customized to order request more flexible production systems. The advent of the Industry 4.0 paradigm has brought changes to the simulation modelling paradigm as well. Part of the Industry 4.0 paradigm is modelling the manufacturing systems using the concept of a virtual factory or Digital Twin. These three points summarize the main changes to the simulation and modelling paradigm in the change from stand-alone simulation-based decision support system to the Digital Twin:

- Connectivity and integration in a wider IS (manufacturing or enterprise resource planning (MRP, ERP) is the norm,
- The modelled system is modelled with a holistic, multi-level/resolution approach, which includes physical modelling. Aspects of the simulation model require a high level of details and low level of abstraction,
- Modelling and modification of models is automated (data-based).

The concepts themselves are not new, the novel aspect is the unifying paradigm of Industry 4.0, which promises standardization of methods and the new technologies that make their implementation easier. In Industry 4.0 paradigm every instance of an individual product or production system produces a "digital shadow", which is the name for the structured collection of data generated by operation and condition data, process data, etc. Hence an instance of a Digital Twin consists of: a unique instance of the universal Digital Master model of an asset, its individual Digital Shadow and an intelligent linkage (algorithm, simulation model, correlation, etc.) of the two elements above [9].

Schematics of a decision support system incorporating the concept of Digital Twin is shown in Figure 2. In such systems, the Business System Simulator contains a Digital Twin model of the Business process. The Digital Twin is used to supply the array of decision support tools with a detailed, dynamically update digital representation of the real-life business process (e.g. a manufacturing plant). The process data is gathered real-time by the array of sensors and smart machines in the business process, stored in the business database and then transferred to the Digital Shadow. The Digital Master model's operation is adjusted according to the data in the Digital Shadow, allowing on-line optimization and decision support, and control of the process automation, creating a controlling feedback loop, which is the basis of cybernetic systems [4].

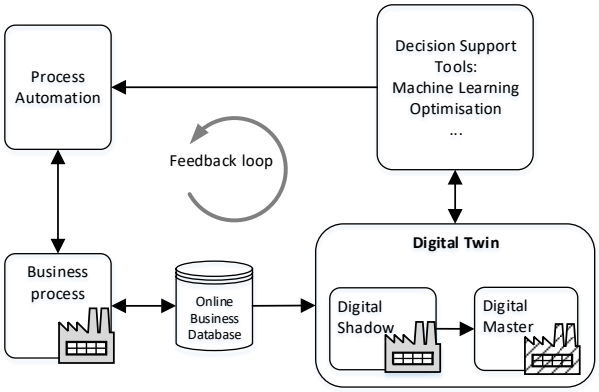


Figure 2: Schematic of a simulation modelling based decision support system implementing the Digital Twin

3 Implementing the simulation modelling paradigm

Implementing the new simulation modelling paradigm and the Industry 4.0 remains a serious challenge for researchers and companies. There are however novel ways to improve the integration of models built in general purpose simulation modelling tools, automate their construction and modification, and implement such solutions without major financial investments, which is a very attractive prospect especially for the SMEs. A number of solutions have been developed for automated generation of DES simulation models corresponding to manufacturing systems, with a good overview of solutions presented in [10]. We will present two cases of the implementation of the aspects of new simulation modelling paradigm, ranging from the development of new methods for data-driven automated simulation model construction to the development of a Digital Twin concept for SMEs.

3.1 Example 1: case of automated model building using XML and Java

This case involves a novel automated DES model construction method, using the customer order data obtained with SQL queries to modify the XML (Extensible Markup Language) file containing the simulation model structure and data. The method was applied in a manufacturing process optimisation project. Authors used discrete event simulation (DES) to build a model that reflects the current manufacturing processes and allows them to test optimisation methods.

Construction of a DES simulation model requires that the data that describe the manufacturing processes are

obtained, analysed, extracted and prepared in a suitable format for the model. In order to maintain model accuracy despite changes in manufacturing processes, integration of simulation software, auxiliary applications and databases is necessary. System optimisation through modification of model structure can be performed by constructing several versions of the model and input data (i.e. scenarios) and comparing simulation results. To accelerate the development of model versions and scenarios one can construct algorithms that build or modify simulation models according to model input data. This is especially useful in cases of large simulation models and if the model variants are prepared by an algorithm, e.g. an optimisation algorithm. Automated model building and modification however requires that the model structure can be modified with an algorithm, without manual interventions. [11]

Developing a static simulation model that would cover all possible (i.e. 30,000) products that may appear in client's orders is not realistic as it takes approximately 15 minutes to complete a model of a process for each product, and a model containing 30.000 process also exceeds the memory limitations of the modelling tool used (Anylogic, http://www.anylogic.com/). While other DES tools would also serve the purpose, Anylogic was used as its licence is already owned by the faculty and the models are stored in XML files, allowing relatively easy modifications to model structure. Manual modifications of the simulation model can be time consuming, especially if a large set of variations of the model needs to be built. In Anylogic, simulation model is typically constructed by adding different blocks and connections to the canvas by "click and drag" technique.

Instead, a method for ad-hoc model construction using a set of open orders was developed using Java. The Java application builds the model from a model template, the database of technical procedures and the database of currently open orders. Based on the list of ordered products and technical procedures only the necessary machines are placed in the model. Anylogic models are standard XML files, which allows easy manual or algorithmic modifications of the model. Anylogic XML simulation model file stores information on standard and user-defined blocks and agents, connectors between blocks, statistical monitors, input readers, output writers, etc. The data are stored as elements (nodes) and nested in a tree-like structure. An element can contain several attributes, describing type of the element and all the parameters describing element properties. The attributes can contain several lines of programming code describing how the block operates in different situations and states. [11]

The developed Java application manipulates XML code to change data on machines and all other relevant abstract objects such as connectors, sources and sinks that are connected to the blocks of machines. Specifically, the Java application reads the blocks in the template file and copies them according to input data. A new element (block) is added to the model by the following procedure:

- find a node representing a template block in XML tree according to the searched attributes,
- copy the node and connect it to the parent of the original node,
- change the data of the copied block (name of the block, position on the canvas, properties of the block, part of the programming code, etc.).

The resulting XML structure is then saved to a new Anylogic file. Products and carts play a role of transactions in DES and are therefore constructed dynamically during simulation. The resulting modelling and simulation system, shown in Figure 3 is composed of four main elements [11]:

- Core manufacturing process simulation model in Anylogic environment.
- Java application that constructs XML Anylogic model from a template file.
- MS Excel as an intermediate input and output data storage, and analysis tool. MS SQL server database describing technical procedures and client's orders.

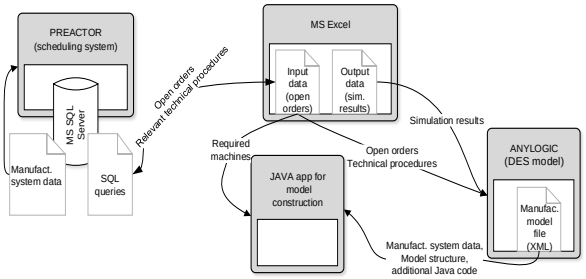


Figure 3. Automated DES model generation system schematic

3.2 Example 2: A Digital Twin for SMEs

Authors [12] present a concept for the realization of a Digital Twin of the production system within SMEs. Their concept is feasible by assuring sufficient data quality with minimized investment costs, and without compromising the advantages of the Digital Twin and of the CPPS. Their concept proposes a database structure and guidelines for the implementation of the Digital Twin in production systems in SMEs. The further concept of the Digital Twin for a production process enables a coupling of the production system with its digital equivalent as a base for an optimization with a minimized delay between the time of data acquisition and the creation of the Digital Twin. This allows the construction of a CPPS, opening up powerful applications. To ensure a maximum concordance of the cyber-physical process with its real-life model a multimodal data acquisition and evaluation has to be conducted.

Even though Industry 4.0 is one of the most prevalent subjects in production engineering, the methods of Industry 4.0 are still under-represented within manufacturing operations in SMEs. Authors furthermore describe the following main difficulties in the course of the realization of the Digital Twin in SMEs [12]:

- manual acquisition of motion data is widely used, but not with real-time availability, which limits the potential of simulation,

- in-house implementation of Industry 4.0 is frequently insufficient,
- slow standardization of data acquisition in productions systems hinders agile and adaptable system implementations,
- high costs for new IT and a centralised IS inhibit the application of vertical Industry 4.0,
- coupling of simulation and optimization is not sufficiently ensured to take full advantage of near real-time models, and
- data security concerns.

As the database of production data in SME is extremely heterogeneous, and its quality regularly insufficient for the realization of the Digital Twin, the authors [12] introduced sensor based tracking and machine vision for manufacturing process data acquisition. Sensor-based tracking provides information regarding routes and position of production employees and routes and position of large and highly mobile production devices, e.g. forklifts. The required technologies, i.e. sensor-based tracking systems and extensive program libraries for the machine vision implementation are commercially available, and therefore the implementation of the proposed concept is feasible.

In the last step, the data acquisition hardware is implemented into a real model process and linked with the data layer. This forms the final and major step within the realization of the CPPS in SME as part of the presented concept. The schematic of the proposed Digital Twin concept is shown in Figure 4.

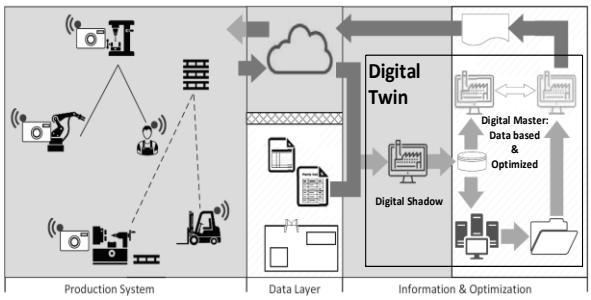


Figure 3. Concept of the implementation of a Digital Twin in SMEs (adapted from [12])

The described concept is novel in comparison to approaches prevalent in large enterprises, which focus on full automation. Automated gathering of machine data is not considered, as the low degree of digitalisation of manufacturing in SMEs data does not allow it. Furthermore, the collection of detailed machine data is not required with the presented concept. The innovative aspect of the concept is in the integration of well adopted and commercially available components, which are already available as isolated solutions [12].

4 Conclusion

The research presented in this paper includes novel solutions, that allow researchers and engineers to develop solutions that automate the model generation and solution seeking aspects of simulation based decision support and engineering systems. Presented methodologies and solution allow the development of Industry 4.0 automation solutions using the Digital Twin concept with widely available sensor technologies.

The adoption of new simulation modelling paradigm in research environment requires closer cooperation with industry partners, and diversification of knowledge of researchers, in order to build integrated, multi-level models of systems. As shown by the presented examples, lack of tools is not a problem, as the current generation of general purpose simulation modelling tools offer sufficient integration options. As the CPPS concept involves the integration of diverse information systems and multi-level simulation models, the Industry 4.0 and Digital Twin concept present researchers with a new motivation for closer cooperation with industry and transfer of knowledge between research groups and institutes.

Literature

[1] KPMG, "The Factory of the Future - Industry 4.0: The challenges of tomorrow," 02 06 2016. [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/pdf/2016/05/factory-future-industry-4.0.pdf. [Accessed 20 4 2017].

[2] S. Jain and D. Lechevalier, "Standards based generation of a virtual factory model," in *Proceedings of the 2016 Winter Simulation Conference (WSC '16)*, Piscataway, 2016.

[3] J. W. Forrester, Industrial Dynamics, Cambridge, MA: MIT Press, 1961.

[4] M. Kljajić, Teorija sistemov, Kranj: Moderna organizacija, 2002.

[5] N. Gilbert, Quantitative Applications in the Social Sciences: Agent-Based Models, London: SAGE, 2007, p. 1–20.

[6] A. Borshchev, The Big Book of Simulation Modeling, AnyLogic North America, 2013.

[7] A. Fereidunian, A. Lesani, M. Zamani, A. Kolarijani, N. Hassanpour in S. Mansouri, „A Complex Adaptive System of Systems Approach to Human–Automation Interaction in Smart Grid," v *Contemporary Issues in Systems Science and Engineering*, Piscataway, IEEE Press, 2015, pp. 425-485.

[8] M. Kljajić, I. Bernik and A. Škraba, "Simulation Approach to Decision assessment in Enterprises," *Simulation*, vol. 75, no. 4, pp. 199-210, 2000.

[9] R. Stark, S. Kind in S. Neumeyer, „Innovations in digital modelling for next generation manufacturing system design," *CIRP Annals - Manufacturing Technology*, 2017.

[10] P. Barlas and C. Heavey, "Automation of Input Data to Discrete Event Simulation for Manufacturing: A

Review," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 7, no. 1, 2016.

[11] B. Rodič in T. Kanduč, „Optimisation of a complex manufacturing process using discrete event simulation and a novel heuristic algorithm," *International Journal Of Mathematical Models And Methods In Applied Sciences*, Izv. 2015, št. 9, 2015.

[12] T. H.-J. Uhlemann, C. Lehmann in R. Steinhilper, „The Digital Twin: Realizing the Cyber-Physical Production System for Industry 4.0," v *Procedia CIRP: Proceedings of The 24th CIRP Conference on Life Cycle Engineering*, Kamakura, 2017.

[13] J. D. Sterman, Business dynamics: Systems Thinking and Modeling for a Complex World, Boston: Irwin/McGraw-Hill, 2000.

[14] S. Robinson, Simulation: the practice of model development and use 2nd ed., Basingstoke/New York: Palgrave Macmillan, 2004.

Analysis of mesh density in electrical simulations of heterojunction silicon solar cells

Jošt Balent, Janez Krč and Marko Topič

University of Ljubljana, Faculty of Electrical Engineering,
Tržaška 25, 1000 Ljubljana
E-mail:jost.balent@fe.uni-lj.si

Abstract. *Solar cells convert solar energy into electricity. Numerical modelling has been proven as an efficient tool in the process of optimisation of solar cell performance. In this contribution we focus on electrical modelling of a silicon heterojunction solar cell, employing the Sentaurus TCAD simulation tool. In particular we studied the impact of an adaptive mesh density on accuracy of the J-V characteristics and the main output parameters of the solar cell. It is shown that, for a broad range of densities (134 – 15962 discrete elements), parameters vary below 3 % for the analysed solar cell structure. Relatively low variation is a consequence of the adaptive character of the mesh – denser mesh in the vicinity of interfaces and in the regions of high gradient changes of the internal quantities, such as defect regions.*

1 Introduction

Energy demand is on the rise as population soars above 7.6 billion. Meeting this demand by fossil or nuclear fuels proved to be unsustainable. Photovoltaics (PV) presents a promising renewable energy solution to supply the world with clean and low-cost electrical energy as it uses solar cells to convert sunlight directly into electrical energy. Among various PV technologies [1] crystalline silicon (c-Si) wafer-based solar cells are dominating the market today (>90 %) [2], with commercial PV modules reaching conversion efficiency in the range of 17–20 %, whereas the research cell record efficiency is currently 26.7 % [3]. This record-braking Si cell design has a structure of a so called heterojunction (HJ-Si) device. The term hetero refers to the use of different energy band-gap materials. In particular, HJ-Si combines the advantages of Si thin-film and wafer-based technology. The HJ-Si solar cells are also the subject of our numerical investigation in this paper. State-of-the-art HJ-Si devices are quite complex, both optically and electrically, as they are comprised of ultra-thin (few nanometres) and thick layers (hundreds of micrometres), made of semiconductor, oxide and metal materials. Various textures and contacting schemes are applied to obtain superior performance of the device [1]. Numerical modelling proved to be an indispensable tool in the R&D cycle as it offers an in-depth analysis of such complex devices which consequently leads to cost reduction and improvement of efficiency.

In this paper we are going to use both optical (Sun*Shine* [4]) and electrical (Sentaurus TCAD [5]) numerical simulations, with the emphasis on the latter, to simulate optoelectronic properties of the HJ-Si solar cell. In particular we will show the importance of meshing density when doing finite element method analysis (FEM) on such layered structures. This analysis is a key step towards reliable simulations of realistic devices and further optimization of their operational performances.

2 Structure and principle of operation of a HJ-Si device

The HJ-Si cell (Figure 1) employs a lightly doped p or n type crystalline silicon (in our case n-type) as an absorber with a thickness of around 100 – 200 μm. Both sides of the absorber are capped with a thin (1-5 nm) passivation layer (typically intrinsic hydrogenated amorphous silicon, i-a-Si:H or other alternative material) in order to passivate c-Si surface defects - fill the free, dangling bonds of Si. Passivation layers are usually covered with a highly doped p-type (one side) and n-type (other side) a-Si:H material (10–50 nm thick) to assure selectivity of electron/hole extraction and to supply them to the front (top) and rear (back) contact.

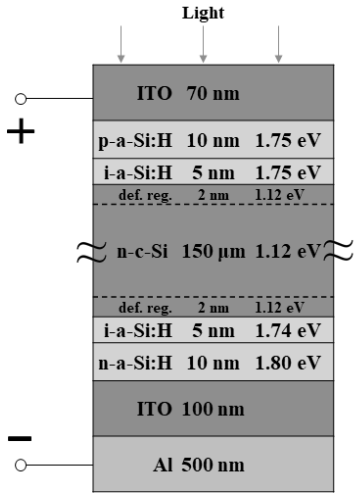


Figure 1: Structure of a typical HJ-Si solar cell.

A highly doped n-type transparent conductive oxide (TCO) is used as a front contact in combination with a metal grid (the grid is not shown in the figure). Indium tin oxide (ITO) is usually employed in HJ-Si devices as the TCO. Also at the rear side a TCO layer is often used in combination with Al or Ag metal contacts to ensure good optical and electrical characteristics of the rear part of the device. A cell with contacts solely at the rear has also been developed in the form of an Inter-digitated-back contact (IBC) scheme (not shown here). The device employs some form of light trapping by

interface micro-texturing and anti-reflection coating (ARC) to maximize the absorption of light in c-Si and minimize the reflection loss (in Figure 1 the texturing of the interfaces is not shown, but it was considered in optical simulation of the device). Optical simulations were performed on the entirety of the schematic shown in Figure 1, whereas electrical simulations omitted the ITO layers at this stage and assumed ideal ohmic contacts instead.

3 Numerical models

In electrical simulations with the Sentaurus TCAD we used the drift-diffusion based model to solve the transport equations of free electrons and holes in the HJ-Si device. The densities of electrons and holes were calculated with Fermi-Dirac statistics in both low and highly doped regions. Mobility and diffusivity were modelled by the Einstein relation. We also included the Masetti model for doping dependence of mobility [6]. Band-gap narrowing effects were considered negligible and were not included in the simulations. We incorporated Auger, Shockley-Read-Hall (SRH) and radiative recombination mechanisms. Passivation quality at front and rear surfaces of the c-Si absorber was modelled by the 2 nm thick defective regions (def. reg. in Figure 1). In these two regions amphoteric Gaussian-like trap distributions were introduced, similar to those used in modelling of the a-Si:H layers, which included exponential tail-states and Gaussian-like dangling bonds. Amphoteric nature of dangling-bonds was modelled as donor and acceptor Gaussian functions whose peak trap densities were apart by correlation energy. Quantum tunnelling was included at layers thinner than 10 nm. The model used is based on Wentzel–Kramers–Brillouin (WKB) approximation and takes into account only intra-band tunnelling. Doping ionization was assumed to be 100 %. All material parameters were taken from our previous publication [7]. The generation rate profile across the structure, due to absorbed sunlight, was calculated by the Sun*Shine* simulator.

The performed electrical simulations of the solar cell were two-dimensional (2-D), although the structure of the device would allow for a 1-D simulation (device planar dimensions are much larger than its thickness and all parameters are also uniform in the planar directions). However, a 2-D model was used instead since it will be required in our further work, where the device structure will be changed (e.g. different lateral contacting schemes). Here, a 2-D model with appropriately small planar dimensions was applied.

In finite element method (FEM) linear interpolation between the calculated points (nodes of the discrete triangles) was applied, which was in our previous studies indicated to be an effective solution compared to higher order interpolations. In the analysis presented here the attention was paid on variation of mesh density - variation of number of discrete elements in 2-D FEM simulations with the Sentaurus TCAD tool. In case of a 2-D simulation these elements are represented by triangles. There is a trade-off between validity and accuracy of simulations on one hand and computational time and processing power of computing on the other. Many simulations have to be performed during the optimization process of a device, therefore it is very important to find optimal settings for reliable, accurate and time-efficient simulations.

4 Simulation results

Figures of merit of a solar cell performance are the current density - voltage (J-V) characteristic and the corresponding parameters under full sun (1000 W/m^2) illumination:
the short-circuit current density (J_{sc}), the open-circuit voltage (V_{oc}), the maximum power-point (MPP) current density and voltage (J_{mpp}, V_{mpp}), the fill-Factor (*FF*, see definition in Figure 2) and the conversion efficiency (η)

To demonstrate these characteristics we show in Figure 2 the *J-V* curve obtained by the optoelectrical simulations of the HJ-Si solar cell where 5654 discrete triangular elements were used in electrical simulations. As shown later, this number of elements used in an adaptive mesh is recognized as a proper discretization setting in simulations of the device under scope.

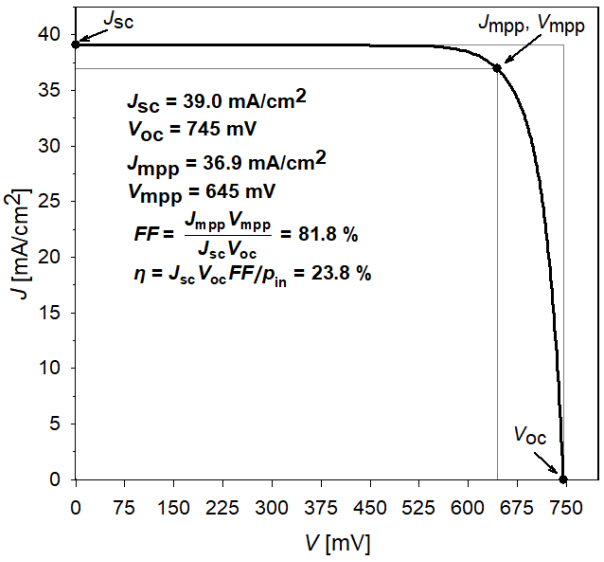


Figure 2: Simulated *J-V* curve and parameters of the HJ-Si cell. 5654 of discrete elements were used in the simulation.

The J_{sc} is the maximum current density a cell can produce under given illumination (short-circuit condition, load with zero resistance). The V_{oc} is the maximum voltage obtained (under open circuit condition-no load). Fill-factor tells us what percentage of the product of J_{sc} and V_{oc} we can utilize when applying optimal load to obtain the maximum power point (MPP) condition. Conversion efficiency, as a final parameter, defines how much of the incoming solar energy is being converted into electricity. All parameter values shown in Figure 2 are given for the simulated solar cell reaching a conversion efficiency of 23.8 % (the cell was not optimized).

4.1 Analysis of mesh density in electrical simulations

In FEM analysis mesh density can play a crucial role when it comes to validity and accuracy of simulations. It is prudent to use denser mesh near material interfaces or regions with high gradients of internal quantities, such as defect densities near interfaces.

The Sentaurus TCAD tool allows for a definition of an adaptive mesh according to various conditions (doping density gradients, material or region interfaces etc.). Meshing starts by defining maximum and minimum triangle element dimensions (*x*-lateral and *y*-vertical axis for 2-D). In our case minimum and maximum values for the *x* axis were both equal to 1 μm. Depth of the device was assigned to the *y* axis and element dimensions were restricted to the interval between 1 pm–10 μm. We implemented an adaptive mesh for all material interfaces by defining a starting triangle element height–y_0 and an enlargement factor-*k* when going away from the interface or high gradient region. The adaptive mesh algorithm uses a material interface as its starting point where it applies triangles with the height y_0 (see Figure 3). Next step repeats the procedure by multiplying y_0 with k and forming larger triangles. The height *y* remains constant when it reaches the maximum value defined at the beginning of this section.

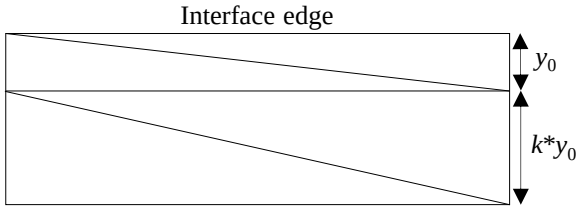


Figure 3: First two steps of the adaptive mesh algorithm.

In our analysis 20 different mesh densities were included with the total number of discrete elements ranging from 134 to 15962 obtained by varying the factor *k* from 1.005 to 200 (y_0 = 1 pm and was kept constant) An example of low, medium and higher schematic mesh densities can be seen in Figure 4.

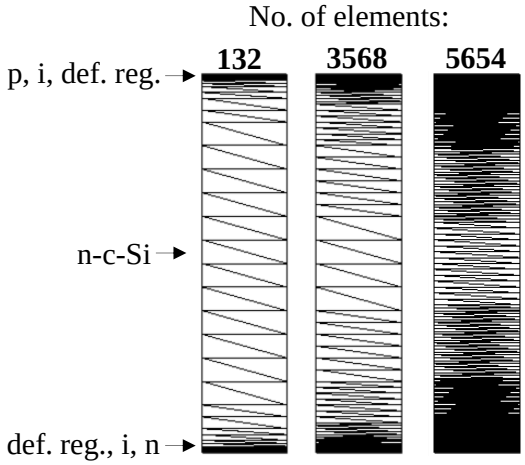


Figure 4: Examples of different mesh densities.

Results of our analysis are shown in Figure 5 (a-d). Simulated values of the parameters J_{sc}, V_{oc}, *FF* and η are presented as a function of number of mesh elements. The span of No. of elements is relatively broad. We can see that the results of all parameters vary the most for the smallest No. of elements.

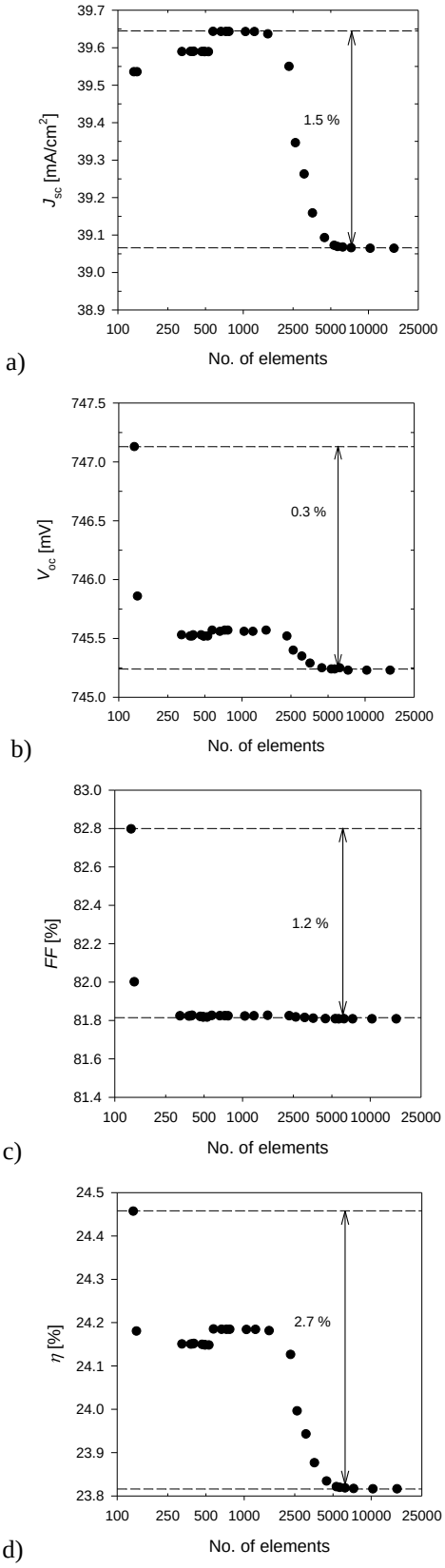


Figure 5: Parameters of the solar cell vs No. of elements.

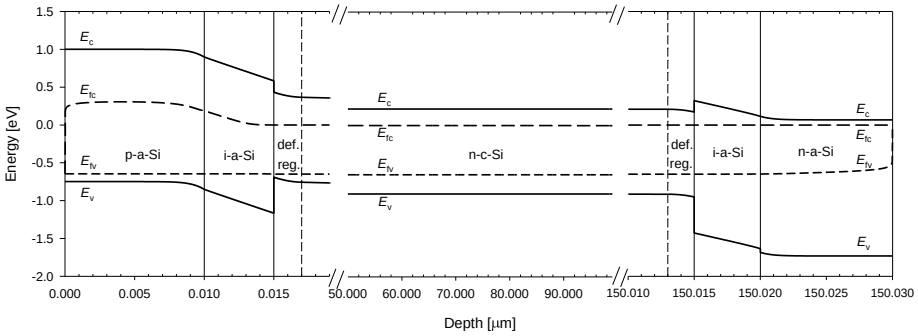


Figure 6: Energy-band diagram of a reference cell at MPP with 5654 elements.

The values of J_{sc} (Figure 5a) and η (Figure 5d) settle down in two regions (250 – 500 and 550 – 1100) before they finally reach their constant value after 5654 No. of elements.

The values of V_{oc} vary slowly after the first two points up until 2500 No. elements when the value changes to the final and constant value at 5654.
Values for *FF* are quite constant right after the first two points due to it being a kind of composed parameter (in the definition products of *J* and *V* are present) and the variations of the components can cancel each other out.

Maximum relative variation was observed for η and was 2.7 %, followed by J_{sc} with 1.5 %, FF with 1.2 % and V_{oc} with 0.3 %. We have to point out that these values are really low (low sensitivity to the mesh density) due to the utilization of an adaptive mesh. Mesh density as low as 134 elements can be used when doing preliminary research, but for an accurate and still time efficient simulation we recommend at least 5654 elements (triangular element dimensions range from 1.15e-6 μm^2 -1.17 μm^2). We also have to take special care if the gradients of internal parameter distribution change significantly during optimization – in this case the mesh has to be changed accordingly during the optimization process.

4.2 A view into internal quantities

After our simulations pass the test of validity and accuracy we can use them to get an inside view in the device. Distributions of internal quantities, which often cannot be accessed by measurements, can be determined by simulations. This presents a key advantage when we do further optimisation of the devices. These internal quantities are for instance the energy band diagram, free charge carrier densities, recombination rates, electric field and others. Here we show, as an example, only the energy band diagram (Figure 6) calculated with 5654 elements and under MPP. The result of the 2-D simulation is presented as a function of the vertical direction (depth) only, belonging to a vertical cross-section taken at the middle position of the 2-D structure. Results corresponding to cross-sections taken at different positions differ only in the range of numerical errors. From the diagram we can see, for example, how conduction and valence bands change throughout the front and rear heterojunctions. We can also see the splitting of quasi-fermi levels which is the voltage we have at the terminals (V_{mpp}). The shape of the bands at material interfaces (i-a-Si regions in our case) can tell us the level of carrier selectivity (energy barrier for electrons in p,i-a-Si regions at the front and an energy barrier for holes in i,n-a-Si at the rear). We can also see regions where there are high gradients of internal quantities which allows us to redefine the mesh at this critical points to insure accurate simulations (defect region interfaces in our case)

5 Conclusions

In numerical simulation of solar cells, mesh type and density play an important role trading off accuracy and validity for time efficiency and computational power. We showed, that by using an adaptive mesh in electrical simulations of a HJ-Si solar cell, main output parameters vary with mesh density (134-15962 No. of elements) up to 2.7 % of maximum relative difference. The values of all parameters become constant when we use at least 5654 elements in the mesh. When an adaptive mesh is used, we can do preliminary optimization with as few as 134 points for a structure shown in Figure 1. To perform time efficient and accurate simulations, an adaptive mesh of at least 5654 points are to be used.

Literature

[1] A. Luque and S. Hegedus, Eds., *Handbook of photovoltaic science and engineering*, 2nd ed. Chichester, West Sussex, U.K: Wiley, 2011.

[2] F. ISE, 'Photovoltaics report', vol. 2016, p. 45.

[3] K. Yamamoto *et al.*, 'Record-breaking efficiency back-contact heterojunction crystalline si solar cell and module', *33rd European Photovoltaic Solar Energy Conference and Exhibition*, vol. 2017, p. 4.

[4] J. Krč, S. Franc, and T. Marko, 'One-dimensional semi-coherent optical model for thin film solar cells with rough interfaces', in *Informacije MIDEM 32*, Ljubljana, vol. 2002.

[5] Sentaurus Device, https://www.synopsys.com/silicon/tcad/device-simulation/sentaurus-device.html

[6] G. Masetti, S. Maurizio, and S. Sandro, 'Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-Doped Silicon', *IEEE Transactions on Electron Devices*, no. 7, p. 6, Jul. 1983.

[7] J. Balent, J. Krč, A. Čampa, F. Smole, and M. Topič, 'Electrical simulations of hetero-junction silicon solar cells', in *53rd International Conference on Microelectronics, Devices and Materials & the Workshop on Materials for Energy Conversion and their Applications: Electrocalorics and Thermoelectrics*, Ljubljana, Slovenia, 2017.

Formiranje potenciala pred ravno elektrodo, ki oddaja elektrone

Tomaž Gyergyek[1,2], Jernej Kovačič[2], Iñaki Gomez[2]

[1]*Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija*
[2] *Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia*
E-pošta: tomaz.gyergyek@fe.uni-lj.si

Potential formation in front of an electron emitting collector planar electrode

Abstract. *Formation of a virtual cathode in front of an electron emitting floating collector that terminates a bounded plasma system is studied with a particle-in-cell (PIC) computer simulation. As the electron emission from the collector is increased the depth of the potential well increases also. As long as the emission is below approximately 40 times the critical emission level the depth of the potential well follows very well a simple logarithmic formula that is derived. For larger emissions the depth of the virtual cathode increases faster than predicted by the formula. The reasons for this discrepancy remain to be further investigated.*

1 Uvod

Modeliranje plazemskih diod, kjer so hitrostne porazdelitvene funkcije nabitih delcev odrezane Maxwellove je precej živahno področje [1,2] predvsem zaradi praktičnih aplikacij (na primer emisijske sonde [3], v zadnjem času pa tudi fuzijske aplikacije [4]) Krajši pregled literature je v [5]. V tem delu preučujemo formiranje tako imenovane virtualne katode, ki nastane pri zelo močni emisiji elektronov iz elektrode.

2 Model

Eno dimenzionalna plazemska dioda je omejena z dvema zelo velikima planarnima elektrodama, ki se nahajata na mestih $x = 0$ (na levi) in $x = L$ (na desni) – slika 1. Desna elektroda je ozemljena in njen potencial vzamemo za ničlo merjenja potenciala. Ta elektroda (izvir) v sistem vbrizgava protone (indeks *i*) in elektrone (indeks 1). Hitrostna porazdelitev obeh vrst vbrizganih delcev je polmaksvelska z različnima temperaturama T_i in T_e. Levo elektrodo imenujemo kolektor, ker absorbira vse delce, ki jo zadenejo. Obenem pa ta elektroda emitira elektrone (indeks 2), ki izstopajo iz nje s polmaxwellsko hitrostno porazdelitvijo s temperaturo T_2. Leva elektroda je električno izolirana od desne in se glede na njo nabije na potencial lebdenja, ki je odvisen od tega koliko pozitivnih in negativnih delcev jo zadene oziroma zapusti na časovno enoto. Predpostavljamo, da v prostoru med elektrodama delci ne trkajo med seboj in da se torej skupna kinetična in potencialna energija vsakega delca ohranja. Ion, ki desno elektrodo zapusti z ničelno začetno hitrostjo ima na mestu *x* hitrost:

$$v_{mi} = -\sqrt{-\frac{2e_0\Phi(x)}{m_i}}, \qquad (1)$$

v smeri proti kolektorju. Zato ima ta hitrost negativen predznak. Pri tem je m_i masa iona, e_0 je osnovni naboj, $\Phi(x)$ pa je potencial na mestu *x*. Potencial desne elektrode postavimo na ničlo, $\Phi(x=L) = 0$ kot referenčni potencial, potencial na katerem lebdi kolektor pa je Φ_C. Ker so elektroni mnogo bolj gibljivi, kot protoni, lahko pričakujemo, da je potencial Φ_C negativen glede na potencial desne elektrode. Zanimajo nas samo takšni potencialni profili $\Phi(x)$, ki monotono pojemajo od desne proti levi. Potencial $\Phi(x)$ je torej povsod v sistemu negativen. Porazdelitveno funkcijo za ione zapišemo takole:

$$f_i = n_{Si}\sqrt{\frac{m_i}{2\pi kT_{Si}}}\exp\left(-\frac{e_0\Phi(x)}{kT_{Si}}\right)\exp\left(-\frac{m_i v^2}{2kT_{Si}}\right)H(-v+v_{mi}). \qquad (2)$$

Pri tem je n_{Si} gostota, T_{Si} temperatura ionov tik ob izviru, *v* pa hitrost. Boltzmannova konstanta je *k* in *H* je Heavisideova funkcija enotske stopnice.

Elektron, ki zapusti kolektor z ničelno začetno hitrostjo, ima na mestu *x* hitrost

$$v_{me} = \sqrt{\frac{2e_0\left(\Phi(x)-\Phi_C\right)}{m_e}}, \qquad (3)$$

v smeri proti izviru in ta hitrost je torej pozitivna. Porazdelitveno funkcijo elektronov, ki so prišli z izvira torej zapišemo takole:

$$f_1 = n_{S1}\sqrt{\frac{m_e}{2\pi kT_{S1}}}\exp\left(\frac{e_0\Phi(x)}{kT_{S1}}\right)\exp\left(-\frac{m_e v^2}{2kT_{S1}}\right)H(-v+v_{me}). \qquad (4)$$

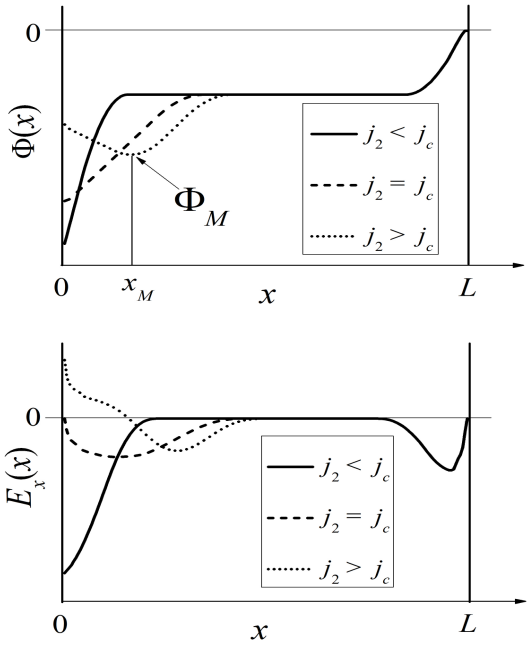
Tu je n_{S1} gostota, T_{S1} pa temperatura elektronov ob izviru. Masa elektrona je m_e. Porazdelitvena funkcija elektronov, ki jih emitira kolektor, je:

$$f_2 = n_{C2}\sqrt{\frac{m_e}{2\pi kT_{C2}}}\exp\left(\frac{e_0\left(\Phi(x)-\Phi_C\right)}{kT_{C2}}\right)\exp\left(-\frac{m_e v^2}{2kT_{C2}}\right)H(v-v_{me}). \qquad (5)$$

Podobno, kot prej je n_{C2} gostota, T_{C2} pa temperatura emitiranih elektronov ob kolektorju.

Vse zgoraj povedano velja, dokler je potencialni profil v sistemu monoton. Če pa emisija elektronov iz kolektorja močneje narase, lahko potencialni profil postane nemonoton. Znano je, da potencial lebdenja

elektrode, ki emitira elektrone narašča, če narašča emisija. V nekem trenutku dosežemo tako imenovano kritično emisijo elektronov, ko električno polje ob elektrodi doseže ničlo. Če pa emisijo elektronov še povečamo, nastane pred elektrodo potencialna jama, ki ji pogosto pravimo virtualna katoda. Električno polje pred elektrodo takrat spremeni predznak v pozitivnega in začne pospeševati emitirane elektrone nazaj proti elektrodi. Tiste elektrone, ki zapustijo elektrodo s premajhno začetno hitrostjo, celo vrne nazaj na elektrodo. Ti profili potenciala in električnega polja so shematično prikazani na sliki 1. Pri tem j_2 pomeni gostoto toka emitiranih elektronov iz leve elektrode, j_c pa kritično emisijo, pri kateri je električno polje ob kolektorju enako 0.



Slika 1: Shematični prikaz modela. Izvor se nahaja pri $x = L$, kolektor pa pri $x = 0$. Ko emisija postane kritična, je električno polje ob kolektorju enako nič. Ko pa emisija še narase, nastane pred kolektorjem potencialna jama – virtualna katoda.

Označimo vrednost minimuma potencialne jame Φ_M. Ta minimum se nahaja na mestu $x_M > 0$. Radi bi ocenili, kolikšen delež emitiranih elektronov lahko pride skozi potentialno bariero v področje $x > x_M$, če je njihova porazdelitvena funkcija podana z izrazom (5). Tik ob kolektorju je $v_{me} = 0$, ker je $\Phi = \Phi_C$. Porazdelitvena funkcija (5) torej dobi naslednjo obliko:

$$f_{2M} = n_{C2}\sqrt{\frac{m_e}{2\pi k T_{C2}}}\ \exp\left(-\frac{m_e v^2}{2kT_{C2}}\right)H(v). \quad (6)$$

Gre torej za enodimenzionalno polmaxwellsko porazdelitev, ki je shematično prikazana na sliki 2. Samo elektroni, ki imajo hitrost večjo kot

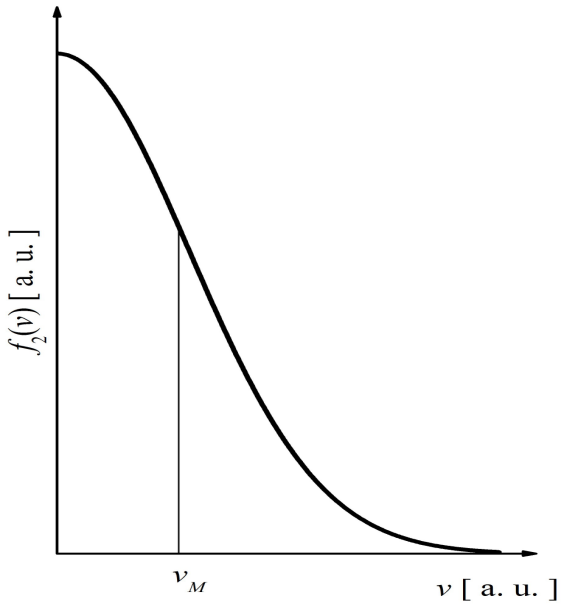
$$v_M = \sqrt{\frac{2e_0\left(\Phi_C - \Phi_M\right)}{m_e}}, \qquad (7)$$

lahko pridejo v področje $x > x_M$.

Uvedemo naslednje spremenljivke:

$$\mu = \frac{m_e}{m_i},\ \tau = \frac{T_{Si}}{T_{S1}},\ \sigma = \frac{T_{C2}}{T_{S1}},\ \Psi = \frac{e_0\Phi(x)}{kT_{S1}},\ \Psi_C = \frac{e_0\Phi_C}{kT_{S1}}, \quad (8)$$

$$\Psi_M = \frac{e_0\Phi_M}{kT_{S1}},\ \ \varepsilon = \frac{n_{C2}}{n_{S1}},\ v_0 = \sqrt{\frac{2kT_{S1}}{m_e}},\ u = \frac{v}{v_0}.$$



Slika 2: Porazdelitvena funkcija emitiranih elektronov ob kolektorju je polmaxwellska. Samo tisti elektroni, ki imajo hitrost večjo od v_M lahko pridejo v področje $x > x_M$.

S spremenljivkami (8) zapišemo porazdelitveno funkcijo (6) takole:

$$F_2 = \frac{\varepsilon}{\sqrt{\pi\,\sigma}}\ \exp\left(-\frac{u^2}{\sigma}\right)H(u).$$

Porazdelitvena funkcija F_2 je normirana na n_{S1} takole:

$$n_{S1}\sqrt{\frac{m_e}{2\pi k T_{S1}}}\int_{-\infty}^{\infty}\exp\left(-\frac{m_e v^2}{2kT_{S1}}\right)dv = n_{S1} \equiv 1.$$

Zapišimo še brezdimenzijske hitrosti:

$$u_{mi} = \frac{v_{mi}}{v_0} = -\sqrt{-\mu\Psi},\ u_{me} = \frac{v_{me}}{v_0} = \sqrt{\Psi - \Psi_C},$$

$$u_M = \frac{v_M}{v_0} = \sqrt{\Psi_C - \Psi_M}.$$

Gostoto toka emitiranih elektronov J_{2a} podaja izraz:

$$J_{2a} = \int_{-\infty}^{\infty} uF_2(u)du = \frac{\varepsilon}{2}\sqrt{\frac{\sigma}{\pi}}. \qquad (9)$$

Gostoto toka elektronov J_{2b}, ki pridejo v področje $x > x_M$ pa podaja integral:

$$J_{2b} = \int_{\sqrt{\Psi_C - \Psi_M}}^{\infty} uF_2(u)du = \frac{\varepsilon}{2}\sqrt{\frac{\sigma}{\pi}}\exp\left(-\frac{\Psi_C - \Psi_M}{\sigma}\right). \quad (10)$$

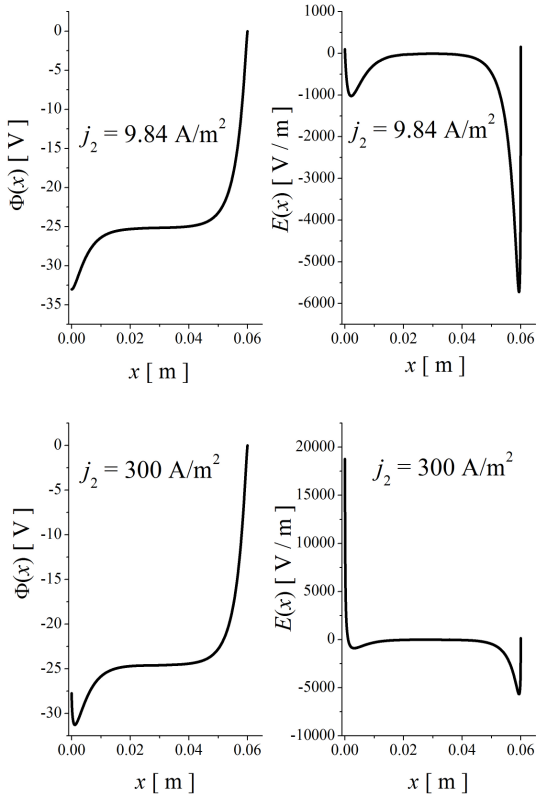
Ko med seboj delimo izraza (9) in (10) dobimo:

$$\frac{J_{2a}}{J_{2b}} = \exp\left(\frac{\Psi_C - \Psi_M}{\sigma}\right), \quad \Psi_C - \Psi_M = \sigma \ln\left(\frac{J_{2a}}{J_{2b}}\right). \quad (11)$$

Izraz (11) razumemo takole. Če poznamo gostoto toka elektronov, ki jih oddaja kolektor J_{2a} (na primer iz znane temperature kolektorja in Richardsonove formule) in če predpostavimo, da je J_{2b} enaka kritični gostoti toka J_c, pri kateri je električno polje ob kolektorju enako 0, potem lahko formulo (11) uporabimo za oceno globine potencialne jame, če seveda poznamo temperaturo emitiranih elektronov. Eksperimentalno je te podatke zelo težko izmeriti. V delčni simulaciji pa je to mnogo preprosteje. Zato v naslednjem razdelku primerjamo rezultate delčne simulacije s formulo (11).

3 Delčne simulacije

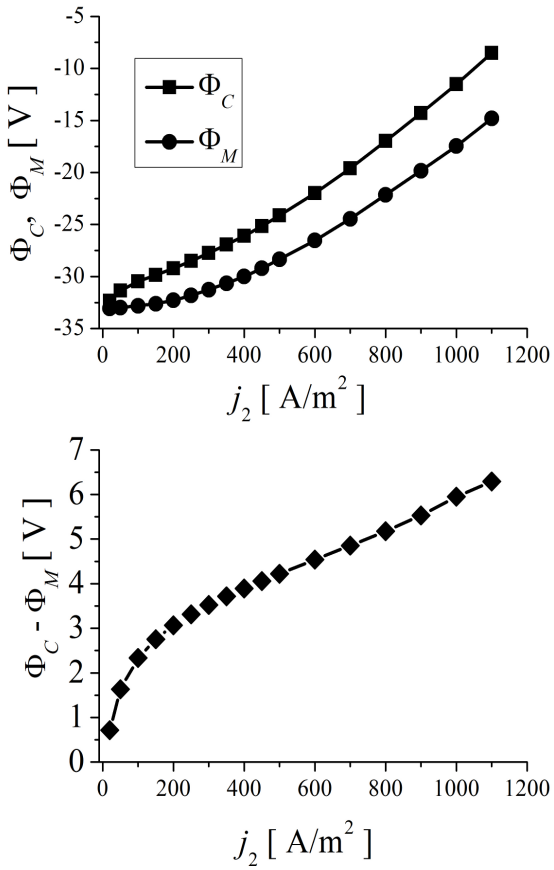
Plazemsko diodo, ki smo jo opisali v prejšnjem razdelku simularamo z delčno kodo XPDP1 [6]. Za pozitivne ione izberemo protone, njihovo temperaturo pa postavimo kT_{Si} = 0.1 eV. Temperaturo elektronov izberemo kT_{S1} = 20 eV. Pri emitiranih elektronih izberemo več različnih temperatur kT_{C2}. Dolžino diode nastavimo na 6 cm. Vbrizgavanje elektronov z desne elektrode nastavimo na j_1 = 60 A/m^2, vbrizgavanje emitiranih elektronov z leve strani j_2 pa počasi povečujemo. Pri vsakem paru j_1 = 60 A/m^2 in j_2 nato nastavimo vbrizgavanje ionov j_i tako, da je električno polje E_S ob desni elektrodi enaki nič. To električno polje določamo iterativno s strelsko metodo, kar je časovno zelo zamudno.



Slika 3: Dva primera profilov potenciala $\Phi(x)$ (levo) in električnega polja $E(x)$ (desno), ki ju dobimo iz simulacij za parametre kT_{S1} = 20 eV, kT_{C2} = 1 eV, kT_i = 0.1 eV, j_1 = 60 A/m^2 in pri 2 vrednostih j_2.

Na sliki 3 vidimo dva primera profilov potenciala in električnega polja, ki ju dobimo iz simulacij pri: kT_{Si} = 0.1 eV, kT_{S1} = 20 eV in kT_{C2} = 1 eV. Gostota toka vbrizganih elektronov je j_1 = 60 A/m^2, izberemo pa tudi 2 vrednosti za j_2 in sicer j_2 = 9.84 A/m^2 in j_2 = 300 A/m^2. Pri vsakem j_2 poiščemo takšno vbrizgavanje ionov j_i, da je električno polje E_S ob desni elektrodi kolikor je le mogoče blizu ničle. Vidimo, da je pri j_2 = 9.84 A/m^2 tudi električno polje ob levi elektrodi zelo blizu ničle, zato to vrednost emisije proglasimo za kritično emisijo j_c. Na spodnji levi sliki je formiranje virtualne katode zelo dobro vidno. Potencial lebdenja je Φ_C = -27.7 V potencial virtualne katode pa Φ_M = -31.3 V. Globina potencialne jame je 3.6 V.

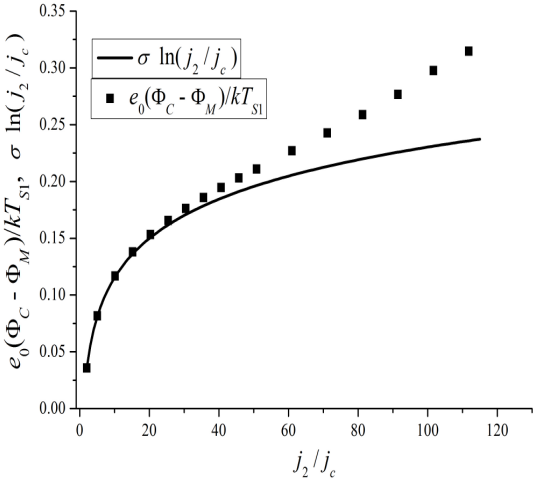
Na podoben način izvedemo še mnogo simulacij, kjer spreminjamo j_2 parametrov kT_{Si} = 0.1 eV, kT_{S1} = 20 eV, kT_{C2} = 1 eV in j_1 = 60 A/m^2 pa seveda ne. Emisijo j_2 spreminjamo med 20 A/m^2 in 1100 A/m^2. Vse te vrednosti krepko presegajo kritično emisijo j_c = 9.84 A/m^2. Pri vsaki vrednosti j_2 izmerimo potenciala Φ_C in Φ_M. Tako dobimo sliko 4. Z naraščajočim j_2 oba potenciala naraščata, narašča pa tudi razlika med njima - spodnji graf na sliki 4. Emisijo povečujemo do j_2 = 1100 A/m^2. Pri večji emisiji začne potencial v sistemu oscilirati in pride tudi do povratnega toka ionov proti desni elektrodi.



Slika 4: Potencial lebdenja kolektorja Φ_C in potencial virtualne katode Φ_M dobljena iz delčnih simulacij v odvisnosti od j_2 pri: kT_{S1} = 20 eV, kT_{C2} = 1 eV, kT_i = 0.1 eV in j_1 = 60 A/m^2.

Da bi mogli primerjati rezultate delčnih simulacij s formulo (11), moramo rezultate s slike 4 najprej

normirati. Potencial normiramo na kT_{S1}/e_0 = 20 V. Temperaturi kT_{S1} = 20 eV in kT_{C2} = 1 eV data σ = 0.05, emisijo j_2 pa normiramo s kritično emisijo j_c = 9.84 A/m^2, – glej sliko 3. Normirane rezultate simulacij prikazujemo z majhnimi kvadratki na sliki 5. Polna črta na sliki 5 pa kaže napoved formule (11). Vidimo, da je do $j_2/j_c \approx 40$ ujemanje med simulacijami in računskim modelom skoraj popolno. Pri večji emisiji pa formula (11) napoveduje manjšo globino potencialne jame Ψ_C - Ψ_M, kot jo dobimo iz simulacij. Razloga za to razhajanje zaenkrat še ne poznamo.



Slika 5: Normirana globina potencialne jame pred elektrodo v odvisnosti od normirane gostote toka emitiranih elektronov pri istih parametrih kot na sliki 4. Polna črta prikazuje rezultat formule (11), kvadratki pa rezultat delčne simulacije.

4 Sklep

S kinetičnim računskim modelom in delčnimi simulacijami smo preučevali nastanek virtualne katode pred električno lebdečo elektrodo, ki emitira elektrone. Za simulacije smo uporabili kodo XPDP1 [6]. Ko elektronska emisija preseže kritično vrednost pride do nastanka potencialne jame – virtualne katode. Izpeljali smo preprosto formulo, ki povezuje emisijske parametre z globino potencialne jame pred elektrodo. Njene napovedi smo primerjali z rezultati delčnih simulacij. Dokler emisija elektronov ne preseže približno 40-kratne kritične emisije, je ujemanje med simulacijami in napovedjo formule (11) skoraj popolno, pri večji emisij pa formula (11) napoveduje manjšo globino potencialne jame, kot jo pokažejo delčne simulacije. Razlage za to razliko zaenkrat še nimamo.

Simulacije pokažejo še en zanimiv rezultat. Ko narašča emisija, narašča tudi potencial lebdenja in ni opaziti nobenega nasičenja potenciala lebdenja ob veliki emisiji. Pri emisijskih sondah [3] namreč praviloma pride do nasičenja potenciala lebdenja ob naraščajočem gretju sonde, ko se potencial lebdenja dovolj približa potencialu plazme.

Zahvala

Zahvaljujemo se za finančno podporo ARRS v okviru programa P2-0073 in projekta BI-FR/CEA/17-19-002.

Literatura

[1] L. A. Schwager and C. K. Birdsall, “Collector and sorce sheaths of a finite ion temperature plasma”, *Phys. Fluids B*, vol. 2, no. 5, o. 1057-1068, 1990.

[2] L. A. Schwager, “Effects of secondary and thermionic electon emission on the collector and source sheaths of a finite ion temperature plasma using kinetic theory and numerical simulation”, *Phys. Fluids B*, vol. 5, no. 2, p. 631-645, (1993).

[3] J. P. Sheehan, N. Hershkowitz, “Emissive probes”, *Plasma Sources Sci. Techol.* vol. 20, no. 6, p. 063001, (22 pages), 2011.

[4] J.P. Gunn, S. Carpentier-Chouchana, F. Escourbiac, T. Hirai, S. Panayotis, R.A. Pitts, Y. Corre1, R. Dejarnac, M. Firdaouss, M. Kočan, M. Komm, A. Kukushkin, P. Languille, M. Missirlian, W. Zhao and G. Zhong, “Surface heat loads on the ITER divertor vertical targets” Nucl. Fusion vol. 57, p. 046025 (35 pages), 2017.

[5] T. Gyergyek, B. Jurčič-Zlobec, M. Čerček, “Potential formation in a one-dimensional bounded plasma system containing a two-electron temperature plasma: Kinetic model and PIC simulation”, *Phys. Plasmas*, vol. 15, no. 6, p. 063501 (28 pages), 2008.

[6] J. P. Verboncoeur, M. V. Alves, V. Vahedi and C. K. Birdsall, "Simultaneous Potential and Circuit Solution for 1d bounded Plasma Particle Simulation Codes," *J. Comput. Phys*., vol 104, no 2, p. 321-328, 1993.

Univerza v Ljubljani
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST IN ŠPORT
EVROPSKA UNIJA
EVROPSKI SOCIALNI SKLAD
NALOŽBA V VAŠO PRIHODNOST
Javni štipendijski, razvojni, invalidski in preživninski sklad Republike Slovenije

Identifikacija varne pristajalne cone

Aljaž Blažič[1], Klemen Kotnik[1], Kristina Nikolovska[1], Miha Ožbot[1], Martin Pernuš[1], Uroš Petkovič[1], Nika Hrušovar[2], Matic Verbič[2], Irena Ograjenšek[2], Andrej Zdešar[1], Matevž Bošnak[1], Tine Tomažič[3], Gregor Klančar[1]

[1] Fakulteta za elektrotehniko, Univerza v Ljubljani
[2] Ekonomska fakulteta, Univerza v Ljubljani
[3] PIPISTREL VERTICAL SOLUTIONS d.o.o

Identification of a safe landing zone

Abstract. *Recently, there has been a growing interest in developing autonomous lending systems which is the greatest challenge of the whole autonomous flight mission. This paper presents market research results of currently available sensors for environment recognition during landing. Main focus is given to development of algorithms for the safe-landing zone estimation using LIDAR and stereo-camera sensors. And finally, the paper highlights challenges in this research field.*

1 Uvod

Ob pojavu revolucionarne inovacije, kot je bil v preteklosti npr. avtomobil, ki je omogočil razvoj množične mobilnosti, se sedaj bliža nova inovacija na področju transporta, ki bo omogočala potovanje po zraku in občutno zmanjšala tako stroške kot porabljen čas. Eden ključnih izzivov avtonomnega delovanja letala VTOL (vertical take off and landing) je avtomatizacija pristajanja. Tu je potrebno identificirati primerno površino oziroma območje, na katerem lahko letalo zanesljivo pristane. Ta ne more biti odvisno samo od pilotovega pogleda iz kabine, ki lahko ima nepopolne informacije. Zato smo razvili predlog senzorskega sistema, ki bo sprva pilotu v pomoč v prihodnje pa omogočal avtonomno delovanje. Odločili smo se za povezovanje in nadgradnjo obstoječih rešitev zaradi ekonomije obsega, kjer se cena končnega izdelka s povpraševanjem zmanjšuje. Namreč vizija vključenega podjetja je izdelek tržiti na globalnem trgu.

Predstavljeno delo je nastalo v okviru PKP projekta (po kreativni poti do znanja) katerega osnovni cilj je vključitev študentov v reševanje konkretnih problemov katere definira sodelujoče podjetje. Sodelovalo je osem študentov iz Fakultete za elektrotehniko in Ekonomske fakultete Univerze v Ljubljani, ter pet mentorjev iz omenjenih fakultet in podjetja PIPISTREL VERTICAL SOLUTIONS d.o.o.

V prispevku je najprej predstavljena raziskava trga obstoječih zmogljivih senzorjev za namen razpoznavanja pristajalne cone. Nato sledi opis razvoja algoritmov, ki so uporabljeni za obdelavo podatkov, identifikacijo pristajalne cone iz meritev in ustrezne predstavitve rezultatov uporabniku. Pri razvoju smo se osredotočili na 3D LIDAR in stereo kamero, ki sta se izkazala za najprimernejša. Vrednotenje delovanja predlaganih sistemov je izvedeno na realističnih meritvah. Na koncu so izpostavljeni bodoči izzivi na tem področju raziskav.

2 Opis problema in predstavitev uporabljenih orodij

Cilj projekta je predlog avtonomnega sistema za pristajanje letala VTOL. Predlog vključuje izbiro ustreznih senzorjev, njihove primerne postavitve na letalo, obdelavo podatkov in izvedba osnovnih algoritmov. Najprej smo definirali osnovne predpostavke, izhodišča in omejitve. Sistem mora delovati od višine 50 metrov in manj. Končna odločitev za pristanek naj bo podana iz višine vsaj 15 metrov. Senzorji naj ne vsebujejo vrtečih delov zaradi robustnosti, njihov vpetje naj bo fiksno oz. je dovoljena le rotacija v eni smeri. Lahko se uporabi več senzorjev po trupu. Celotna cena senzorskega sistema naj ne preseže 10000€.

Pri delu smo uporabili sledeča orodja.

Matlab, kjer je potekal osnovni razvoj algoritmov. Okolje je primerno v fazi razvoja, saj vsebuje številne knjižnice, omogoča enostavno delo z matrikami, hitro implementacijo algoritmov, enostavne prikaze, digitalno obdelavo signalov, ipd. To programsko okolje je bilo uporabljeno za implementacijo algoritmov in obdelavo podatkov LIDARja in slik.

PCL (Point cloud library) je odprta knjižnica za obdelavo 2D in 3D slik in točk oblakov. Vsebuje algoritme za filtriranje, rekonstrukcije površine (modeliranje in segmentiranje), razpoznavanje objektov iz oblik. Celotna knjižnica je sestavljanja iz manjših knjižnic. Podpira stereo kamere, 3D skenerji, Kinect itd. PLC knjižnica je bila uporabljena pri obdelavi slik, ki so bile dobljene iz kamer.

Fusion 360 je program za 3D risanje. To je prvo 3D orodje, ki povezuje celoten proces razvoja izdelka (CAD, CAM in CAE) v eno platformo. Je odprt program za študente. Uporabljen je bil pri risanju predlaganih 3D modelov in postavitev kamer na 3D model letala.

3 Pregled senzorjev za namen avtonomnega pristajanja

Senzorji so glavni gradniki za izvedbo avtonomnega pristajanja letala. Pri izberi so imeli prednost senzorji, ki imajo vsaj 30 metrov dometa, nimajo vrtljivih mehanskih delov ter so dovolj robustni.

Cilj opravljene analize je bil ugotoviti, kaj trg senzorjev ponuja, katere tipe, kaj od tega je za nas uporabno, kaj je že v prodaji in kaj šele v razvoju. Pregledana je bila celotna ponudba (3D, 2D in 1D LIDAR, kamere, stereo kamere, IR kamere, radarji, ToF kamere, ultrazvočni senzorji itd.) od teh pa izbran najbolj primeren nabor za naš problem (glej Tabelo 1).

Ugotovljeno je bilo, da veliko ponudnikov trži rešitve, ki so še v razvoju in zato dejansko nedostopne in s pomanjkljivimi specifikacijami. Tovrstno razočaranje so bili Solid-State LIDARji, saj nimajo mehansko premikajočih delov in so zato bolj robustni in bi posledično bili najprimernejša izbira. Poleg nedostopnosti jim trenutno večinoma uporabo omejuje še majhno vidno polje, domet, resolucija in cena.

Zelo velik potencial imajo kamere (stereo), zaradi dostopnosti in informativnosti meritve oz. slik. Večji izziv pa predstavlja kakovostno obdelava slik in pridobivanje zanesljivih 3D slik oziroma oblaka točk na večjih razdaljah.

Delovanje IR kamer je odvisno od temperaturnega kontrasta okolice, ki lahko zelo niha v določenem območju. A kljub vsemu vrne 2D sliko, primerno za obdelavo in hitro analizo potencialnih objektov.

ToF kamere delujejo le na (pre-) kratkih razdaljah in v dobrih razmerah (lepo vreme). Dobra stran pa je, da neposredno meri 3-D informacijo.

Radarji lahko delujejo tako na kratkih, kot tudi dolgih razdaljah, a so omejeni s slabo resolucijo.

Izbira trenutno najprimernejših senzorjev za našo rešitev in dostopnih na trgu prikazuje Tabela 1.

Tabela 1. Izbrani senzorji

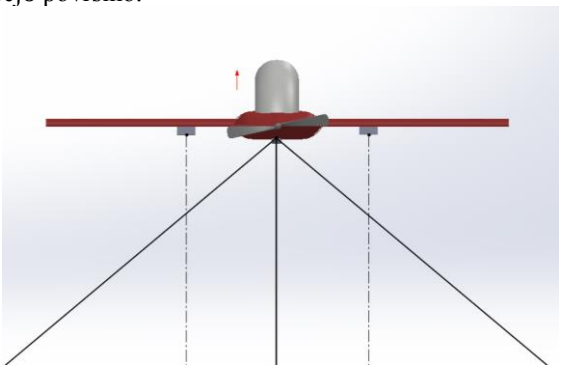
Tip senzorja	Senzor
Solid-State LIDER	Cepton HR80W
Stereo kamere	MultiSense S2
Infrardeca kamera	NightHawk 2
Time of Flight	03D303
Radar	Continental ARS4-A

Predlagani so štiri različni scenariji za zaznavanje okolice pristanka. Vsem je skupna osnovna ideja v pridobitvi 3D oblaka točk prostora, katerega nato obdelamo z namenom identificiranja dobrih in slabih ravnin oz. območij. Pri klasifikaciji ravnin se je za najprimernejšo izkazala metoda RANSAC (RANdom SAmple Consenses). Vsak scenarij vsebuje kombinacijo dveh različnih principov senzorjev, kjer slabosti enega odpravijo prednosti drugega. Narejene so naslednje kombinacije senzorjev:

- solid-state LIDAR kombiniran s stereo kamero,
- IR kamera kombinirana s stereo kamero,
- radar kombiniran s stereo kamero ter
- ToF kamera kombinirana s stereo kamero.

Vsi možni pristopi temeljijo na postopnem merjenju, kjer je uporaba senzorja namenjena za različne višine, s tem se izboljša natančnost in zanesljivost zajete meritve. Večje število senzorjev doda k redundanci, s tem se poveča varnost ob okvari in se pridobi več virov informacij za primerjavo z drugimi viri informacij.

Nekateri senzorji, npr. solid state LIDAR, imajo zelo ozko področje merjenja (FoV) kot je npr. 60°x24° pri modelu Cepton HR80W (Wide angle). Možna rešitev (manj robustna) je mehanski mehanizem, ki bi senzor pomikal v krožnem loku pri tem posnel več posameznih slik, katere naknadno združimo v globalno sliko terena. Druga rešitev (dražja in izvedbeno zahtevnejša) je postavitev več enakih senzorjev, ki hkrati pokrivajo večjo površino.

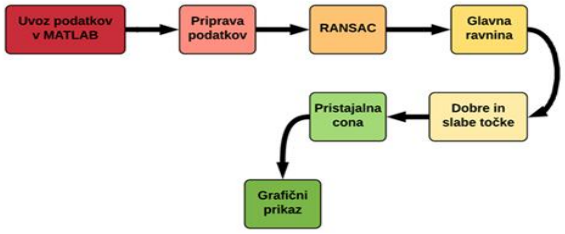


Slika 1. Postavitev senzorjev na letalu (LIDAR senzor na sredini in par stereo kamer na krilih)

4 Ocena varne pristajalne cone

4.1 Osnovni pristop

Bločni diagram potega obdelave podatkov od zajema meritev do grafičnega prikaza je podan na sliki 2. Le-ta je osnovan tako, da je skupen tako meritvam LIDARja kot stereo kamere s katerima je bila izvedena konstrukcija globinske slike.



Slika 2: Shema poteka obdelava podatkov

4.2 Izvedba meritev

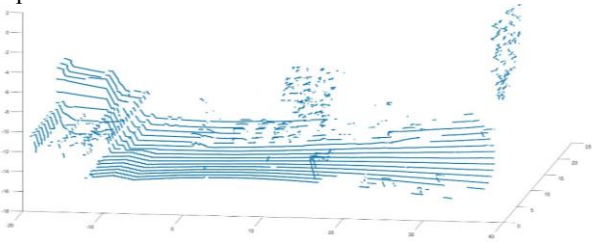
Razvoj in delovanje algoritmov smo izvajali izključno na realnih podatkih. Zajeti so bili z dveh različnih senzorjev LIDAR (Velodyne VLP-16) in stereo kamera (2x ACME VR06 na razdalji 1m). Meritve smo posneli iz okna v drugem nadstropju Fakultete za elektrotehniko (pogled na parkirišče, Slika 3).



Slika 3: Prikaz območja, ki je bilo merjeno.

4.3 Pristop z senzorjem LIDAR

Podatki iz LiDAR Velodyne VLP-16 so bili dobljeni v obliki pcap, katere se s programom VeloView preoblikuje v format csv (comma-separated values). Podatke uvozimo v MATLAB s pomočjo funkcije csvread, kjer je rezultat oblak točk v kartezičnem koordinatnem sistemu LIDAR senzorja (Slika 4). Velodyne VLP-16 meri 360° vidnega polja, a zanimivo je le specifično področje, tako da odstranimo nepomembne točke.


Slika 4: 3D globinska slika meritve LIDARja

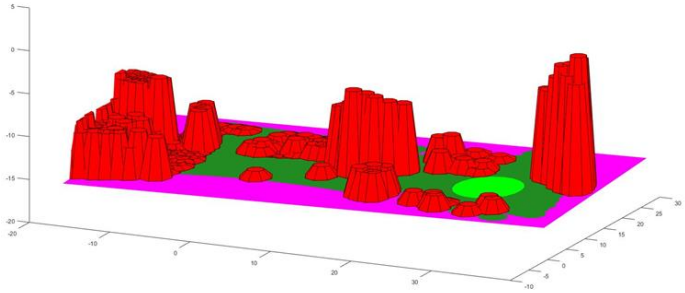
Iz meritve določimo meje območja in ga razdelimo na podobmočja. V posameznem podobmočju iščemo ravnino, ki se najbolj prilega točkam z RANSAC metodo. RANSAC je iterativni postopek za oceno parametrov matematičnega modela iz podatkov, ki vsebujejo osamelce, za katere ne želimo, da vplivajo na oceno parametrov. V algoritmu se najprej preveri, če je dovolj točk, sicer je podobmočje označeno za nedefinirano območje, nato se naključno izbere tri točke, katere eksaktno določajo eno ravnino. V naslednjem koraku se izračunajo razdalje vseh točk podobmočja do ravnine, kjer se shranijo vse točke, ki so v določeni bližini (npr. vsaj 10 cm) ploskve, te smatramo kot dobre točke. Opisani postopek se *n*-krat ponovi. Seznam točk shranimo le, ko je teh več kot v vseh prejšnjih poskusih. Dobre točke se lahko klasificira tudi kot točke katere so bile več kot *m*-krat označene za dovolj blizu naključno najdene ravnine. Po končanih iteracijah je shranjen največji seznam dobrih točk in po metodi najmanjših kvadratov se določi ravnino, katera se najbolj prilega dobrim točkam. Dobljeno ravnino se oceni na podlagi njene vodoravnosti oz. poravnanosti normale ravnine z *z*-osjo ter skupno oddaljenostjo dobrih točk do nje. Na podlagi odstopanja se dobljeno ravnino oceni za spremenljivo ali nespremenljivo, torej dobra ali slaba cona.

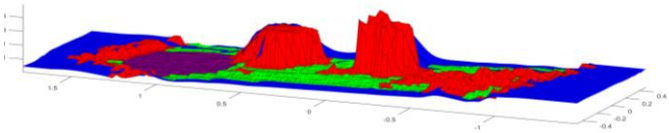
V naslednjem koraku se z upoštevanjem vseh dobrih ravnin podobmočij lahko določi glavno ravnino (na Sliki 5-6 označeno s temno zeleno) na več različnih načinov. Težave povzročajo dobre ravnine na različnih višinah.

Sledi klasificiranje vseh točk na podlagi oddaljenosti od glavne ravnine, kjer prag (npr. 10 cm) določa točko kot dobro ali slabo. Dodatno se preveri okolico vseh slabih točk, kjer se išče prvo dobro točko okoli dane slabe, katero se naknadno interpretira tudi kot slabo točko. S tem se pridobi informacija o robovih ovir. Vse dane točke se preslika na glavno ravnino, kjer se na področjih, kjer so točke preveč oddaljene med seboj, doda nove "nedefinirane" točke, katere v radiju *R* ne vsebujejo nobene druge točke.

Nato poiščemo največji radij dobrih točk za potencialno najboljšo pristajalno površino. Le-ta ne sme zajeti nobene slabe ali "nedefinirane" točke, dodatno mora biti celotno iskano področje znotraj mej glavne ravnine in ne sme biti v kontaktu z nedefiniranim področjem. Iskanje je iterativni postopek, kjer se iskano področje premika po *x*- in *y*-osi in se shrani najbolj uspešno področje.

Končni rezultat razpoznave je grafični prikaz (Slika 5, prva metoda) glavne ravnine z osvetljenim krogom najprimernejše cone pristajanja, malo temnejša zelena barva na ravnini prikazuje dobre točke in rdeča barva prikazuje izbočene ovire. Za drugi način (slika 5) predstavitve rezultatov se najprej v vsakem podobmočju poračuna mediana srednje vrednosti točk, v nedoločenih območjih se ta vrednost določi z mediano sosednjih vrednosti median, ter se vse te točke poveže in oblikuje nepretrgano in nezvezno površino.


Slika 5: Prikaz rezultatov s prvo metodo


Slika 6:Predstavitev rezultatov po drugo metodi

4.4 Pristop s stereo kamero

S stereo kamero, sestavljeno iz dveh kamer ACME VR06, so bili tako kot s senzorjem LIDAR, zajeti zunanji posnetki. Z uporabo kalibriranih stereo kamer se je lahko nato pridobilo ustrezno globinsko sliko. Izkazalo se je, da je kvalitetno izvedena kalibracija izrednega pomena za pridobitev ustrezne globinske slike. V nadaljevanju je opisan postopek pridobivanja globinskih slik.

Prvi korak pri obdelavi posnetkov je kalibracija stereo kamere. Z njo se pridobi parametre, ki opisujejo notranje parametre kamere in zunanje parametre kamere. Notranji parametri predstavljajo optični center in goriščno razdaljo kamere. Zunanji parametri predstavljajo orientacijo in oddaljenost med obema kamerama. Zaradi uporabe kamer, ki imata širok zorni kot, je pomembna uporaba radialnih koeficientov distorzije, ki upoštevajo večje lomljenje svetlobnih žarkov ob robovih leče kot ob optičnem centru. Kalibracijo se lahko izvede s pomočjo slike šahovnice, ki je v zornem kotu obeh kamer. Na vsaki sliki se poišče oglišča šahovnice. Z uporabo več slik šahovnic z označenimi oglišči in informacije o resnični razdalji med posameznimi oglišči se lahko izračuna potrebne notranje in zunanje parametre kot tudi parametre distorzije z algoritmom, opisanim po članku [1].

Z uporabo kalibracijskih parametrov se lahko odpravi distorzijo na slikah in jih izravna. Z izravnavo slik se doseže, da pripadajoče točke iz obeh slik ležijo v isti vrstici, oziroma isti vrednosti *y* koordinate slike. Za dosego tega se uporabi transformacijo slike, ki doseže, da se zdi, kot da sta kameri poravnani paralelno. S pridobitvijo izravnanih slik se lahko na teh izračuna sliko disparitete med obema slikama. Dispariteta predstavlja razliko v slikovnih točkah med stolpcema korespondenčnih točk na levi in desni sliki.

Računanje disparitete je potekalo s pol-globalno metodo prileganja (angl. Semi-global matching) [2], ki se od običajnih metod računanja disparitete razlikuje po tem, da zahteva podobno dispariteto med bližnjimi bloki slikovnih elementov, zaradi česar se lahko delno znebimo šuma, ki nastane zaradi neujemanja slikovnih elementov pri iskanju disparitete.

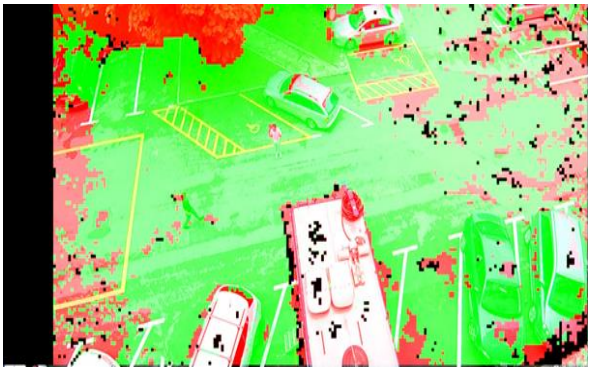
Oblak točk iz najdenih parov slikovnih elementov dobimo preko enačbe:

$$\begin{bmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/T_x & (c_x - c_x')T \end{bmatrix} \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix},$$

kjer c_x ter c_y predstavljata središče senzorja leve kamere, f je goriščna razdalja leve kamere, T je razdalja med kamerama, c_x' je x-koordinata središča desne kamere, d je dispariteta na koordinatah x ter y. Ker imamo opravka s homogenimi koordinatami, je pozicija slikovne točke v 3D koordinatah enaka (X/W, Y/W, Z/W).

Ko je oblak točk dobljen ga obdelujemo na enak način kot z LIDAR podatki (Slika 2 in podpoglavje 4.3). Rezultat obdelave je prikazan na Sliki 7.

Vsaka točka na Sliki 7 predstavlja 10x10 pikslov na originalni sliki. Glede na to ali je dobra ali slaba so točka obarva zeleno ali rdeče. V primeru, ko točka ni določena pa na Sliki 7 pripadajoče slikovne elemente prikažemo s črno barvo.



Slika 7: Prikaz dobrih (zelenih), slabih (rdečih) in nedoločenih (črnih) con na sliki leve kamere

5 Zaključek

V delu je predstavljen predlog zasnove avtonomnega sistema za pristajanje letala in avtomatične ocene varne cone pristanka, ki temelji na 3D LIDARju in stereo-kameri. Opravljen pregled senzorjev je pokazal, da je veliko senzorjev, ki bi bili primerni (npr. solid-state LIDAR) še v fazi razvoja. Iz opravljene študije vidimo predvsem omejitve in prednosti senzorjev, ki so trenutno na trgu.

Rezultati, dobljeni pri vrednotenju delovanja predlaganih algoritmov so dobri, glede na pričakovane težave strojnega vida celo nad pričakovanji. Algoritmi trenutno ne delujejo v realnem času, saj je bil primarni cilj potrditev predlaganih konceptov, algoritmov in nabora senzorjev. Opravljene raziskave nakazujejo, da je mogoče z dostopnimi in cenovno sprejemljivimi senzorji ter obstoječimi algoritmi razviti zanesljiv sistem samostojnega zaznavanja cone pristajanja.

Nadaljnje izboljšave in koraki bodočega razvoja bi bili sledeči. Preizkus metodologij na bolj kvalitetnih senzorjih, predvsem kamerah. Nadgradnje algoritmov v smislu večje robustnosti, izbira zanesljivejših značilk pri določanje disparitet slike in delovanje v realnem času. Izvedba meritev na dejanskem letalu s predlaganimi scenariji senzorjev ter izvedba izboljšav v postavitvi senzorjev, njihovi izbiri ter v algoritmih.

Zahvala

Delo je nastalo v okviru PKP projekta, ki ga sofinancirata Evropska unija iz Evropskega socialnega sklada in Republika Slovenija.

Literatura

[1] Z., Zhengyou. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22 , str.1330-1334, 2000.

[2] H. Hirschmuller. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. *International Conference on Computer Vision and Pattern Recognition*, vol. 2, str. 807-814, 2005.

Zaznavanje napak na sistemu HVAC z metodama PCA in ICA

Anže Stražar, Goran Andonovski

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: strazar@gmail.com, goran.andonovski@fe.uni-lj.si

Fault detection on the HVAC system with PCA and ICA methods

Abstract. *This paper presents a study of fault detection methods on HVAC system. The HVAC system is part of almost every factory and has the function to maintain and provide the desired conditions for comfortable indoor living. The main purpose of the paper is to detect various faults on the HVAC system based on the statistical data. In the paper we briefly introduce the HVAC system and the possible faults that can occure on the system. We tested and compared two statistical fault detection methods, PCA (Principal component analysis) and ICA (Independent component analysis). Moreover we selected the most influentilal components for detection of an individual fault. We also tested different settings of the initial parameters and we choosed the best ones for each fault. Both methods, PCA and ICA, give satisfactory and comparable results.*

1 Uvod

Konkurenčne razmere na trgu zahtevajo od proizvajalcev boljšo kakovost in učinkovitost proizvodnje. Posledično je čedalje pomembnejša tudi zahteva po pravilnem delovanju različnih sistemov. Napake v delovanju lahko povzročijo poslabšanje kakovosti proizvodnje oziroma celo izpad proizvodnje, kar je lahko povezano z visokimi stroški ali je nevarno okolju. Pri povečanih zahtevah za učinkovitost proizvodnje in kvalitete delovanja procesov v industriji, je torej pomembno vseskozi spremljati celoten proces in zaznavati morebitne napake. V članku bomo obravnavali zaznavanje napak na klimatskem sistemu HVAC – sistem gretja, prezračevanja in hlajenja (ang. *Heating, Ventilation and Air Conditioning*), pri čemer bomo napake zaznavali na podlagi podatkov, ki poročajo o stanju procesov znotraj sistema [1].

V klimatskih procesih uspešno delovanje temelji na dobrem delovanju vseh podsistemov. V reguliranih procesih obravnava regulacijski algoritem izpad ali slabše delovanje enega od sestavnih elementov kot motnjo. Nekatere napake v delovanju lahko regulacijski algoritem izniči. S tem se običajno močno spremenijo regulacijski signali, lahko se poslabša kakovost regulacije, a brez vpogleda v notranje stanje regulacijskega algoritma, napaka ni očitna. Tu nastopi problem, saj se napaka ne odstrani, proces pa se trajno nahaja izven optimalnih razmer, za katere je bil skonstruiran. To lahko povzroči delovanje v nasičenju, previsoko obrabo aktuatorjev, preveliko uporabo energije itd [2] [3].

HVAC sistem se uporablja za izmenjavo in pripravo zraka. Je del skoraj vsake tovarne in drugih večjih večnamenskih zgradb, kot so izobraževalne ustanove, bolnišnice, poslovne zgradbe ipd.. Njegova naloga je vzdrževanje in zagotavljanje željenih pogojev udobnega bivanja v zaprtih prostorih. Sistem sestoji iz več strojnih elementov, ki imajo različne funkcionalnosti. Zahtevane pogoje v prostorih dosegamo z različnimi gradniki, kot so ventilator, grelnik, hladilnik in vlažilnik zraka [2] [3].

Eden od glavnih namenov vodenja sistema HVAC so regulacije željenih oziroma zahtevanih parametrov, kot so temperatura, vlaga in tlak. Z navedenimi gradniki lahko vplivamo na te parametre. Elemente posredno vodi algoritem vodenja, katerega namen je priprava zraka z želenimi lastnostmi [2] [5].

Za zaznavanje napak smo uporabili pristopa, ki temeljita na metodi glavnih komponent (ang. *principle componenet analysis - PCA*) in metodi neodvisnih komponent (ang. *independent componenet analysis - ICA*). Na voljo so še drugi pristopi, med drugimi dinamična metoda glavnih komponent (Dynamic PCA, DPCA), ki je v splošnem razširitev prej omenjene metode glavnih komponent [2] [4].

Pri zaznavanju napak smo se omejili na zaznavanje napak na primerih ventilov grelca, hladilnika ter rekuperatorja. S spoznanimi metodami (PCA in ICA) smo ugotavljali napako stalne odprtosti oziroma zaprtosti ventilov na omenjenih treh gradnikih sistema. Za potrebe testiranja uporabimo podatke, ki so bili pridobljeni s pomočjo Simulink modela HVAC sistema.

2 Teoretične osnove

2.1 Predvidene napake na HVAC sistemu

Za pravilno regulacijo sistema je pomembno, da ima operater na voljo pravilne dejanske vrednosti stanj v sistemu. Da zagotovimo pravilen odziv sistema na napake, poskušamo predvidevati vpliv najpogostejših napak in pripraviti sistem na pravilen odziv ob prisotnosti le-teh. Predvidimo več vrst napak na signalih:

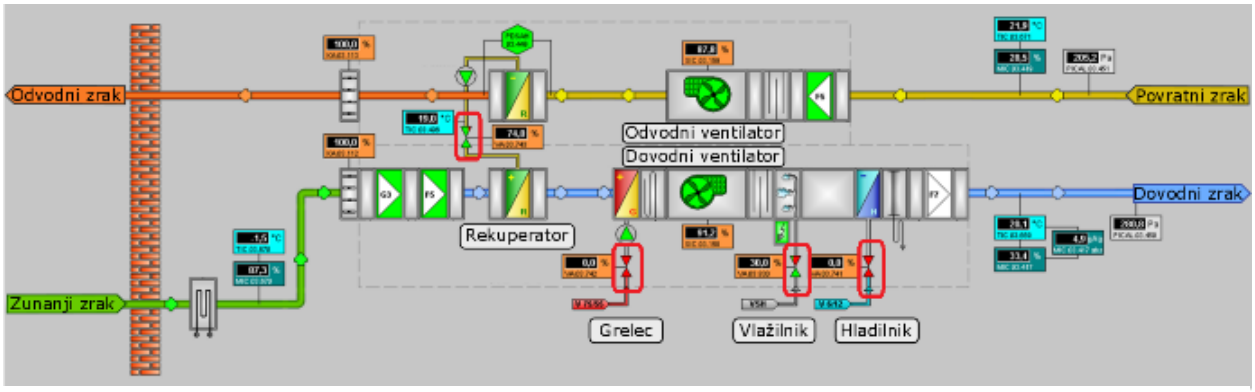
- zakasnjen signal,
- premik signala po amplitudi,
- izpad komunikacije,
- nepričakovan preklop v ročni način,
- odpoved HVAC elementov (grelec, hladilnik, ventilator, vlažilec).

Napake na signalih lahko vplivajo na:

- odprtost/zaprtost ventila rekuperatorja, grelca, hladilnika,
- moč dovodnega ventilatorja,
- moč odvodnega ventilatorja,

- idr. [3] [5].

Na spodnji sliki (slika 1) so poudarjena mesta znotraj strukture HVAC sistema, pri katerih želimo preizkusiti zaznavanje morebitnih napak. Izbrana mesta predstavljajo ventile rekuperatorja, grelca in hladilnika,



Slika 1. Shema HVAC sistema

pri katerih bomo pozorni na napake v obliki odprtosti oziroma zaprtosti ventilov.

2.2 Metoda glavnih komponent

Za gradnjo detektorjev različnih napak smo preizkusili metodo glavnih komponent (ang. principle componenet analysis - PCA). Metoda poišče linearna razmerja med spremenljivkami ter omogoči preslikavo osnovnih podatkov v prostor glavnih komponent, katerega osi so med seboj ortogonalne. Spremenljivke meritev v novem prostoru so med seboj nekorelirane. Metoda je uporabna, ko imamo opravka z visoko-dimenzionalnimi podatki, saj omogoča predstavitev podatkov z manjšim številom spremenljivk [2] [7].

Uporabna lastnost PCA je, da omogoča uporabo manjšega števila komponent za predstavitev zadetkov, kot jih je potrebnih za zapis izvirnih podatkov. Transformacijska matrika, ki podatke preslika iz izvirnega prostora v prostor zadetkov, je v osnovi kvadratna. Lahko se zmanjša in izbere le nekaj glavnih komponent. Nekatere komponente nosijo zelo malo informacije (nizke lastne vrednosti) in se jih zato lahko zanemari. S tem se zmanjša dimenzijo matrike zadetkov. Število izbranih komponent določa parameter CPV. Uporabnik določi *CPV*, ki določa, koliko komponent moramo izbrati za preslikavo v prostor glavnih komponenet – s tem določimo dimenzijo prostora glavnih komponent. Tipična vrednost je *CPV* = 0.95 = 95 % [2] [7] [8].

2.3 Metoda neodvisnih komponent

Metodo neodvisnih komponent lahko obravnavamo kot posplošitev metode PCA, pri čemer nam metoda ICA podaja več informacij o ločljivosti razredov. Uporabna je za podatke, za katere se domneva, da so linearne kombinacije nekaterih neznanih latentnih spremenljivk. Šteje se, da so latentne spremenljivke ne-Gaussove in neodvisne. ICA najde te neodvisne komponente [9] [10].

Metodo neodvisnih komponent smo vpeljali v obliki algoritma FastICA.

2.4 Metoda neodvisnih komponent

Za zaznavanje napak s PCA in ICA smo uporabili dve metriki: Hotelling-T2 in Q-mera. Napaka je zaznana v tistih trenutkih, ko je mera za delovanje z napako nižja kot mera za delovanje brez napake [2] [7].

2.4.1 Hotelling-T²

Mera, poznana tudi kot Mahalanobisova razdalja, je definirana z enačbo:

$$h_i\,(x_i - m)^T S^{-1} (x_i - m) \qquad (1)$$

kjer je i oznaka indeksa vzorca, torej h_i predstavlja mero za trenutni vzorec, x_i predstavlja vrstico v matriki vzorcev (X), m je (vrstični) vektor srednjih vrednosti vzorcev v matriki X. S je kovariančna matrika:

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^T (x_i - m) \qquad (2)$$

N predstavlja število vzorcev (vrstic matrike X) [2] [7].

2.4.2 Q-mera

Predstavlja kvadrat evklidske razdalje med merjenim vzorcem in nazaj v osnovni prostor preslikanim zadetkom [2] [7].

$$q_i = (x_i - t_i P_{SK}^T)^T (x_i - t_i P_{SK}^T) \qquad (3)$$

3 Metodologija dela

3.1 Detektor napak

Za gradnjo detektorja napak smo izbrali pristop pri katerem se v koraku učenja detektorju poda dva seta podatkov: set meritev, ki podaja tipično delovanje brez napak ter set meritev sistema za obdobje, ko je prisotna napaka, ki jo želimo zaznati. Detektor preverja, kateremu načinu delovanja meritve ustrezajo. Za vsak vzorec (vektor meritev) se izračuna mera odstopanja od obeh načinov delovanja (z in brez napake). Vzorec se razvrsti v tisti 'razred' (način delovanja), kjer je mera manjša. Za izgradnjo in testiranje detektorjev napak potrebujemo množico učnih in testnih podatkov, ki so bili za potrebe tega članka že generirani s pomočjo simulacije modeliranega sistema [2] [7].

Za vsak detektor napake smo preiskali širok nabor nastavitev: različne množice simulacijskih signalov oziroma vplivnih komponent, različne vrednosti pragu

CPV ter izbrali različne mere za izračun odstopanja (mera T^2 in Q). Preiskali smo širok nabor nastavitev za vsak detektor napake ter vsako množico nastavitev ovrednotili. Upoštevni sta tako splošna uspešnost, izračunana na podlagi matrike zamenjav, kot ocena robustnosti. Slednja je izračunana iz razdalje med obema merama v različnih trenutkih [2] [7].

4 Rezultati

4.1 Napaka na ventilu rekuperatorja – stalno zaprt

Izbrana je bila naslednja množica simulacijskih signalov oziroma vplivnih komponent:

- odprtost ventila rekuperatorja,
- odprtost ventila grelca,
- temperatura grelca,
- temperatura vpiha,
- vlaga vpiha.

Tabela 1 povzema uporabljene metode in kriterije s katerimi smo v kombinaciji z izbrano množico vplivnih komponent gradili detektorje napak ter jih zatem tudi ovrednotili.

Tabela 1. Uporabljeni kriteriji

Metoda	PCA & ICA
CPV	95%, 80% in 70%
Mera	T^2 in Q-mera

V Tabeli 2 in Tabeli 3 je predstavljeno vrednotenje nastavitev zgrajenih detektorjev napak, za napako zaprtega ventila rekuperatorja. Opazimo, da so v splošnem deleži uspešnosti razpoznavanja napak višji, ko za uporabljene kriterije za gradnjo detektorjev uporabimo mero Q, kot metriko za zaznavanja napak in prag CPV z vrednostjo 95%. V Tabeli 2 in Tabeli 3 so navedeni rezultati za prag CPV ob uporabi vrednosti 95% in vrednosti 80%. Ob uporabi pragu z vrednostjo manjšo od 80% se rezultati niso bistveno razlikovali od že uporabljenega pragu CPV z vrednostjo 80%. Le iz zaznavanja napak na ventilu rekuperatorja je težko določiti ali se bolje obnese metoda PCA ali ICA, saj ne dobimo konstantnih enoličnih ugotovitev ob uporabih različnih kriterijev, ki jih zajema gradnja detektorja napake.

Tabela 2. Prikaz uspešnosti zaznave napak pri izbranih metodah in kriterijih

	Uspešnost [%]			
	Mera: T^2			
	CPV: 95 %		CPV: 80 %	
	PCA	ICA	PCA	ICA
Zaprt ventil rekuperatorja	42,7 %	56,6 %	42,7 %	42,8 %

Tabela 3. Prikaz uspešnosti zaznave napak pri izbranih metodah in kriterijih

	Uspešnost [%]			
	Mera: Q			
	CPV: 95 %		CPV: 80 %	
	PCA	ICA	PCA	ICA
Zaprt ventil rekuperatorja	87,7 %	60,4 %	81,8 %	57,3 %

4.2 Pregled skupnih rezultatov

Rezultatom razpoznavanja napake na ventilu rekuperatorja dodamo še rezultate razpoznavanja napake na ventilih grelca in hladilnika. Kot napako, pri vseh upoštevamo ventil, ki je stalno odprt. Za ventil rekuperatorja upoštevamo množico vplivnih komponent, ki je predstavljena v poglavju 4.1. Za ventil grelca oziroma ventil hladilnika, pa sta bili izbrani naslednji množici simulacijskih signalov oziroma vplivnih komponent:

Stalno odprt ventil grelca:

- odprtost ventila rekuperatorja,
- odprtost ventila grelca,
- temperatura vpiha,
- temperatura grelca,
- vlaga vpiha.

Stalno odprt ventil hladilnika:

- odprtost ventila hladilnika,
- temperatura vpiha,
- temperatura hladilnega medija,
- sprememba specifične vlažnosti.

Tabela 1 povzema uporabljene metode in kriterije s katerimi smo v kombinaciji z izbrano množico vplivnih komponent gradili detektorje napak ter jih zatem tudi ovrednotili.

V Tabeli 4 in Tabeli 5 je predstavljeno vrednotenje nastavitev zgrajenih detektorjev napak za napako stalno odprtega ventila rekuperatorja, grelca in hladilnika. Opazimo, da so v splošnem deleži uspešnosti razpoznavanja napak višji, ko za uporabljene kriterije za gradnjo detektorjev uporabimo mero Q, kot metriko za zaznavanja napak in prag CPV z vrednostjo 95%. V splošnem dobimo boljše rezultate ob uporabi metode PCA.

Tabela 4. Prikaz uspešnosti zaznave napak pri izbranih metodah in kriterijih

	Uspešnost [%]			
	Mera: T2			
	CPV: 95 %		CPV: 80 %	
	PCA	ICA	PCA	ICA
Odprt ventil grelca	66,9 %	58,8%	72,0 %	58,8%
Odprt ventil hladilnika	66,3 %	47,6 %	66,5 %	67,2%
Odprt ventil rekuperatorja	42,7 %	42,8 %	42,7 %	54,2 %

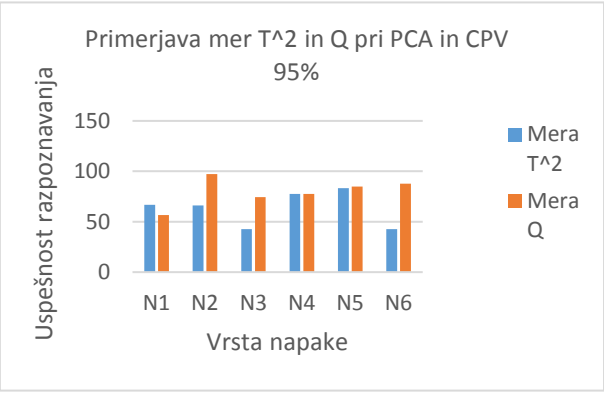
Tabela 5. Prikaz uspešnosti zaznave napak pri izbranih metodah in kriterijih

	Uspešnost [%]			
	Mera: Q			
	CPV: 95 %		CPV: 80 %	
	PCA	ICA	PCA	ICA
Odprt ventil grelca	56,5%	57,3%	44,1 %	42,7%
Odprt ventil hladilnika	97,3%	47,5%	97,0 %	47,5%
Odprt ventil rekuperatorja	74,5 %	47,1 %	73,5 %	47,1 %

V tabeli 6 je predstavljena legenda napak, ki so predstavljene na sliki 2. Na sliki 2 vidimo, kako izbira metrike za zaznavanje napake pri metodi PCA in izbranem pragu CPV je 95%, vpliva na uspešnost razpoznavanja napake. Opazimo, da je uspešnost razpoznavanja večja v primeru, ko uporabimo za metriko razpoznavanja uporabimo mero Q.

Tabela 6. Legenda napak

Napaka	Oznaka
Odprt ventil grelca	N1
Odprt ventil hladilnika	N2
Odprt ventil rekuperatorja	N3
Zaprt ventil grelca	N4
Zaprt ventil hladilnika	N5
Zaprt ventil rekuperatorja	N6



Slika 2. Primerjava mer T^2 in Q pri PCA in CPV 95%

5 Zaključek

V članku je predstavljeno zaznavanje napak na klimatskem sistemu HVAC. Zaznavanje napak deluje na osnovi metode glavnih komponent in metode neodvisnih komponent. Glavna lastnost naučenega detektorja je transformacijska matrika, ki preslika iz prostora meritev v prostor zadetkov. Predlagani sta dve metriki za ocenjevanje vhodnih vzorcev in s tem zaznavanje tistih z napakami.

Pri metodah PCA in ICA, za njuno realizacijo ne potrebujemo modela procesa, temveč le zgodovino merjenih spremenljivk procesa. Metodi odlikuje tudi relativno preprosta realizacija. Zavedati se moramo tudi njunih omejitev. Metodi potrebujeta za uspešno izolacijo napake učno množico podatkov z določeno napako v delovanju. Ti podatki niso vedno dostopni, še posebej če gre za napake v procesu, ki drastično spremenijo njegovo delovanje. V tem primeru si lahko pomagamo le s simulacijo na modelu, ki te napake upošteva.

Predstavljen način zaznavanja napak za večino predvidenih napak deluje z zadovoljivo uspešnostjo. Opazimo, da imata pomembno vlogo pri zaznavanju napake vrednost parametra CPV in izbira metrike za zaznavanje napak (Hotelling, Q-mera). Načrtani algoritem za zaznavanje napak se je najbolje obnesel, ko smo za parameter CPV uporabljali vrednost CPV = 95% in kot metriko uporabili mero Q.

Literatura

[1] G. Klančar, I. Škrjanc: *Metoda glavnih komponent pri odkrivanju in izolaciji napak: primer hidravličnega procesa in procesa fermentacije*, Elektrotehniški vestnik št. 69, str. 311-316, 2002.

[2] Ž. Stržinar: *Modeliranje in zaznavanje napak v klimatskih sistemih*, Fakulteta za elektrotehniko, Ljubljana, 2017.

[3] M. Bastašič: *Vodenje in nadzor večjega sistema HVAC*, Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko, 2016.

[4] V. Štruc: *Biometrični pristopi na osnovi računalniškega vida*, Fakulteta za elektrotehniko, Ljubljana, 2006.

[5] Ž. Stržinar: Prileganje Simscape modelov, Fakulteta za elektrotehniko, Ljubljana, 2017.

[6] A. Øyvind: *HVAC-systems: Modeling, simulation and control of HVAC-systems*, 2011.

[7] I. Škrjanc: Inteligentni sistemi za podporo odločanju, Založba FE in FRI, Univerza v Ljubljani, Fakulteta za elektrotehniko.

[8] K. Košmelj: *Metoda glavnih komponenet: osnove in primer*, Acta agriculturae Slovenica, 89-1, str. 159 – 172, 2007.

[9] Ž. Zaplotnik: Uporaba analize neodvisnih komponent v fizikalnih aplikacijah, Fakulteta za matematiko in fiziko, Ljubljana, 2014.

[10] „ICA“ [Elektronski]. Available: https://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml [Poskus dostopa januar 2018].

Močnostna elektrotehnika

Power Engineering

Primerjava različnih izvedb regulacije v močnostni elektroniki

Mitja Nemec

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

Comparison of various control approaches in power electronics

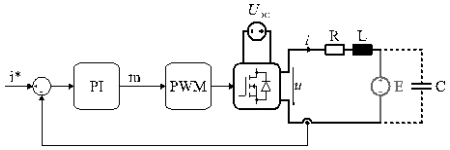
Abstract. *This paper presents different implementations of PID controller that can be found in power electronics applications. In the beginning general application of current and/or voltage control is presented. Then the fundamental difference between serial and parallel architecture of PID controller is described. Furthermore the details regarding the implementation of integral and derivative and the application of feed–forward control is explained. Finally the difference between normalized and absolute output of the controller is described and how this impacts the implementation of supply voltage rejection mechanisms.*

1 Uvod

V močnostni elektrotehniki je regulacija toka in/ali napetosti že dodobra obdelana tema. Pričujoči članek se tako loteva praktičnih vidikov implementacije regulacije in njihovih medsebojnih vplivov, česar v strokovni literaturi ne zasledimo .

Regulacija toka in/ali napetosti je lahko izvedena na več različnih načinov bodisi popolnoma analogno kot histerezna, časovno diskretna ali tudi kot zvezna regulacija [1], [2], vendar pa se te rešitve umikajo digitalnim pristopom [3]. In pri teh je daleč najbolj razširjena regulacija povprečne vrednosti (ang. average mode control) kateri je tudi posvečen ta članek.

Splošno močnostno stopnjo za regulacijo toka ali napetosti predstavlja shema na sliki 1. Z močnostno stopnjo, ki je lahko sestavljena iz kakršne koli topologije (pretvornik navzdol, pretvornik navzgor, poln mostič, …), reguliramo tok skozi induktivnost in posredno napetost na kondenzatorju. Pri tem so lahko upornost R, induktivnost L in napetost E, oziroma kapacitivnost C že del bremena. Lahko pa je induktivnost z upornostjo del močnostne stopnje za omejevanje strmine toka, napetost E pa je napetost bremena. Tako sestavo srečamo pri razsmernikih oziroma aktivnih usmernikih [4].



Slika 1. Shema tipične regulacijske zanke v močnostni elektroniki.

Prenosno funkcijo bremena v primeru tokovne regulacije lahko zapišemo kot

$$F_{BR}(s)=\frac{1}{R+L\cdot s}=\frac{1/R}{1+\tau_{BR}\cdot s}, \qquad (1)$$

kjer je τ_{BR} časovna konstanta bremena. Prenosna funkcija bremena v primeru regulacije napetosti pa je:

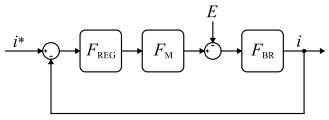
$$F_{BR}(s)=\frac{1}{R+L\cdot s}\cdot\frac{1}{C\cdot s}=\frac{1/R}{1+\tau_{BR}\cdot s}\cdot\frac{1}{C\cdot s} \qquad (2)$$

Prvi približek prenosne funkcije močnostne stopnje in PWM modulatorja je prav tako člen prvega reda [3], kjer je časovna konstanta povezana s preklopno periodo:

$$F_M(s)=\frac{1}{1+\tau_M\cdot s},\tau_M=\frac{1{,}5}{f_{SW}} \qquad (3)$$

Izkaže se, da je ta približek dovolj dober [3], saj je tipično preklopna perioda bistveno krajša kot časovna konstanta bremena in tako nima velikega vpliva na sistem. V primerih, ko pa to ne drži, se lahko poslužujemo Padéjevih približkov višjega reda.

V kolikor pri meritvi toka in/ali napetosti ne nastopa kakšna zakasnitev, ki jo ne moremo zanemariti, imamo tako pri regulaciji toka opravka z zelo enostavno regulacijsko zanko, ki jo sestavljata dva zaporedno vezana člena 1. reda (slika 2).



Slika 2. Blokovna shema regulacijske zanke

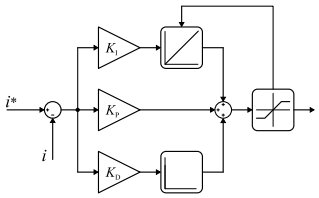
Parametri sistema, ki so uporabljeni pri simulacijah v nadaljevanju članka se nahajajo v tabeli 1.

Tabela 1. Parametri uporabljeni pri simulaciji

Opis	Simbol	Vrednost
Induktivnost	L	1 mH
Upornost	R	0,1 Ω
Inducirana napetost oz. napetost bremena	E	2 V
Napajalna napetost	U_{DC}	24 V
Frekvenca vzorčenja	f_{SW}	10 kHz

2 Struktura PID regulatorja

Osnovna struktura PID regulatorja, ki jo lahko zasledimo v literaturi, je tako imenovana paralelna struktura, ki je prikazana na sliki 3. Pri tem so proporcionalni, integralni in diferencialni del medsebojno popolnoma neodvisni, kar olajša teoretično obravnavo in to je tudi razlog, da je ta arhitektura tipično uporabljena v akademskih tekstih.



Slika 3. Paralelna arhitektura PID regulatorja

Vsak realen regulator, ki vsebuje integralni člen, ima izvedeno tudi preprečitev integralnega pobega (ang. anti-windup), s čimer se prepreči nepotrebna akumulacija napake v integralnem členu med prehodnimi pojavi, ko je izhod regulatorja v nasičenju. Nasičenje je lahko implementirano v samem regulatorju, ali pa v katerem izmed blokov ki mu sledijo. Bistveno je, da se zazna, kdaj je zanka v nasičenju in se nato tudi ustrezno ukrepa. Prenosna funkcija paralelnega PI regulatorja je v osnovi

$$F_{PI}(s) = K_P + \frac{1}{T_I \cdot s} + T_d \cdot s\,, \qquad (4)$$

Kjer sta K_P ojačanje proporcionalnega člena, T_I pa časovna konstanta integratorja. Pogosto pa integralni del raje zapišemo z ojačanjem namesto časovne konstante, saj je bolj razumljivo, ko želimo integralni člen izklopiti (ojačanje K_I=0, časovna konstanta T_I=∞)

$$F_{PI}(s) = K_P + \frac{K_I}{s} + K_d \cdot s \qquad (5)$$

Na tem mestu naj še omenimo alternativni zapis s skupnim imenovalcem, kateri je bolj nazoren pri nastavljanju parametrov regulatorja toka, ko uporabljamo samo PI regulator [3], [5], [6].

$$F_{PI}(s) = K_P \frac{1 + T_{IP} \cdot s}{T_{IP} \cdot s} + K_d \cdot s\,, \qquad (6)$$

kjer velja $T_{IP} = T_I \cdot K_P$.

V kolikor pogledamo prenosno funkcijo za tokovno regulacijsko zanko (7) pridemo do načina nastavljanja parametrov regulatorja, ki je najbolj pogosto v uporabi [7].

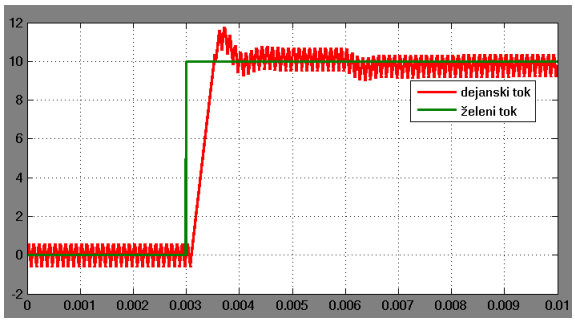
$$H_S(s) = K_P \frac{1 + T_{IP} \cdot s}{T_{IP} \cdot s} \cdot \frac{1}{1 + \tau_M \cdot s} \cdot \frac{1/R}{1 + \tau_{BR} \cdot s} \qquad (7)$$

Ker velja $\tau_{BR} >> \tau_M$, je najbolj smiselno, da izberemo časovno konstanto T_{IP} tako, da izničimo pol z najdaljšo časovno konstanto.

$$T_{IP} = T_{BR} \qquad (8)$$

Ojačanje proporcionalnega regulatorja pa se potem nastavi z ozirom na fazno oz. amplitudno rezervo.

Ob tako nastavljenem regulatorju dobimo ustrezen odziv ob skočni spremembi želene vrednosti (slika 5) odziv ob spremembi napetosti bremena (ob času 0,006 s) pa ni optimalen.



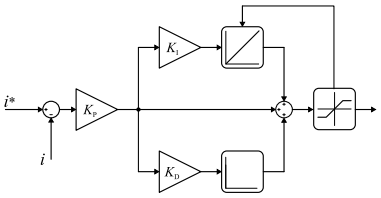
Slika 4. Odziv ob spremembi želene vrednosti toka in ob spremembi napetosti bremena

3 Praktične izvedbe regulacije

Čeprav je paralelna arhitektura PID regulatorja enostavnejša za razlago in posledično tudi za razumevanje vloge posameznega regulacijskega člena, pa je eksperimentalno parametriranje najbolj težavno ravno pri paralelni arhitekturi. V praksi se namreč regulatorje raje nastavlja eksperimentalno (ali je to zaželeno ali ne, presega okvire tega članka).

3.1 Serijska arhitektura PID regulatorja

V praksi tako večkrat zasledimo uporabo serijske arhitekture [5], [6], pri kateri je delovanje integralnega dela odvisno tudi od proporcionalnega dela (slika 5).



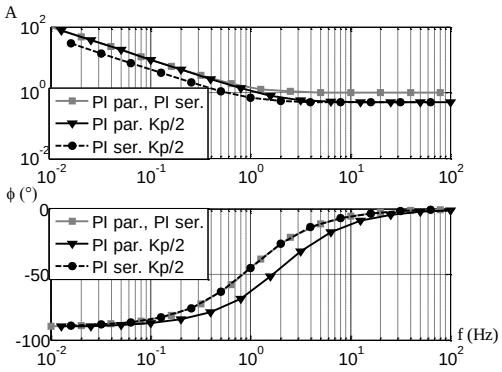
Slika 5. Serijska arhitektura PID regulatorja

Čeprav je delovanje obeh regulatorjev enako je pri nastavitvah treba upoštevati relacijo:

$$K_{I_serijski} = \frac{K_{I_paralelni}}{K_P} \qquad (9)$$

Prednost serijske arhitekture je prikazana na sliki 6. V izhodišču imamo tako serijski kot paralelni PI regulator z enako frekvenčno karakteristiko. Ko pri paralelni arhitekturi znižamo proporcionalno ojačanje se frekvenčna karakteristika spremeni tako amplitudno, kot tudi fazno. Pri takem obnašanju lahko v določenih primerih naletimo na neintuitivno obnašanje, saj lahko sistem postane nestabilen tako v primeru povečevanja ojačanja (pričakovano) kot tudi v primeru zmanjševanja ojačanja (nepričakovano).

Pri serijski arhitekturi, pa se ob spremembi proporcionalnega ojačanja, spremeni samo amplitudna karakteristiki, fazna karakteristika pa ostane enaka. Zaradi tega je serijska arhitektura regulatorja bolj primerna za eksperimentalno nastavljanje parametrov.



Slika 6: Frekvenčni odziv serijske in paralelne arhitekture PI regulatorja ob spremembi proporcionalnega ojačanja.

3.2 Izvedba integralnega člena

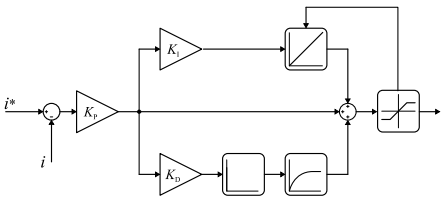
Ne glede na uporabljeno numerično ločljivost je pri izvedbi integralnega regulatorja potrebna pazljivost. Integralni člen se v časovno diskretni obliki zapiše kot

$$y(n) = y(n-1) + \frac{\Delta t}{T_i} x(n). \qquad (10)$$

V primeru, ko je perioda vzorčenja Δt zelo kratka je lahko drugi člen (ne glede na velikost časovna konstante T_i oziroma vhoda $x(n)$) bistveno manjši kot akumuliran izhod ($y(n)$ oziroma $y(n\text{-}1)$). V tem primeru se zaradi numerične napake vhodna vrednost preneha akumulirati. Tipična rešitev tega problema je, da se za akumulirano vrednost izbere zapis z dvojno ločljivostjo.

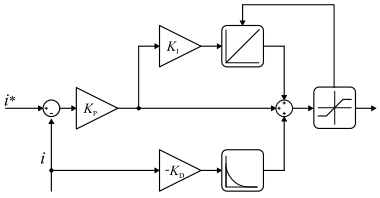
3.3 Izvedbe diferencialnega člena

V realnih primerih, je v signalu napake vedno prisoten tudi šum, katerega visoke frekvence ojača diferencialni člen. Tipična rešitev je, da se v serijo doda nizkopasovno sito 1. ali 2. reda (slika 7). To sicer delno ali v celoti izniči težave s šumom, vendar pa se na ta račun poslabša stabilnost regulacijske zanke in tudi parametriranje regulatorja je težje.



Slika 7: Diferencialni člen s filtrom

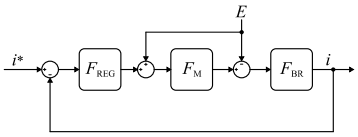
Ker se diferencialni člen uporablja za izboljšanje stabilnosti zanke, ki je posledica prenosne funkcije sistema, pogosto srečamo rešitev kjer diferencialni člen deluje samo na povratni veji (slika 8). Tako se ohrani prispevek regulatorja k stabilnosti sistema, vpliv hitrih sprememb želene vrednosti pa se ne prenese preko regulatorja.



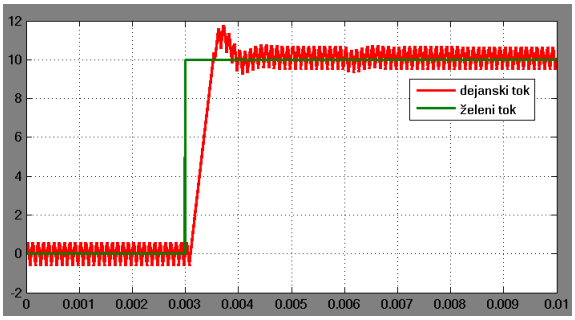
Slika 8: Diferencialni člen samo na povratni veji

3.4 Vpeljava kompenzacijske veje

Da dosežemo hitrejši odziv ob spremembi napetosti bremena je najbolj pogosto uporabljena rešitev vpeljava kompenzacijske veje (slika 9). Tu moramo napetost bremena meriti ali pa vsaj zanesljivo in hitro oceniti [4]. V kolikor imamo ta podatek znan, ga lahko direktno prištejemo k izhodu regulatorja in s tem bistveno pospešimo odziv (slika 10).

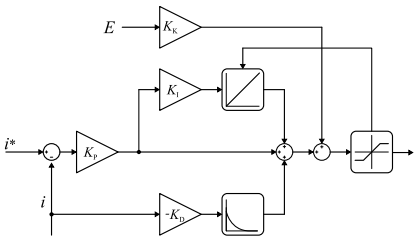


Slika 9. Vpeljava regulacijske krmilne veje



Slika 10. Odziv ob vpeljavi direktne regulacijske veje

Zelo pogosto je kompenzacijske veja izvedena v samem regulatorju. To sicer zahteva dodatni vhod v regulator, vendar pa je v tem primeru tudi kompenzacijska veja vključena v preprečitev integralskega pobega zaradi nasičenja izhoda regulatorja (slika 11).

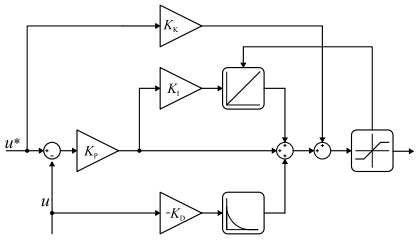


Slika 11. Vpeljava direktne regulacijske veje v PI regulator

Pri vpeljavi kompenzacijske veje pa je treba biti pozoren na dve stvari: šum in zakasnitev kompenzacijskega signala. Šum, ki nastopa pri meritvi signala se preslika na izhod regulatorja, kar lahko bistveno vliva na delovanje regulacije. Sicer lahko šum filtriramo, vendar pa se s tem vnese zakasnitev, kar lahko naredi cel sistem nestabilen.

3.5 Vpeljava direktne krmilne veje

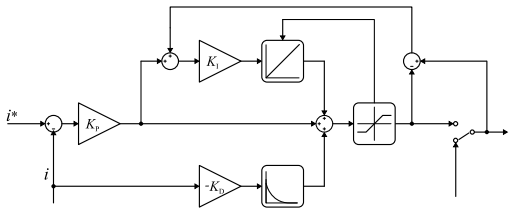
S paralelno direktno krmilno vejo se združi tako krmiljenje kot regulacijo. Pogosto je uporabljena pri regulaciji izhodne napetosti, ko je izhodna napetost odvisna od preklopnega razmerja močnostne stopnje, regulator pa samo popravlja krmilno veličino zaradi ostalih vplivov (breme, napajalna napetost, …) (slika 12). Tudi v tem primeru je zaželeno, da direktna krmilna veja vstopa v regulator z namenom izboljšanja preprečitve integralnega pobega.



Slika 12. Direktna krmilna veja

3.6 Zvezni preklop

V primerih, ko želimo preklapljati med ročnim nastavljanjem krmilne veličine in regulacijo, je za doseganje zveznega preklopa potreben dodaten poseg v regulator (slika 13). Na ta način dosežemo, da izhod regulatorja vedno sledi nastavljeni krmilni veličini, in ob preklopu na regulacijo ni nepotrebnega prehodnega pojava.

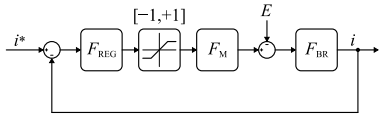


Slika 13. Zvezni preklop

3.7 Normirani izhod

V praksi zelo pogosto srečamo izvedbo regulatorja pri katerem so tako vhodne kot izhodne veličine normirane na območje [-1, +1] (slika 14). Taka izvedba je še posebej primerna za izvedbo s fiksno vejico z nizkim številom bitov zaradi kontroliranega numeričnega razpona. Dodatna prednost te izvedbe je bolj enostavna ponovna uporaba še posebej v aplikacijah, kjer morajo biti vse komponente preverjene in potrjene.

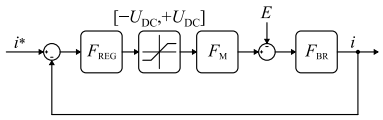
Normirana izvedba regulatorja pa zahteva dodatno pripravo, saj je treba ojačanja regulatorja, ki jih dobimo na podlagi teorije ustrezno prilagoditi.



Slika 14. Normirani izhod regulatorja

3.8 Absolutni izhod

V primeru, ko za izvedbo regulacijskega algoritma uporabljamo plavajočo vejico, ni več potrebe po normirani izvedbi, pa je izhod regulatorja omejen z omejitvami ki so prisotne v sistemu (slika 15). V primeru napetostno napajanih pretvornikov je to napajalna napetost.

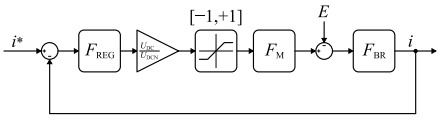


Slika 15. Absolutni izhod regulatorja

3.9 Upoštevanje sprememb napajalne napetosti

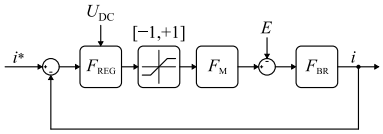
Pri aplikacijah, kjer se napajalna napetost spreminja (baterijsko napajanje, enofazno napajanje s slabšim glajenjem, …), to v regulacijski zanki povzroča spremembo ojačanja sistema. V primeru absolutnega izhoda iz regulatorja to ne predstavlja težave, saj močnostna stopnja poskrbi, da je na izhodu želena napetost ne glede na napajalno napetost.

Pri normiranem izhodu regulatorja pa to nekoliko zakomplicira izvedbo. Ena izmed rešitev je, da se v zanko doda spremenljivo ojačanje (slika 16), ki je odvisno od napajalne in napetosti s čimer dosežemo, da je ojačanje sistema konstantno.



Slika 16. Spreminjanje ojačanja

Pri alternativni rešitvi, ki gre pod tujim imenom »gain scheduling«, pa se spreminja ojačanja regulatorja na podlagi napajalne napetosti (slika 17). Ta rešitev je relativno preprosta v primeru serijske arhitekture regulatorja, pri paralelni izvedbi pa je nekoliko bolj zakomplicirana.



Slika 17. Spreminjanje ojačanja

4 Zaključek

Čeprav je bil klasičen PID regulator teoretično obravnavan že iz vseh zornih kotov vidimo, da obstaja več načinov praktične realizacije. Pričujoči članek opiše osnovne lastnosti posamezne izvedbe končna izbira pa je seveda odvisna od specifične aplikacije.

Literatura

[1] S. Buso, L. Malesani, in P. Mattavelli, "Comparison of current control techniques for active filter applications," *IEEE Trans. Ind. Electron.*, vol. 45, no. 5, pp. 722–729, 1998.

[2] J. Holtz, "Pulsewidth modulation-a survey," *IEEE Trans. Ind. Electron.*, vol. 39, no. 5, pp. 410–420, 1992.

[3] S. Buso in P. Mattavelli, "Digital Control in Power Electronics," *Synth. Lect. Power Electron.*, vol. 1, no. 1, pp. 1–158, Jan. 2006.

[4] M. Nemec, "Uporaba alfa-beta filtra v močnostni elektroniki," in *ERK 2016*, 2016, p. str. 253-256.

[5] A. O'Dwyer, *Handbook of PI and PID controller tuning rules*. World Scientific, 2009.

[6] K. J. Åström in T. Hägglund, *PID controllers: theory, design, and tuning*, vol. 2. ISA Research Triangle Park, NC, 1995.

[7] V. Ambrožič in P. Zajec, *Električni servo pogoni*. V Ljubljani: Slovensko združenje elektroenergetikov CIGRÉ-CIRED, 2016.

Tok enosmernega tokokroga pri večfaznih pogonih

Mitja Nemec

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

Instructions and Template for ERK Conference (ERK_Title)

Abstract. *This paper presents the impact of parallel multiphase drive topologies on DC link current. Utilizing correct phase shift between parallel power stages the AC components of the DC link current can be significantly reduced, but as the simulation results show the attenuation depends on the drive operating point and is subject to law of diminishing returns with more parallel stages. Furthermore parallel operation changes the frequency spectra of DC link current moving the main frequency components higher but modulation algorithm used also has significant impact on the frequency spectra.*

1 Uvod

Področje večfaznih pogonov je v zadnjem desetletju doživelo velik preboj predvsem na področju raziskav in le vprašanje časa je, kdaj se bodo pričeli v večji meri tudi pojavljati v praksi.

V raziskavah je bilo največ pozornosti namenjene osnovnemu principu delovanja [1], [2], redundanci, ki jo tak pogon lahko omogoča [3] in zmanjševanju motenj predvsem na strani stroja (tako zmanjševanju harmonske vsebine, kot tudi sofazne komponente) [4], [5].

Vendar pa lahko z večfaznimi pogoni tudi zmanjšamo vplive, ki jih pretvornik povzroča na strani enosmernega tokokroga. Čeprav so bile v preteklosti že opravljene raziskave, kako je tok enosmernega tokokroga odvisen od obratovalne točke [6] in modulacijskega algoritma [7], raziskav na temo vplivov večfaznih pogonov na tok enosmernega tokokroga ni bilo veliko [8].

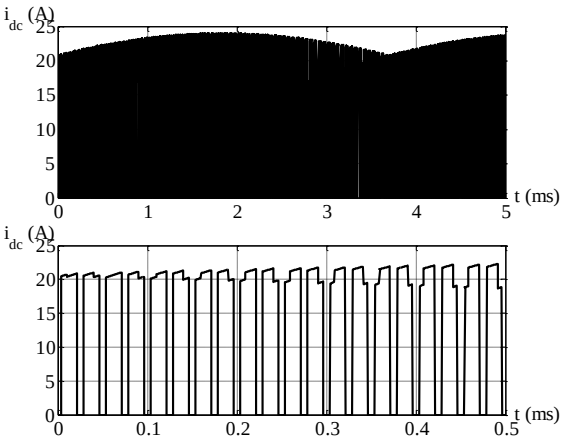
V tem članku si bomo pogledali kakšen tok lahko pričakujemo na enosmerni strani in kako z večfaznim pogonom zmanjšamo izmenično komponento toka v enosmernem tokokrogu, ter kak vpliv imajo različni modulacijski algoritmi.

2 Opis večfaznega pogona

Pri klasičnem trifaznem pogonu z B6 močnostnim pretvornikom, se močnostni pretvornik obnaša kot pretvornik navzdol. To pa pomeni, da je tok v močnostni pretvornik trgan, kar lahko vidimo na sliki 1. Na obliko toka imajo zelo malo vpliva parametri stroja (upornost, induktivnost) več vpliva pa ima delovna točka pogona (hitrost, navor). Taka oblika toka povzroča veliko obremenitev kondenzatorjev enosmernega tokokroga in hkrati tudi povzroča motnje, ki se širijo predvsem preko napajalnega voda.

Obremenitev enosmernega tokokroga z izmenično komponento toka lahko zmanjšamo v primeru, da imamo več paralelnih močnostnih stopenj od katerih vsaka prevzame del moči (toka). V primeru, ko je proženje močnostnih stopenj fazno zamaknjeno, se fazno zamakne tudi enosmerni tok posamezne močnostne stopnje in se lahko izmenična komponenta toka posameznih močnostni stopenj med seboj odšteje.

Omenjeno topologijo lahko dosežemo z večfaznim pogonom [9].

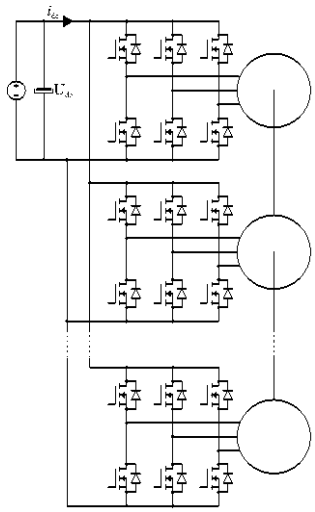


Slika 1: Oblika toka v enosmernem tokokrogu klasičnega B6 močnostnega pretovornika

Vendar pa so s stališča zmanjševanja vplivov na enosmernem tokokrogu paralelnega obratovanja s faznim zamikanjem nosilnega PWM signala v praksi zanimive samo izvedbe brez sklopitve med posameznimi fazami [10]. V strokovni literaturi jih srečamo pod izrazom simetrični večfazni stroji oziroma več-sistemski pogoni. V takem primeru lahko smatramo, da gre za več popolnoma ločenih pogonov na isti gredi (slika 2) in s takim pogonom se bomo ukvarjali v nadaljevanju.

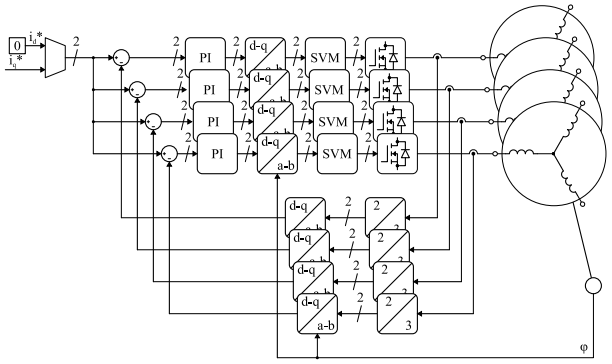
3 Simulacijski rezultati

Za potrebe prikaza vpliva več sistemskega pogona na izmenično komponento toka v enosmernem tokokrogu smo opravili simulacije s pomočjo programskega paketa MATLAB/Simulink. Simulirali smo klasičen trifazni pogon in dvo, tri in štiri-sistemski pogon z ustrezno prilagojenimi električnimi parametri in v enakih delovnih točkah (slika 2). V kolikor ni izrecno navedeno je v simulacijah uporabljena klasična (simetrična oz zvezna) modulacija prostorskega vektorja napetosti (CSVM) [11]. Prisotnost vseh izmeničnih komponent je bila ocenjena s standardno deviacijo.



Slika 2: Shema več-sistemskega pogona

Sama regulacija toka je bila izvedena ločeno za vsak sistem posebej, kot je to za štiri-sistemski pogon prikazano na sliki 3.



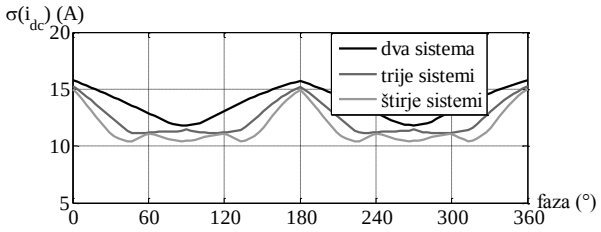
Slika 3: Regulacija toka pri štiri-sistemskem pogonu

Podatki pogona pa so navedeni v tabeli 1:

Tabela 1. Podatki pogona

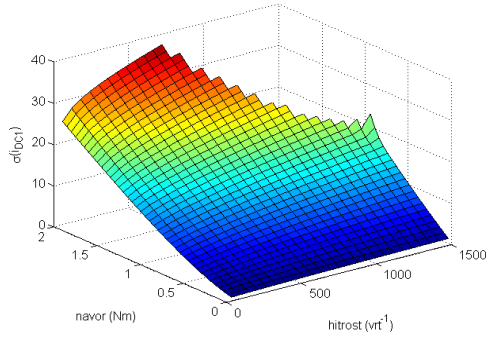
Statorska upornost	0,0567 Ω
Statorska induktivnost	68 μH
Nazivni navor	1.5 Nm
Nazivna hitrost	900 min^{-1}
Napetost enosmernega tokokroga	10 V
Preklopna frekvenca	20 kHz

Prvi set rezultatov (slika 4) prikazuje odvisnost izmenične komponente toka enosmernega tokokroga od faznega zamika med modulacijskimi signali posameznega sistema v eni delovni točki (1 Nm, 900 min^{-1}). Zaradi uporabe simetričnega prožilnega signala je najprimernejša fazna zakasnitev med posameznimi sistemi 90° za dvo-sistemski pogon, 60° za tri-sistemski pogon in 45° za štiri-sistemski pogon. V primeru uporabe nesimetrične modulacije prostorskega vektorja napetosti , so te premaknitve 180°, 120° in 90°.

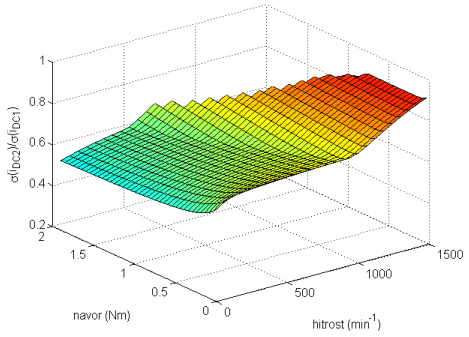


Slika 4: Odvisnost izmenične komponente toka v enosmernem tokokrogu od faznega zamika modulacijski signalov za dvo-sistemski , tri-sistemski in štiri-sistemski pogon.

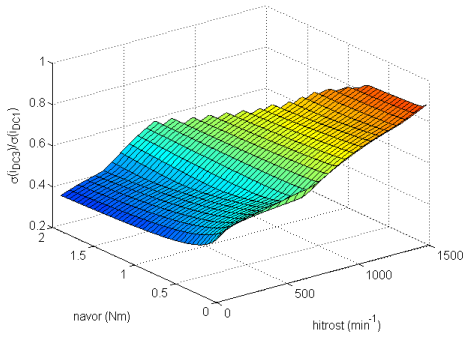
Nadaljnje si lahko ogledamo kako se vsebnost izmenične komponente toka v enosmernem tokokrogu spreminja z obratovalno točko pogona. Na sliki 5 vidimo kako se le ta spreminja za klasičen eno-sistemski pogon. Vidimo da je obremenitev enosmernega tokokroga največja pri najvišjih navorih in je bolj ali manj neodvisna od vrtilne hitrosti. Slike 6-8. prikazujejo relativno zmanjšanje obremenitve za dvo, tro in štiri-sistemski pogon v primerjavi z eno-sistemskim pogonom. Vidimo lahko, da že dvo-sistemski pogon zmanjša obremenitev enosmernega tokokroga v večjem obsegu pod 70%, nikoli pa le-ta ni višja kot 90%. Pri tri-sistemskem pogonu je obremenitev tako vedno manjša ko 85% in v večjem delu pod 60%. Pri štiri-sistemskem pogonu, pa je obremenitev večjem delu pod 60% in nikoli nad 85%. Poleg tega lahko vidimo, da je zmanjšanje izmenične komponente najbolj učinkovito pri nizkih hitrostih in visokih navorih, potem pa ta učinek pada predvsem z naraščajočo hitrostjo. Tako z več-sistemskim pogonom relativno učinkovito zmanjšamo izmenično komponento prav tam kjer je le ta največja t.j. pri visokih navorih. V splošnem lahko vidimo, da največ pripomore razširitev iz eno-sistemskega na dvo-sistemski pogon, potem pa doprinos z vsakim dodanim sistemom pada.



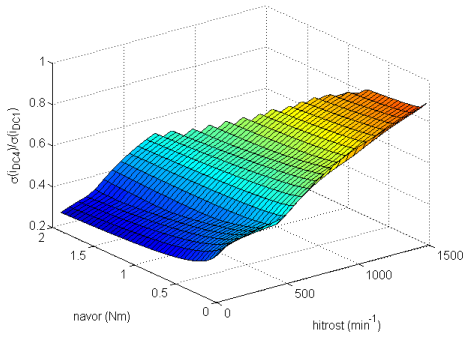
Slika 5: Vsebnost izmenične komponente toka v enosmernem tokokrogu v odvisnosti od delovne točke za eno-sistemski pogon



Slika 6: Relativna odvisnost izmenične komponente za dvo-sistemski pogon.



Slika 7: Relativna odvisnost izmenične komponente za tri-sistemski pogon.



Slika 8: Relativna odvisnost izmenične komponente za štiri-sistemski pogon.

Vendar pa samo vsebnost izmeničnih komponent ocenjena s standardno deviacijo ne pove katere frekvenčne komponente so prisotne. Le te so za eno delovno točko (1 Nm, 900 min^{-1}) prikazane na sliki 9. V primeru uporabe simetričnega proženja, je najmočnejša frekvenčna komponenta (20 kHz) pri eno-sistemskem pogonu dvakratnik preklopne frekvence (40 kHz). Po pričakovanjih se obremenitev enosmernega tokokroga znižuje z večjim številom paralelnih sistemov, prav tako pa se viša tudi najbolj izrazita frekvenčna komponenta. Tako je morebitno odpravljanje teh komponent vedno lažje.

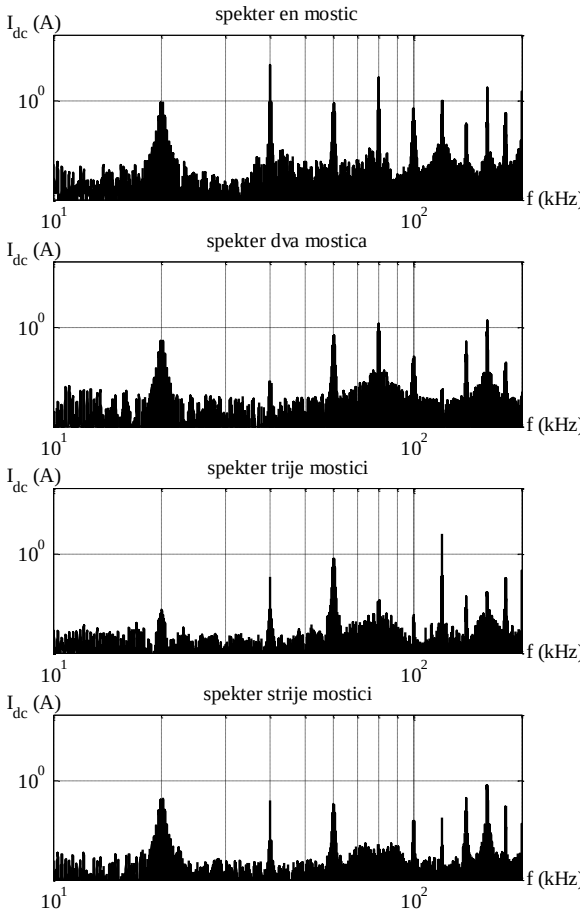
Za primerjavo so priloženi tudi frekvenčni spektri za primer, ko uporabimo asimertrično modulacijo [12] prostorskega vektorja napetosti (slika 10). Kljub temu, da je efektivno gledano obremenitev enaka ne glede na izbrani modulacijski algoritem, pa so frekvence komponente z najvišjo amplitudo za polovico nižje kot pri uporabi simetrične modulacije, kar nekoliko oteži morebitno naknadno odpravljanje teh komponent.

4 Zaključek

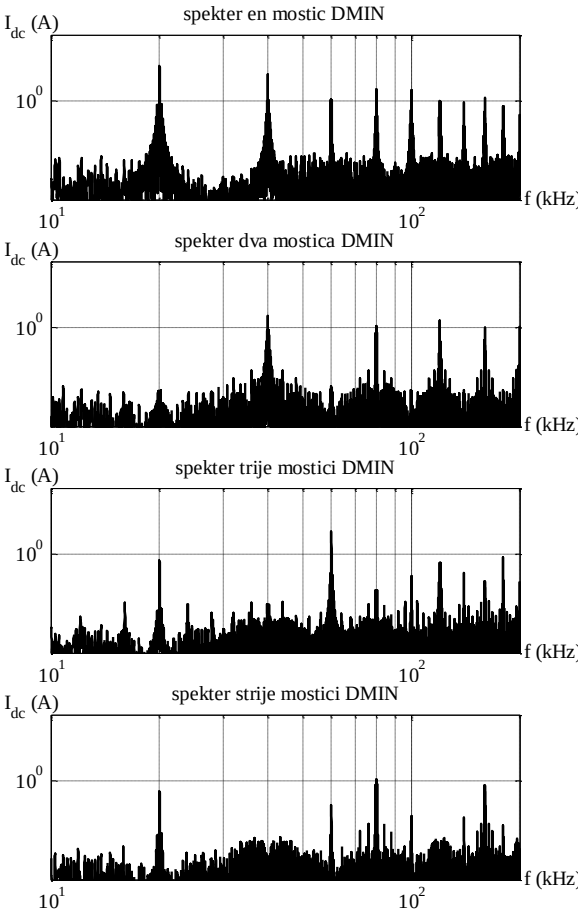
Čeprav je zaradi največjega doprinosa s stališča enosmernega tokokroga najbolj smiselna dvo-sistemska rešitev so več-sistemske rešitve vredne razmisleka v primeru, ko se povečuje moč pogona ali ko je treba prisotnost višjih harmonskih komponent reševati z dodatnimi prijemi. Več-sistemske rešitve te harmonske komponente pomaknejo višje, tako da jih je lažje odpravljati s pasivnimi komponentami.

Prav tako vidimo, da ima izbira modulacijskega algoritma nezanemarljiv vpliv na frekvenčni spekter izmeničnih komponent v enosmernem tokokrogu.

Pričujoči prispevek obravnava obremenitev enosmernega tokokroga z izmeničnimi komponentami, ko lahko posamezne elemente med seboj ločimo in so povezave med njimi brez parazitnih parametrov (induktivnost, kapacitivnost). V praksi, pa je to nemogoče, saj kondenzator enosmernega tokokroga ni en element temveč jih je več, pri čemer se stremi, da se jih postavi čim bližje posamezni tranzistorski veji. V takem primeru pa bi za oceno obremenitve posameznega elementa morali simulirati celoten sklop, pri čemer bi se parazitne parametre najlažje zajelo z FEM izračunom v 3D prostoru, kar pa simulacije naredi bistveno bolj kompleksne in specifične za točno določen pogon.



Slika 9: Frekvenčni spekter toka enosmernega tokokroga za eno, dvo, tri in štiri-sistemski pogon ob uporabi simetrične modulacije



Slika 10: Frekvenčni spekter toka enosmernega tokokroga za eno, dvo, tri in štiri-sistemski pogon ob uporabi asimetrične modulacije

Literatura

[1] E. Levi, „Multiphase Electric Machines for Variable-Speed Applications“, *IEEE Trans. Ind. Electron.*, let. 55, št. 5, str. 1893–1909, maj 2008.

[2] E. Levi, R. Bojoi, F. Profumo, H. A. Toliyat, in S. Williamson, „Multiphase induction motor drives – a technology status review“, *IET Electr. Power Appl.*, let. 1, št. 4, str. 489, 2007.

[3] M. J. Duran in F. Barrero, „Recent Advances in the Design, Modeling, and Control of Multiphase Machines - Part II“, *IEEE Trans. Ind. Electron.*, let. 63, št. 1, str. 459–468, jan. 2016.

[4] K. Gopakumar in V. T. Ranganathan, „Split-Phase Induction Motor Operation from PWM Voltage Source Inverter“, str. 6.

[5] L. Parsa, „On advantages of multi-phase machines“, 2005, str. 6 pp.

[6] J. W. Kolar in S. D. Round, „Analytical calculation of the RMS current stress on the DC-link capacitor of voltage-PWM converter systems“, *IEE Proc. - Electr. Power Appl.*, let. 153, št. 4, str. 535–543, jul. 2006.

[7] M. Bierhoff in F. W. Fuchs, „DC link harmonics of three phase voltage source converters influenced by the pulse width modulation strategy-an analysis“, v *31st Annual Conference of IEEE Industrial Electronics Society, 2005. IECON 2005.*, 2005, str. 6 pp.-.

[8] P. A. Dahono, A. Satria, in D. Nurafiat, „Analysis of DC current ripple in six-legs twelve-devices inverters“, v *2012 International Conference on Power Engineering and Renewable Energy (ICPERE)*, 2012, str. 1–6.

[9] E. Levi, F. Barrero, in M. J. Duran, „Multiphase machines and drives - Revisited“, *IEEE Trans. Ind. Electron.*, let. 63, št. 1, str. 429–432, jan. 2016.

[10] A. Leban in D. Vončina, „Vpliv sklopljenih dušilk na lastnosti vzporedno vezanih pretvorniških modulov“, predstavljeno na ERK, Portorož, Slovenija, 2006, str. str. 459-462.

[11] X. Wen in X. Yin, „The SVPWM fast algorithm for three-phase inverters“, v *2007 International Power Engineering Conference (IPEC 2007)*, 2007, str. 1043–1047.

[12] M. Thanuja, M. B. Jyothi, in D. M. V. Rao, „Analysis And Simulation Of Bus-Clamping Pwm Techniques Based On Space Vector Approach“, *Int. J. Eng. Sci.*, str. 10, 2012.

Vzroki napak pri digitalni regulaciji toka več–vejnega DC–DC pretvornika s PWM prepletom

Peter Zajec

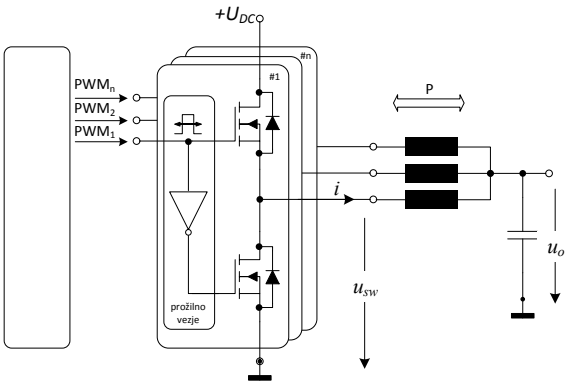
Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-mail: peter.zajec@fe.uni-lj.si

Causes of errors in digital control of multi–phase interleaved DC–DC converter

The paper focuses on some common and usually underestimated sources of current measurement errors deteriorating the control of switched mode DC–DC converters. In particular, the effects of measuring resistor's parasitic inductance and the Point-of-Common-Coupling's inductance are addressed. The second inductance is as a rule overlooked in many design guidelines covering the interleaved DC-DC converters. In addition, the impact of the time delay of Sample/Hold signal and, the mitigation solutions including the hardware as well as the software ones are presented and discussed.

1 Več–vejni pretvornik s PWM prepletom

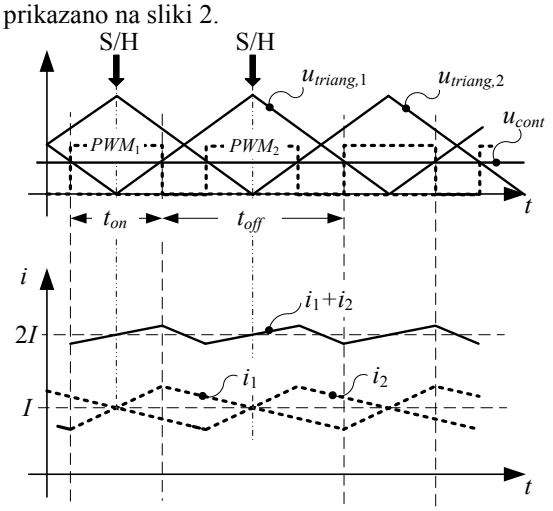
Enosmerni presmerniki (DC-DC pretvorniki) večjih izhodnih tokov so pogosto sestavljeni iz večjega števila (n) vzporedno delujočih pretvornikov (slika 1). Zaradi zmanjšanja tokovne obremenitve vgrajenih komponent je zaželeno, da vsak od njih zagotavlja bremenu enak delež celokupne moči. Močnostna stikala preklapljajo po načelu pulzno širinske modulacije (PWM) z enotno frekvenco (f_{sw}). PWM signali med posameznimi stikali so fazno zamaknjeni (360°/n) s ciljem, da v izhodnem kondenzatorju zmanjšamo amplitudo [1] in povišamo frekvenco ($n{\cdot}f_{sw}$) valovitosti toka (slika 2). Posledično z vgradnjo kondenzatorja bistveno manjše kapacitivnosti zagotovimo enako valovitost izhodne napetosti. Zaradi prepleta toka se izboljšajo tudi tokovne razmere v vhodnem kondenzatorju, s čimer se volumen obeh kondenzatorskih sklopov in pretvornika v celoti zmanjša.



Slika 1. Konceptualna shema več–vejnega pretvornika s PWM prepletom in dvosmernim pretokom moči.

2 Regulacija izhodne napetosti

Izhodno napetost pretvornika se regulira praviloma s kaskadno zasnovo, kjer podrejena tokovna regulacijska zanka primarno določa dinamiko regulirane napetosti [2], [3]. Navkljub ciljani simetriji vgrajenih komponent, tolerančno odstopanje slednjih onemogoča uporabo skupnega tokovnega regulatorja – ki bi generiral enoten krmilni signal (u_{cont}) vsem PWM modulatorjem, kot je to prikazano na sliki 2.



Slika 2. Idealizirano sosledje prožilnih (PWM) in vzorčevalnih (S/H) signalov na primeru dvo–vejnega pretvornika ter potek izhodnih tokov v stacionarnem obratovanju.

Posameznemu tokovnemu regulatorju moramo torej pripeljati informacijo o velikosti toka pripadajoče veje. Srednjo vrednost toka, ki je merodajna preneseni moči, najlažje ugotovimo z ovrednotenjem (Sample/Hold) merjenenega toka na sredini enega ali obeh preklopnih intervalov (t_{on}, t_{off}).

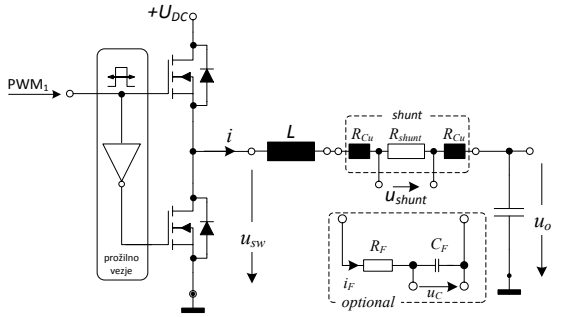
2.1 Napake pri zajemu merilnega signala

Navkljub veliki izbiri tokovnih merilnikov [4], [5] sta – iz povsem ekonomskih razlogov – v nadaljevanju opisani rešitvi še vedno najpogosteje uporabljeni.

2.1.1 Meritev toka s pomočjo merilnega upora

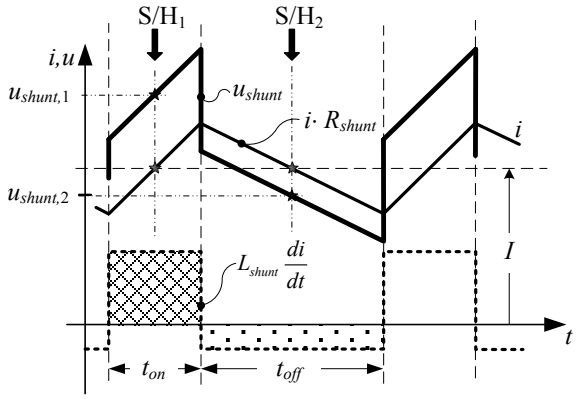
Po izvedbi merilne upore delimo na dvopolne in četveropolne (s t.i. Kelvin-ovima sponkama). Izvedenka s Kelvinovima sponkama je ugodnejša in zahteva manj pazljivosti pri oblikovanju tiskanine, saj zajamemo le padec napetosti na aktivnem delu upora (R_{shunt}) ter ne tudi padca napetosti na upornosti priključnih sponk in na kontaktni upornosti spajke (ponazorjeno z R_{Cu}).

Upornost slednje je pogosto težje določljiva in se s staranjem ter s temperaturo spreminja v mnogo večji meri kot upornost aktivnega dela upora.



Slika 3. Meritev toka z merilnim uporom s Kelvin-ovima sponkama.

Ker je izkoristek pretvorniških naprav pogosto v ospredju želja, smo seveda primorani vgrajevati merilne upore majhnih upornosti, pogosto pod 1 mΩ. V praksi se ob tem pogosto prezre vpliv lastne induktivnosti (aktivnega dela) upora (L_{shunt}), ki je pri SMD komponentah v grobem pogojena z velikostjo njihovega ohišja in znaša nekaj nH. Njen vpliv na potek zajetega signala (u_{shunt}) je tem večji čim večje so strmine tokov.



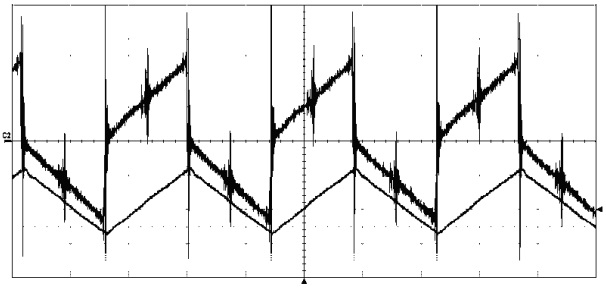
Slika 4: Vpliv lastne induktivnosti merilnega upora na obliko zajetega signala (u_{shunt}).

Iz slike 4 je razvidno, da je padec napetosti (u_{shunt}) enak vsoti padcev napetosti $R_{shunt} \cdot i$ ter $L_{shunt} \cdot \mathrm{d}i/\mathrm{d}t$, ki ima značilno pravokotno obliko. Srednja vrednost slednjega je v stacionarnem stanju enaka nič, vendar pa sta magnitudi signala v intervalu t_{on} in t_{off} različni. Ob zajemu signala na sredini preklopnih intervalov posledično izmerimo vrednosti

$$u_{shunt,1} = I \cdot R_{shunt} + \frac{L_{shunt}}{L + L_{shunt}} \cdot \left(U_{DC} - U_o\right)$$
$$u_{shunt,2} = I \cdot R_{shunt} - \frac{L_{shunt}}{L + L_{shunt}} \cdot U_o \quad , \quad (1)$$

ki odstopata od dejanske srednje vrednosti merjenega toka.

Relevantnost omenjenega padca napetosti je poudarjena s prikazom realnih razmer na sliki 5, ki kaže izmerjene veličine na primeru štiri-vejnega pretvornika (U_{DC} = 48 V) z vklopnim razmerjem $D = t_{on} / (t_{on}+t_{off}) = 0{,}5$ in z induktivnostjo vgrajene dušilke 6,8 μH. Vgrajeni merilni upor ima nazivno upornost 0,5 mΩ in lastno induktivnost 3 nH.



Slika 5: Izmerjeni tok dušilke: zgoraj – padec napetosti merjen neposredno na merilnem uporu (k_u = 2 mV/div); spodaj – izmerjeno s tokovnimi kleščami LEM-PR30 (k_i = 5 A/div), k_t = 5 μs/div.

V prikazanem primeru, kjer je srednja vrednost toka približno 8 A, bi izmerjeni vrednosti padca napetosti po izrazu (1) ustrezali približno toku enakega iznosa (≅ 4 A), a nasprotnega predznaka. V primeru, ko bi vrednosti povprečili bi storili še večjo napako.

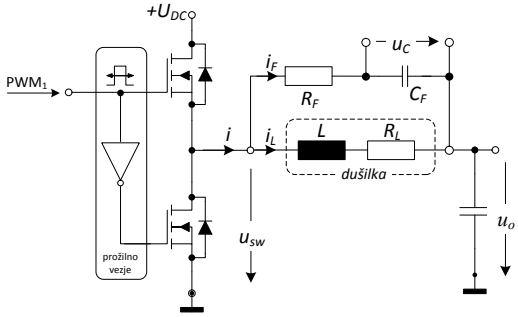
»Induktivni« padec napetosti v izrazu (1) lahko v digitalno zasnovanem regulatorju izničimo s povprečenjem obeh napetosti, ki ju predhodno pomnožimo z vklopnim razmerjem D oziroma z 1-D.

$$u^*_{shunt,1} = \left[I \cdot R_{shunt} + \frac{L_{shunt}}{L + L_{shunt}} \cdot \left(U_{DC} - U_o\right) \right] \cdot D$$
$$u^*_{shunt,2} = \left[I \cdot R_{shunt} - \frac{L_{shunt}}{L + L_{shunt}} \cdot U_o \right] \cdot \left(1 - D\right) \quad (2)$$

Povprečna vrednost napetosti je enaka $I \cdot R_{shunt}$. Ob tem velja opozoriti, da je vklopno razmerje poznano iz PWM modula.

2.1.2 Posredna meritev toka s filtracijo padca napetosti na dušilki

Posredno lahko tok izmerimo – brez dodatno vgrajenih merilnih uporov – izključno s poznavanjem lastne upornosti vgrajenih dušilk ter filtrskih komponent R_F in C_F.



Slika 6. Meritev toka s filtracijo padca napetosti na dušilki.

Padec napetosti na dušilki, podan v obliki kompleksnega zapisa, znaša $U_L = (R_L + j\omega L)I_L$. Ob upoštevanju, da je tok skozi filtrski komponenti zanemarljivo majhen ($i_L >> i_F$), lahko napetost u_C izrazimo kot

$$U_C = \frac{U_L}{1 + j\omega R_F C_F} = \frac{R_L + j\omega L}{1 + j\omega R_F C_F} I \qquad (3)$$

oziroma po preoblikovanju

$$U_C = R_L \left(\frac{1 + j\omega \cdot \tau_L}{1 + j\omega \cdot \tau_F} \right) I; \qquad \tau_L = \frac{L}{R_L}; \quad \tau_F = R_F C_F . \quad (4)$$

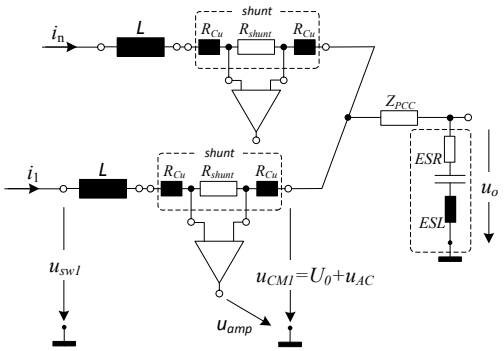
Iz (4) izhaja, da je napetost filtrskega kondenzatorja premosorazmerna merjenemu toku *i*, le če zagotovimo enaki časovni konstanti gladilne dušilke pretvorniškega vezja in dodanega RC filtra. Predpogoj za točno meritev toka je torej poznavanje parametrov dušilke, ki pa se s temperaturo (R_L) in amplitudo toka spreminjajo (nasičenje dušilke).

Z opisano tehniko je možno mnogo učinkoviteje zmanjšati prispevka lastne induktivnosti merilnega upora (slika 3). V ta namen RC filter namestimo preko merilnega upora, ki z lastno induktivnostjo tvori enako nadomestno vezje ($\tau = L_{shunt} / R_{shunt}$) kot je vezje realne dušilke na sliki 6. Primerjaje s parametri dušilke sta L_{shunt} in R_{shunt} od temperature in amplitude toka tako rekoč neodvisna, s čimer je možno pogoju ($\tau_L = \tau_F$) lažje zadostiti. Rešitev je enostavna in za razliko od izničenja po enačbi (2) mikrokrmilnika ne obremeni z dodatnim računanjem.

Neodvisno od uporabljene rešitve moramo merjeni napetosti (u_{shunt}, u_C) ojačiti in ju zaradi zajema z A/D pretvornikom prilagoditi, tako da sta izraženi proti referenčnemu potencialu (masi). Omenjeno izvedemo z merilnimi ojačevalniki [6]. Njihova vhodna ojačevalna stopnja mora vzdržati visoko napetost skupnega načina (*ang. common mode,* u_{CM}) ter jo na izhodu v čim širšem frekvenčnem območju oslabiti (ang. *CMRR*). Sočasno mora ojačiti majhen padec napetosti, ki je običajno več velikostnih razredov manjši. Ker se v praksi *CMRR* z naraščajočo frekvenco manjša, je pomembno, da merilnega upora (in filtrskega kondenzatorja) v vezje ne umestimo tako, da bi bil priključen v vozlišče tranzistorjev. V tem primeru bi namreč vhod merilnega ojačevalnika izpostavili višji napetosti skupnega načina ($u_{CM} = u_{sw}$), ki ima dodatno zaradi velike du/dt mnogo bolj neugoden spekter od napetosti po sliki 3 in 6, kjer je $u_{CM} = u_O$.

2.1.3 Sosedstveni učinki med merilnimi kanali

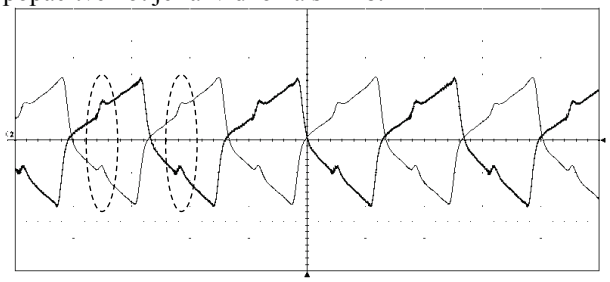
Ko govorimo o vplivu sosednjih tokov na zajem merilne veličine v opazovani veji, imamo najpogosteje v mislih presluh (induktivni ter kapacitivni sklop), ki nastopi zaradi nepazljive razporeditve elementov in povezovalnih vezic med merilnimi upori in ojačevalnikom. Slednji je razviden na sliki 5, kjer se v izmerjenem padcu napetosti, zajetem neposredno na merilnem uporu, opazijo tudi superponirane napetosti v trenutku preklopa v sosednjih vejah.



Slika 7. Nadomestno vezje izhodnega tokokroga za ponazoritev sosedstvenega vpliva.

V primeru močnostnih pretvornikov z velikim izhodnim tokom in posledično majhno upornostjo merilnih uporov ter visokim ojačenjem merilnega ojačevalnika, pa se soočimo tudi z njegovim nezadostnim slabljenjem (*CMRR*) pri visokih frekvencah.

Po sliki 7 napetost skupnega načina, ki jo na svojem vhodu zazna merilni ojačevalnik, ni konstantna (U_O) temveč vsebuje tudi visokofrekvenčne komponente napetosti (u_{AC}). Slednje nastopijo že zaradi končno velike kapacitivnosti gladilnega kondenzatorja. Merilni ojačevalnik na svojem vhodu dodatno zazna padce napetosti na ekvivalentni serijski upornosti in induktivnosti kondenzatorja (*ESR*, *ESL*) ter impedanci skupne priključitve (Z_{PCC}), ki jih povzroča vsota tokov $i_1 + \ldots + i_n$. Podobno kot pri obravnavi merilnega upora, tudi v imenovanem padcu napetosti po amplitudi in harmonskem spektru prevladuje prispevek obeh induktivnosti; *ESL* in Z_{PCC}. Zaradi zmanjšanega slabljenja se njihov prispevek v izhodni napetosti merilnega ojačevalnika manifestira v obliki sofazne popačitve kot je razvidno na sliki 8.

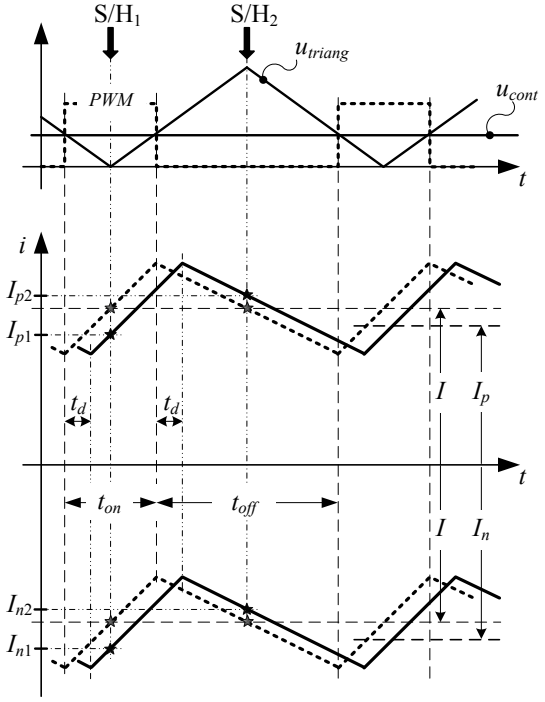


Slika 8: Izhodni napetosti (u_{amp}) merilnih ojačevalnikov v 3. in 4. veji – vhodni signal (u_{shunt}) je po obliki identičen poteku napetosti s slike 3.

2.2 Napake pri vzorčenju merilnega signala

Za stacionarno obratovanje je značilen potek toka skozi posamezno izhodno dušilko kot ga kaže slika 9. Če privzamemo brez–izgubne komponente, potem se ob spremembi izključno smeri pretoka moči, poteka tokov – razen v predznaku – ne razlikujeta (črtkano).

Ker tok niha okoli srednje vrednosti *I*, lahko njeno vrednost enostavno izmerimo, tako da merjeni tok vzorčimo na sredini časovnih intervalov t_{on} in t_{off}.



Slika 9. Vpliv zakasnitev prožilnega signala na vrednost vzorčenega toka.

Trenutek vzorčenja v vgrajenem A/D pretvorniku določata signala S/H_1 in S/H_2, ki ju sinhrono s PWM signalom generira DSP krmilnik. A ker prožilno vezje ob tem zakasni prožilne (PWM) signale, je dejanski potek toka (polna krivulja) zakasnjen za enak čas, s čimer v trenutku vzorčenja zajamemo vrednosti I_{p1} in I_{p2}

$$\begin{aligned} I_{p1} &= I - \frac{U_{DC} - U_o}{L} \cdot t_d \\ I_{p2} &= I + \frac{U_o}{L} \cdot t_d \end{aligned} \quad . \tag{5}$$

Odstopanje obeh od srednje vrednosti toka (I) je različno in pogojeno s strmino toka v pripadajočih intervalih

$$\frac{di}{dt} = \begin{cases} \dfrac{U_{DC} - U_o}{L}; \text{interval } t_{on} \\ -\dfrac{U_o}{L}; \text{interval } t_{off} \end{cases} . \tag{6}$$

Enako ovrednotimo odstopanje v primeru nasprotnega pretoka moči, ko se smer toka spremeni

$$\begin{aligned} I_{n1} &= I + \frac{U_{DC} - U_o}{L} \cdot t_d \\ I_{n2} &= I - \frac{U_o}{L} \cdot t_d \end{aligned} \quad . \tag{7}$$

V nasprotju z uveljavljenim prepričanjem se odstopanju od srednje vrednosti toka (I) ne izognemo niti, če vzorčeni par meritev povprečimo

$$\begin{aligned} I_p &= \frac{I_{p1} + I_{p2}}{2} = I + \frac{t_d}{L}(U_o - \frac{U_{DC}}{2}) \\ I_n &= \frac{I_{n1} + I_{n2}}{2} = I - \frac{t_d}{L}(U_o - \frac{U_{DC}}{2}) \end{aligned} \quad . \tag{8}$$

Zajeti vrednosti po (8) ustrezata pravi vrednosti (I) le v primeru, če je izhodna napetost enaki polovici vhodne. V nasprotnem, če tako zajeto in izračunano vrednost dejanskega toka pripeljemo na tokovni regulator, kjer se primerja z želeno vrednostjo I_{ref} (ki naj ustreza vrednosti I na sliki 9), potem bo seveda regulator skušal izničiti regulacijski pogrešek. Srednja vrednost dejanskega toka bo posledično pravzaprav višja od želene. V primeru negativnega toka ($i \langle 0$) pa nižja od želene vrednosti.

3 Sklepna misel

Od opisanih napak pri zajemu merjenega toka je najbolj nezaželen vpliv lastne induktivnosti merilnega upora in Z_{PCC}. V primeru kaskadne regulacije izhodne napetosti pretvornika, sicer napaka, če izvzamemo potencialno nesimetrijo med kanali, ni problematična, saj jo avtomatično odpravi nadrejeni napetostni regulator. Je pa v primeru dvosmernega pretoka potrebno izpostaviti z meritvijo vneseni izmik (ang. offset) razviden na sliki 9, ki postane mnogo bolj nezaželen v primeru neposredne regulacije toka pri uravnavanju pretoka med dvema izvoroma napetosti (baterijama).

Literatura

[1] D. Christen, S. Tschannen, and J. Biela, 'Highly efficient and compact DC-DC converter for ultra-fast charging of electric vehicles', in *Power Electronics and Motion Control Conference (EPE/PEMC), 2012 15th International*, 2012, p. LS5d–3.

[2] D. K. Saini, A. Reatti, and M. K. Kazimierczuk, 'Average current-mode control of buck dc-dc converter with reduced control voltage ripple', in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016, pp. 3270–3275.

[3] Y. Yan, F. C. Lee, and P. Mattavelli, 'Analysis and Design of Average Current Mode Control Using a Describing-Function-Based Equivalent Circuit Model', *IEEE Trans. Power Electron.*, vol. 28, no. 10, pp. 4732–4741, Oct. 2013.

[4] S. Ziegler, R. C. Woodward, H. H. C. Iu, and L. J. Borle, 'Current Sensing Techniques: A Review', *IEEE Sens. J.*, vol. 9, no. 4, pp. 354–376, Apr. 2009.

[5] A. Leban and M. Nemec, 'Tokovni merilni pretvorniki v vezjih močnostne elektronike', *Elektrotehniski Vestn.*, vol. 2015, no. 82(3), pp. 73–84.

[6] A. D. I. Engineeri, *Linear Circuit Design Handbook*, 1 edition. Amsterdam ; Boston: Newnes, 2008.

Analiza harmonske vsebine signala generiranega z bipolarno sinusno pulzno širinsko modulacijo

Denis Sušin

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

Harmonic content analysis of signal generated with bipolar sinusoidal pulse width modulation

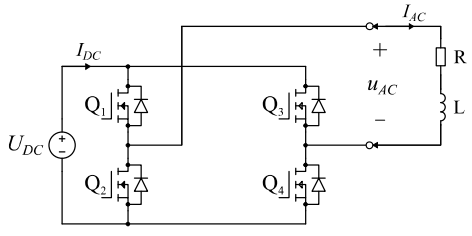
Abstract. *In this paper harmonic content of signal generated with bipolar sinusoidal pulse width modulation (PWM) are analyzed. The impact of four quantities is studied: discretization of the sinusoidal control signal, amplitude and frequency ratio between saw signal and control signal and finally the influence of the dead time. To maximize the amplitude of the first harmonic and to minimize higher harmonics content, the dead time is supposed to be as short as possible and the ratio between PWM frequency and control signal frequency should be high. However, this is not always possible, thus different characteristics are shown based on simulation results.*

1 Uvod

Pulzno širinska modulacija (PWM – angl. Pulse Width Modulation) je predvsem na področju močnostne elektronike zelo razširjena [1]. Zelo pogosto je uporabljena sinhronska sinusna bipolarna PWM, s ciljem generiranja sinusne napetosti oz. toka. Žal pa uporabljena modulacija generira tudi višjeharmonske komponente, ki pa velikokrat ostanejo prezrte oz. zanemarjene. Takšna poenostavitev v določenih aplikacijah ni dovoljena, zato je bil analiziran vpliv diskretizacije krmilnega signala, mrtvega časa ter amplitudnega in frekvenčnega razmerja med žago in krmilnim signalom. Simulacijski rezultati pridobljeni s programskim paketom Matlab so predstavljeni v nadaljevanju.

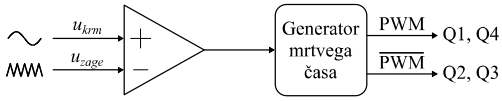
2 Kratek opis in realizacija bipolarne sinusne PWM

Tipična aplikacija za uporabo sinusne PWM je razsmernik.



Slika 1: Topologija enofaznega razsmernika, kjer je uporabljena bipolarna sinusna PWM

Enosmerno napetost U_{DC} in tok I_{DC} želimo pretvoriti v izmenično napetost U_{AC} in tok I_{AC} sinusne oblike. Temu cilju se uspemo približati s topologijo enofaznega mostičnega razsmernika, ki je prikazan na sliki 1. Shematski prikaz proženja tranzistorjev Q1, Q2, Q3 in Q4 je prikazan na sliki 2.



Slika 2: Shematski prikaz delovanja bipolarnega PWM generatorja

Za realizacijo PWM potrebujemo generator sinusnega signala s frekvenco f_{krm}, amplitudo $U_{amp,krm}$ in fazo φ_{krm}

$$u_{krm} = U_{amp,krm} \cdot \cos(2\pi \cdot f_{krm} \cdot t + \varphi_{krm}) \qquad (1)$$

ter generator žagastega signala s frekvenco f_{PWM} in amplitudo $U_{amp,zage}$,

$$u_{zage} = \begin{cases} U_{amp,zage} \cdot \left(1 - 4 \cdot \dfrac{t}{T_{PWM}}\right), \text{pogoj 1} \\ U_{amp,zage} \cdot \left(4 \cdot \dfrac{t}{T_{PWM}} - 1\right), \text{pogoj2} \end{cases} \qquad (2)$$

$$\begin{aligned} &\text{pogoj1}: kT_{PWM} + \frac{T_{PWM}}{2} > t \geq kT_{PWM} \\ &\text{pogoj2}: kT_{PWM} + T_{PWM} > t \geq kT_{PWM} + \frac{T_{PWM}}{2} \end{aligned} \qquad (3)$$

pri čemer je T_{PWM} obratna vrednost frekvence žage f_{PWM}; $k = 0, 1, 2, 3, \ldots$

Predpostavljamo, da je frekvenca žage natančen večkratnik frekvence krmilnega signala (od tu ime sinhronska PWM), kar pomeni, da je razmerje frekvenc m_f celo število:

$$m_f = \frac{f_{PWM}}{f_{krm}} \qquad (4)$$

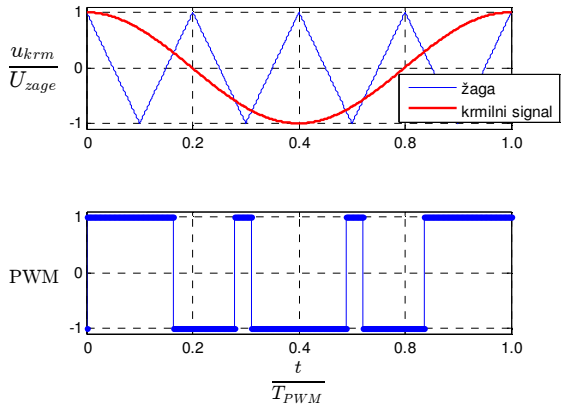
Razmerje amplitud med krmilnim signalom in žago m_a pa je:

$$m_a = \frac{U_{amp,krm}}{U_{amp,zage}} \qquad (5)$$

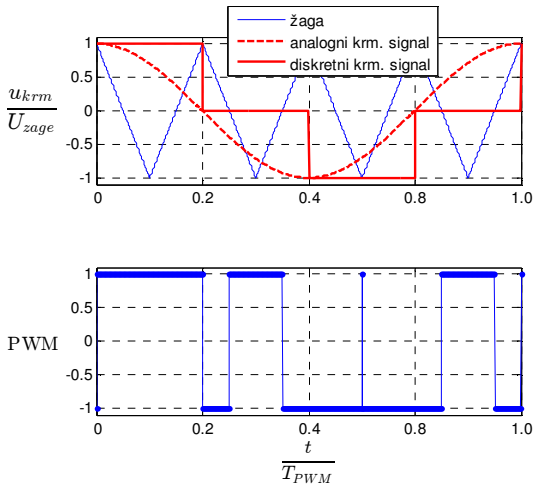
Žago in krmilni signal primerjamo in če je slednji večji, potem prevajata tranzistorja Q1 in Q4. V nasprotnem primeru prevajata tranzistorja Q2 in Q3. Izhodna napetost je tako lahko bodisi U_{DC}, bodisi $-U_{DC}$, od koder izhaja ime bipolarna PWM.

Pri generiranju PWM-a je obvezno še implementacija mrtvega časa t_d ob vsakem preklopu tranzistorjev, s čimer preprečimo hkratno prevajanje obeh tranzistorjev v posamezni veji pretvornika. Žal pa mrtvi čas povzroča parazitne efekte, ki se kažejo tudi v frekvenčnem spektru izhodne napetosti.

PWM je možno realizirati izključno z analognimi elektronskimi vezji (slika 3), lahko pa za generiranje PWM signalov uporabimo tudi mikroprocesor (slika 4). Ta je realiziran tako, da je krmilni signal znotraj periode PWM-a oz. žage, vzorčen le enkrat [2], s čimer zagotovimo le en preklop v posamezni periodi PWM-a. Posledica diskretizacije pa je, da so harmonske komponente izhodne napetosti u_{AC} vedno zakasnjene za $0{,}5 \cdot T_{PWM}$. Med obema pristopoma so razlike majhne, pa vendar ta izbira vpliva na frekvenčni spekter izhodne napetosti u_{AC}.

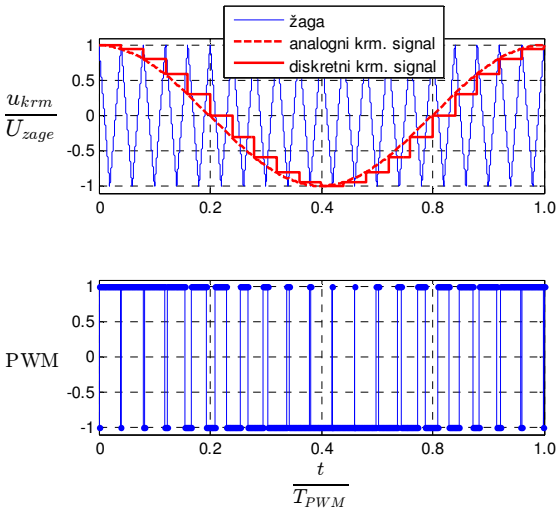


Slika 3: Časovni poteki analogne PWM za $m_f = 4$, $m_a = 1$ in mrtvi čas $t_d = 0$



Slika 4: Časovni poteki diskretne PWM za $m_f = 4$, $m_a = 1$ in mrtvi čas $t_d = 0$

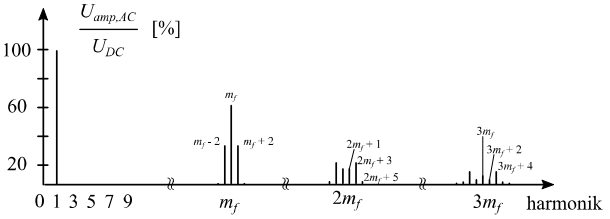
V praksi težimo k razmerju frekvenc $m_f \geq 20$ [3], kar pomeni, da so časovni poteki podobni razmeram s slike 5.



Slika 5: Časovni poteki diskretne PWM za $m_f = 20$, $m_a = 1$, in mrtvi čas $t_d = 0{,}01 \cdot T_{PWM}$

3 Analiza harmonske vsebine izhodne napetosti u_{AC} – simulacijski rezultati

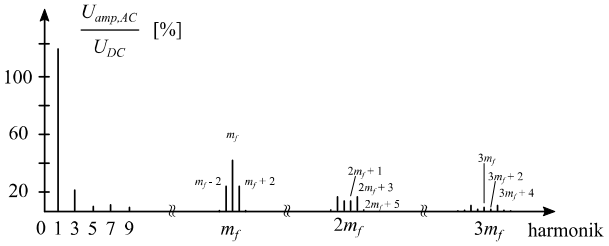
Želja je torej, da na izhodu razsmernika dobimo sinusno napetost (samo osnovno harmonsko komponento brez višjeharmonskih komponent). Žal pa se zaradi narave PWM-a ne moremo izogniti dejstvu, da bodo poleg osnovne harmonske komponente prisotne še vsaj harmonske komponente v okolici m_f, $2m_f$, $3m_f$ itd. Frekvenčni spekter idealne analogne PWM prikazuje slika 6. Slednje lahko omilimo le z uporabo pasivnih analognih močnostnih filtrov, kot so npr. LC, LCL ali CLC filtri [3], kar pa presega namen tega dela. Težimo k čim večjemu frekvenčnemu razmerju m_f, saj je tako lažje zadušiti ta del frekvenčnega spektra.



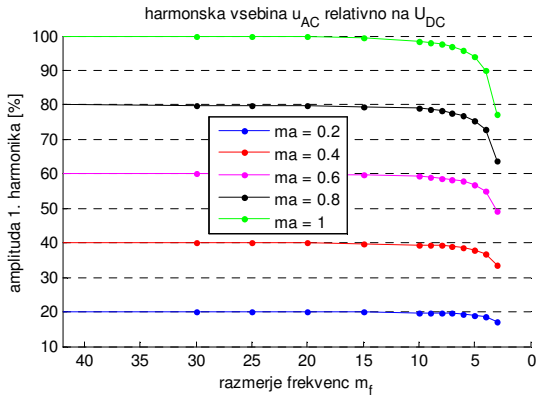
Slika 6: Frekvenčni spekter izhodne napetosti u_{AC} za analogno PWM brez mrtvega časa za $m_f \geq 20$, $m_a = 1$ in mrtvi čas $t_d = 0$, kjer so prikazane amplitude harmonskih komponent izhodne napetosti u_{AC} kot delež od napetosti U_{DC}

Žal pa pasivna vezja relativno slabo dušijo višjeharmonske komponente od 2 do m_f, ki pa so v realnih aplikacijah vsekakor prisotne. Vzroki za pojavljanje teh harmonskih komponent so: uporaba diskretne PWM, vpliv mrtvega časa in morebitno prekrmiljenje (primeri, ko je $m_a > 1$). Vpliv prekrmiljenja na frekvenčni spekter izhodne napetosti

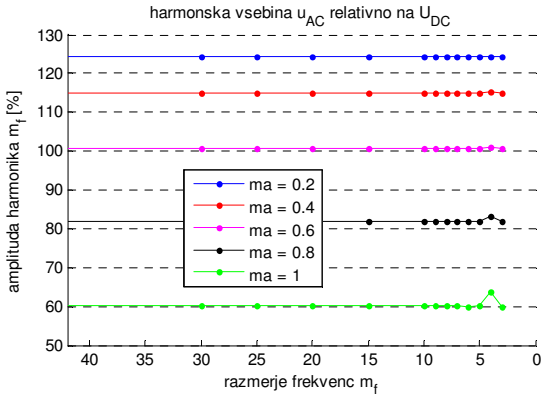
u_{AC} je prikazan na sliki 7. Opazimo lahko, da v primerjavi s sliko 6 amplituda osnovne harmonske komponente rahlo naraste, občutno se povečajo harmonske komponente 3, 5, 7, 9, itd., zmanjša pa se vsebnost harmonskih komponent v okolici m_f, $2m_f$, $3m_f$ itd. Zaradi tega se prekrmiljenja običajno izogibamo.



Slika 7: Frekvenčni spekter izhodne napetosti u_{AC} za analogno PWM brez mrtvega časa za za $m_f \geq 20$, $m_a = 1{,}6$ in mrtvi čas $t_d = 0$, kjer so prikazane amplitude harmonskih komponent izhodne napetosti u_{AC} kot delež od napetosti U_{DC}



Slika 8: Karakteristika amplitude 1. harmonske komponente napetosti u_{AC} v odvisnosti od razmerja frekvenc m_f za diskretno PWM; $t_d = 0$

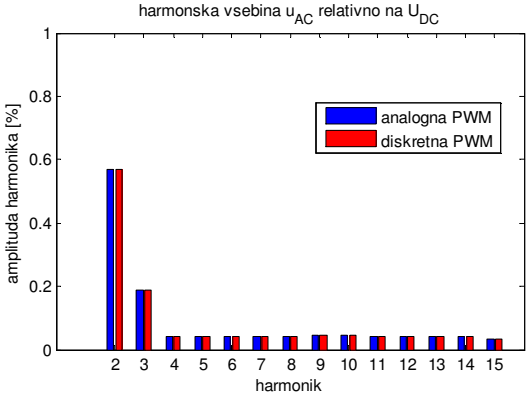


Slika 9: Karakteristika amplitude višjeharmonske komponente m_f napetosti u_{AC} v odvisnosti od razmerja frekvenc m_f za diskretno PWM; $t_d = 0$

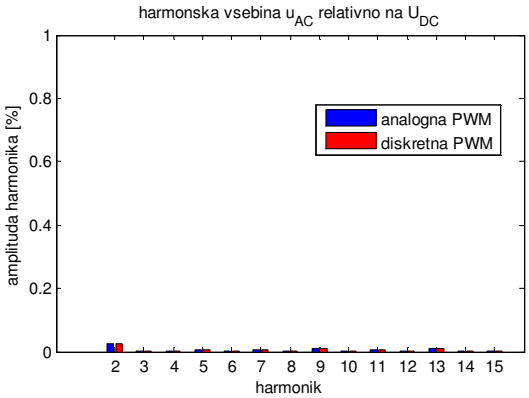
Slika 8 prikazuje amplitudo osnovne harmonske komponente izhodne napetosti u_{AC} v odvisnosti od m_f, za diskretno PWM. Prikazane so različne karakteristike za posamezno amplitudno razmerje m_a. Razmerje frekvenc m_f ne vpliva na amplitudo osnovne harmonske komponente, razen za zelo nizke vrednosti m_f, a samo v primeru, ko je PWM realizirana diskretno. Razmerje frekvenc m_f mora biti večje od 2, da še lahko govorimo o sinusni PWM, v nasprotnem primeru gre za enopulzno PWM (kar pomeni, da ni razlike, če je krmilni signal sinusen ali konstanten). Če pa je m_f zelo velik, razlike med analogno in diskretno PWM praktično ni.

Sprememba m_f na amplitudo višjeharmonske komponente m_f ne vpliva, pač pa občutno vpliva razmerje amplitud m_a. To velja tako za analogno kot tudi za diskretno PWM (slika 9). Manjše kot je razmerje amplitud m_a, višja je vsebnost višjeharmonske komponente m_f in okoliških harmonskih komponent. Prikazane so različne karakteristike za posamezno amplitudno razmerje m_a.

Sliki 10 in 11 (razlika je le v vrednosti m_f) prikazujeta vpliv diskretizacije krmilnega signala, kar se kaže kot povečanje amplitude tako sodih kot lihih višjeharmonskih komponent 2, 3, 4, 5, itd. Tudi to velja tako za analogno kot tudi za diskretno PWM. Če se m_f poveča, se vsebnost teh harmonskih komponent zmanjša.

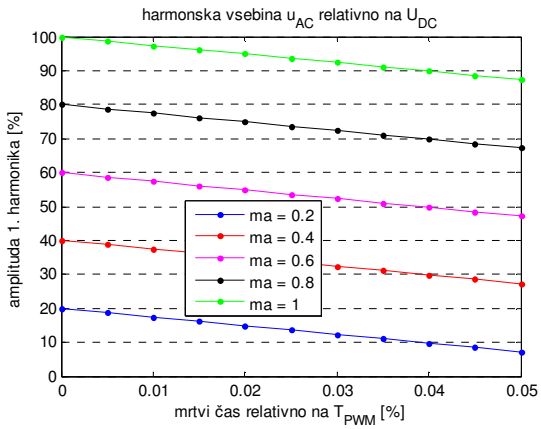


Slika 10: Del frekvenčnega spektra napetosti u_{AC} za $m_a = 1$, $m_f = 20$ ter $t_d = 0$

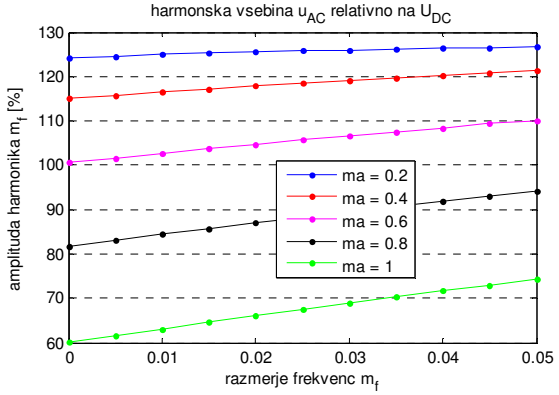


Slika 11: Del frekvenčnega spektra napetosti u_{AC} za $m_a = 1$, $m_f = 100$ ter $t_d = 0$

Vpliv mrtvega časa na amplitudo osnovne harmonske komponente napetosti u_{AC} prikazuje slika 12, na amplitudo višjeharmonske komponente m_f pa slika 13. Povečanje mrtvega časa povzroči zmanjšanje amplitude osnovne harmonske komponente ter povečanje amplitude višjeharmonske komponente m_f.

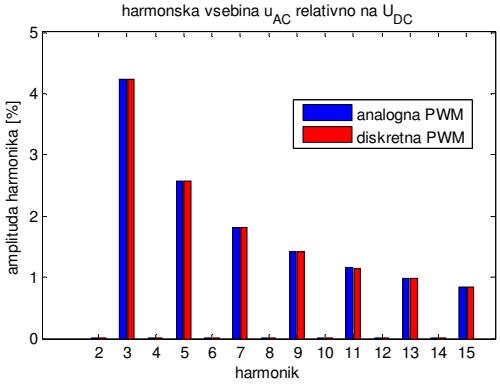


Slika 12: Karakteristika amplitude 1. harmonske komponente napetosti u_{AC} v odvisnosti od mrtvega časa t_d za diskretno PWM; $m_f = 400$

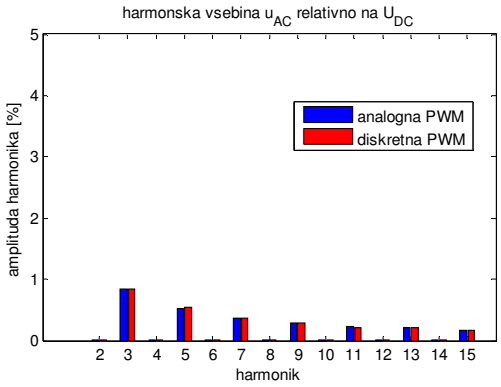


Slika 13: Karakteristika amplitude višjeharmonske komponente m_f napetosti u_{AC} v odvisnosti od mrtvega časa t_d za diskretno PWM; $m_f = 400$

Mrtvi čas povzroči le lihe višjeharmonske komponente 3, 5, 7 itd., kar prikazujeta sliki 14 in 15. Večje vrednosti mrtvega časa povzročijo večje vrednosti amplitud višjeharmonskih komponent.



Slika 14: Del frekvenčnega spektra napetosti u_{AC} za m_a = 1, $m_f = 400$ ter $t_d = 0{,}05 \cdot T_{PWM}$



Slika 15: Del frekvenčnega spektra napetosti u_{AC} za m_a = 1, $m_f = 400$ ter $t_d = 0{,}01 \cdot T_{PWM}$

4 Zaključek

Analiziran je frekvenčni spekter in ovrednoten vpliv določenih parametrov na izhodno napetost enofaznega razsmernika u_{AC}, kjer je uporabljena bipolarna sinusna PWM. Preučen je tudi vpliv diskretne PWM, ki je realizirana v mikroprocesorju. Ta povzroči upad amplitude osnovnega harmonika in zakasnitev za $0.5\,T_{PWM}$ ter prisotnost tako sodih kot lihih višjeharmonskih komponent (2, 3, 4, 5, …), amplituda teh je ranga velikosti 1% (normirano na U_{DC}). Manjše razmerje frekvenc m_f povzroči večjo amplitudo višjeharmonskih komponent. Diskretizacija krmilnega signala pa praktično nič ne vpliva na del spektra v okolici harmonskih komponent m_f.

Precejšen vpliv na frekvenčni spekter u_{AC} ima tudi mrtvi čas, saj povzroči zmanjšanje amplitude osnovnega harmonika ter porast lihih harmonikov (3, 5, 7, …), amplitude harmonikov v okolici višjeharmonskih komponent m_f, $2m_f$, $3m_f$, itd. pa se tudi povečajo.

Prikazani rezultati potrjujejo splošno nenapisano pravilo, da naj bo razmerje frekvenc m_f čim večje ter mrtvi čas t_d čim manjši. Na ta način dosežemo, da bo amplituda osnovne harmonske komponente največja, amplitude višjeharmonskih komponent pa najmanjše.
Ker pa to ni vedno izvedljivo, je zelo koristno poznati vpliv posameznih parametrov na frekvenčni spekter.

Literatura

[1] N. Mohan, T. M. Undeland, in W. P. Robbins, *Power electronics: converters, applications, and design*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2003.

[2] J. Richardson, „Implementation of a PWM Regular Sampling Strategy for AC Drives“, str. 11.

[3] 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.

Načrtovanje, zasnova in izdelava večfaznega razsmernika z MOSFET stikali

Andraž Rihar, Peter Zajec, Mitja Nemec, Marko Petkovšek, Danjel Vončina

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

Design and development of a multiphase high voltage MOSFET inverter

Abstract. *Drivetrain electrification has lately been gaining momentum. Multiphase machines and corresponding inverters promise interesting advantages over ordinary three phase configurations. As an alternative to well-established high-voltage IGBT and SiC MOSFET based inverters, this paper presents design and development of a medium voltage multiphase MOSFET inverter for 40 kW peak power. Each phase of the inverter comprises four parallel branches. The designed inverter offers a compact, yet modular design, having good potential for future research in the field of multiphase MOSFET inverters.*

1 Uvod

V zadnjem obdobju se stopnja elektrifikacije glavnih ali pomožnih pogonov osebnih in drugih vozil opazno povečuje [1, 2], bodisi v obliki hibridnih ali popolnoma električnih rešitev [3]. Poleg trifaznih motorjev se zaradi specifičnih prednosti, kot so manjša valovitost navora, večja zanesljivost, razporeditev moči po fazah in zmanjševanje izgubnih moči, kot vedno bolj primerni zdijo tudi pet, šest in večfazni pogoni [4, 5, 6].

Za napajanje takih motorjev so običajno uporabljeni večfazni dvo- ali večnivojski pretvorniki [7, 8, 9]. Prednosti slednjih so manjše harmonsko popačenje izhodne napetosti (THD, angl. total harmonic distortion), manjše sofazne napetosti (angl. common-mode voltages), manjše preklopne izgube zaradi nižjih strmin napetosti ob preklopih ter predvsem možnost uporabe stikal za nižje napetosti [10]. Do nedavnega se je razvoj razsmernikov v grobem delil na dve področji in sicer na visokonapetostne razsmernike (nad 400 V) z IGBT stikali [11], ki jih šele v zadnjem času počasi nadomeščajo visokonapetostni (nad 600 V) SiC MOSFETi [12, 13], ter nizkonapetostne razsmernike (do 100 V ali 150 V) z MOSFET polprevodniškimi stikali [14]. Razmeroma prosto ostaja področje med obema napetostma zaradi dosedanjih tehnoloških omejitev močnostnih stikal, predvsem omejenih tokovnih zmogljivosti srednjenapetostnih MOSFET-ov [15].

Potrebam po višji moči pogona je moč zadostiti z uporabo višjih napetosti in polprevodniških stikal novih tehnologij, kot sta SiC in GaN [13]. Tokovno zmogljivost je možno povečati z vzporedno vezavo stikal ali tranzistorskih vej. Tako je predvidoma možno poenostaviti prepotrebno odvajanje izgubnih moči, zagotoviti primerno nizko tokovno obremenitev posameznega tranzistorja, izboljšati kompaktnost, nadalje pa tudi poenostaviti montažo razsmernika v sklopu aplikacije. V primeru vzporedne vezave faz se ponuja tudi možnost prepletanja proženja posameznih vej razsmernika, kot je to že dodobra uveljavljena praksa pri DC/DC pretvornikih [16]. Prepletanje vej bi morda lahko dodatno zmanjšalo tudi valovitost navora.

Raziskave v omenjenih smereh in pripadajoče testiranje vplivov posameznih ukrepov so seveda mogoče le s primerno namensko strojno opremo, v prvi vrsti torej primernim razsmernikom. Ta mora ustrezati osrednjim električno-mehanskim zahtevam, med katerimi sta pomembni predvsem primerna kompaktnost z namenom zmanjševanja vpliva parazitnih induktivnosti in kapacitivnosti, ter modularnost sistema, ki omogoča testiranja vpliva števila izbranih faz, različnih kombinacij proženja, profiliranja izkoristka in zmogljivosti različnih polprevodniških stikal.

V literaturi običajno zasledimo opise sistemov, ki so zasnovani namensko glede na specifično aplikacijo [14], optimizirani sistemi za širši spekter uporabe pa so redki. V tem članku so predstavljeni koraki načrtovanja, zasnove in izdelave laboratorijskega vzorca večfaznega razsmernika z MOSFET stikali, pri čemer pa razviti razsmernik predstavlja platformo za nadaljnje raziskave na področju razsmernikov s srednje-napetostnimi MOSFET stikali.

2 Zasnova in izdelava podsklopov večfaznega razsmernika

Načrtovanje, zasnova in izdelava razsmernika so potekali v več zaporednih korakih. Glede na obratovalne zahteve je bilo potrebno definirati okvirne dimenzije razsmernika, izbrati ustrezne komponente, testirati ustreznost posameznih rešitev, nato pa zasnovati, izdelati in sestaviti celoten laboratorijski vzorec.

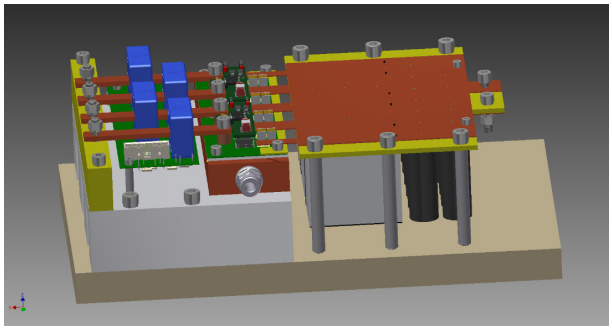
2.1 Obratovalne zahteve

Namen raziskave je zasnovati in izdelati večfazni razsmernik s srednjenapetostnimi MOSFET stikali, ki bo na primer lahko uporabljen za napajanje šestfaznega motorja s trajnimi magneti kratkotrajne moči 40 kW in trajne moči 20 kW. Nazivna napetost enosmernega vmesnega tokokroga je 135 V. Razsmernik naj bo kompakten in čim bolj simetrično zgrajen. Potrebno je kapljevinsko hlajenje za ustrezen odvod izgubnih moči. Stopnja kompaktnosti je lahko rahlo poenostavljena s ciljem zagotavljanja nemotenih laboratorijskih meritev, ki bodo omogočile nadaljnja testiranja vodenja, vpliva izbire polprevodniških elementov, gonilnikov, itn. Na podlagi obstoječih študij na temo ocene izgubnih moči, prednosti in slabosti dvo- in trinivojskih izvedb razsmernikov z MOSFETi [17, 18], je bila za podane električno-mehanske zahteve zaradi nižjih izgubnih

moči, manjše odvisnosti izgub od obremenitve in manjšega števila komponent izbrana 6 – fazna dvonivojska izvedba razsmernika z združenim enosmernim vmesnim tokokrogom. Zasnova posamezne faze iz več tranzistorskih vej bo lahko z namenskim krmiljenjem omogočila tudi več kot 6-fazno delovanje.

2.2 Izdelava in testiranje prototipa ene fazne veje

V prvem koraku je bila izvedena optimizacija prostorske postavitve močnostnih, krmilnih in zaščitnih komponent ter hladilnega sistema v programih Altium Designer in Autodesk Inventor (glej sliko 1) z osrednjim ciljem zmanjševanja razdalj med posameznimi podsklopi in minimizacije parazitnih parametrov sistema. Načrtovanju sistema je sledila izbira komponent, nato izdelava posameznih podsklopov in sestava prototipa ene fazne veje dvonivojskega pretvornika (glej sliko 1).



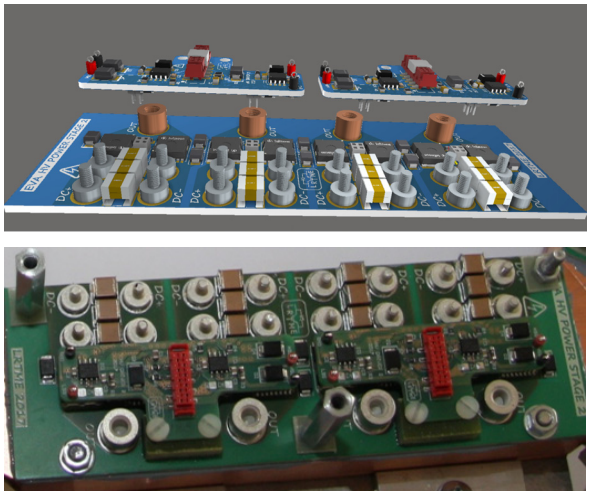
Slika 1. Grafični prikaz prototipa ene faze dvonivojskega pretvornika s štirimi paralelnimi vejami MOSFET stikal. Z leve proti desni so prikazani merilni del, močnostni del in kondenzatorski paket.

Sledilo je testiranje posameznih podsklopov ene fazne veje z naslednjimi osrednjimi cilji: a) ovrednotenje ustreznosti uporabe topologije več paralelnih vej za izbran nivo napetosti in moči, b) ocena uspešnosti minimizacije prenapetosti z izbranimi koncepti optimizacije postavitve komponent in c) evalvacija primernosti izbranega koncepta vodnega hlajenja.

Na podlagi analize omenjenih točk so bili izvedeni popravki nekaterih podsklopov, temu pa je sledilo načrtovanje in izdelava podsklopov večfaznega razsmernika, torej močnostne stopnje z gonilniškim delom, hladilnega sistema, enosmernega vmesnega tokokroga ter krmilne plošče.

2.2.1 Močnostna stopnja z gonilniškim delom

Močnostni del je nameščen na izoliranem metaliziranem substratu (IMS, angl. insulated metal substrate) na aluminijasti osnovi s 105 µm nanosom bakra in je izveden s štirimi vzporednimi vejami s skupno osmimi MOSFET stikali podjetja Infineon (IPT210N25NFD), nazivne napetosti do 250 V (glej sliko 2). Na močnostni del je z zgornje strani možno neposredno namestiti dve gonilniški plošči s skupno 4 gonilniki AUIRS21811S. Pri načrtovanju močnostnega dela je bila posebna pozornost posvečena doseganju majhnih razdalj med elementi in posledično zmanjševanju parazitnih induktivnosti in pripadajočih efektov.

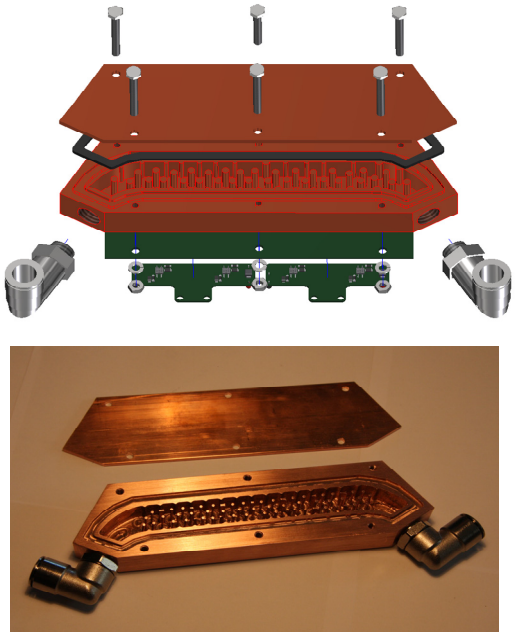


Slika 2. Grafični prikaz (zgoraj) in fotografija (spodaj) močnostnega dela s po dvema gonilniškima ploščama.

Pri načrtovanju gonilniškega dela je bila posebna pozornost posvečena doseganju ustrezne simetrije povezav od priklopa do posameznih gonilnikov, s čimer je zagotovljena čim večja enakost vklopnih trenutkov in posledično delovanja oziroma simetrije vzporedno obratujočih tranzistorskih vej.

2.2.2 Hladilni sistem

Izhajajoč iz podatkov simulacij in termičnih testiranj je bila izvedena posodobitev hladilnih blokov, namenjenih vodnemu hlajenju razsmernika. Načrtani so bili novi hladilni bloki, ki jih je možno simetrično zložiti v heksagon.



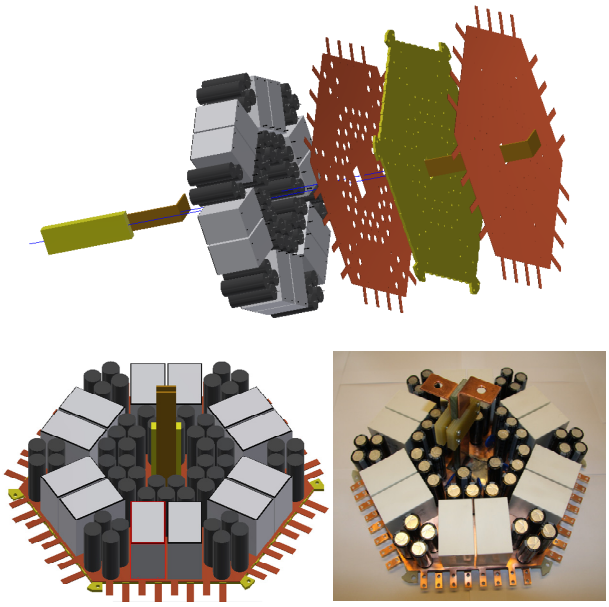
Slika 3. Tridimenzionalni model (zgoraj) in končna izvedba (spodaj) hladilnega bloka z namenskim izrezkanim kanalom za izboljšan odvod toplote.

Prvotna preprosta izvrtina je bila nadomeščena z izrezkanim kanalom z namenskimi pini, ki povečajo neposredno stično površino vode in bakrenega profila, s tem pa dodatno izboljšajo odvod toplote (glej sliko 3). Za tesnjenje bakrenega profila in pripadajočega bakrenega pokrova so uporabljeni namenski O-ringi

podjetja DIHTA, za priklop vode pa namenski priklopniki (angl. fittingi) podjetja FESTO.

2.2.3 Združeni enosmerni vmesni tokokrog

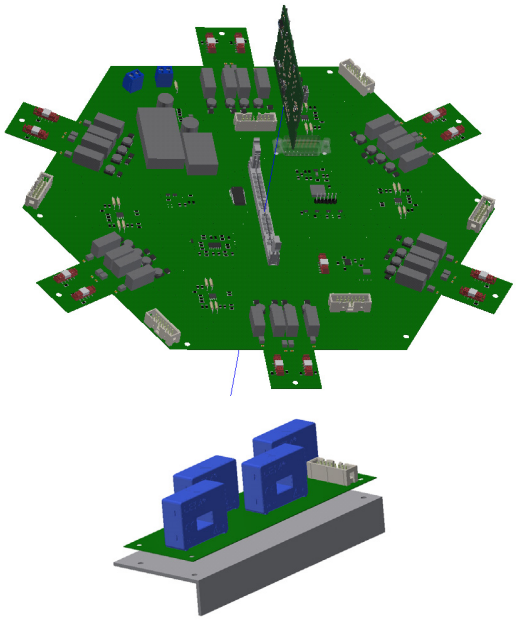
Ker ima vsaka tranzistorska veja svoj gonilnik je z namenskim krmiljenjem predvidoma mogoče vplivati na valovitost toka enosmernega vmesnega tokokroga. V ta namen je bil izdelan združen enosmerni vmesni tokokrog, ki vključuje vzporedno vezane folijske (KEMET, 60 µF) in elektrolitske kondenzatorje (Rubycon, 560 µF) za nazivne napetosti do 220 V in namenske varistorje za zaščito pred morebitnimi prenapetostmi. Uporabljena je tako imenovana sendvič konstrukcija, s katero zmanjšamo parazitne efekte (glej sliko 4). Skupna kapacitivnost paketa je malo manj kot 30 mF. Folijski kondenzatorji so postavljeni bližje priklopnim sponkam razsmernika zaradi večje obremenljivosti in boljšega dušenja povratnih elektromagnetnih motenj. Kondenzatorji so razporejeni v smislu zagotavljanja simetrije posameznih faz. Priklopni priključki so pripravljeni za vijačenje na stebričke močnostnega dela.



Slika 4. Eksplodiran prikaz (zgoraj) tridimenzionalnega modela (spodaj levo) združenega enosmernega vmesnega tokokroga. Končni vzorec kombinacije sivo-belih folijskih in črnih elektrolitskih kondenzatorjev je prikazan spodaj desno.

2.2.4 Krmilna plošča in merilniki toka

Izdelana je bila namenska krmilna plošča z mikrokrmilnikom serije Delfino podjetja Texas Instruments, ki skrbi za napajanje gonilniških plošč in ustrezno krmiljenje sistema (glej sliko 5). Posebna pozornost pri načrtovanju je bila namenjena zagotavljanju dovolj velikega števila signalov za ustrezno krmiljenje posamičnih parov tranzistorskih vej in seveda primerni tokovni zmogljivosti napajalnih povezav ob večji obremenitvi gonilniških vezij. Dodatno je bila upoštevana ločena razporeditev napajalnih in logičnih linij v izogib morebitnim elektromagnetnim motnjam. Pripravljene so bile plošče s po štirimi merilniki toka LEM, ki so namenjene za merjenje izhodnega toka posameznih vej vsake faze (glej sliko 5). Poglavitni namen je možnost analize ustreznosti vzporedne vezave vej in razporeditve obremenitve razsmernika.



Slika 5. Tridimenzionalna modela krmilne plošče z mikrokrmilnikom Delfino (zgoraj) in plošče z merilniki izhodnega toka LEM (spodaj).

2.2.5 Programska oprema in uporabniški vmesnik

Vzporedno z zadnjo točko je potekala preliminarna priprava regulacijskih algoritmov vodenja omenjenega sistema, ki omogoča neodvisno proženje posameznih parov paralelnih vej. Pripravljen je bil namenski grafični uporabniški vmesnik za preliminarno testiranje različnih načinov obratovanja razsmernika. Ta tematika presega namen članka, zato bralca vljudno vabimo k ogledu podrobnejšega opisa v [19].

3 Sestava in zagon večfaznega razsmernika

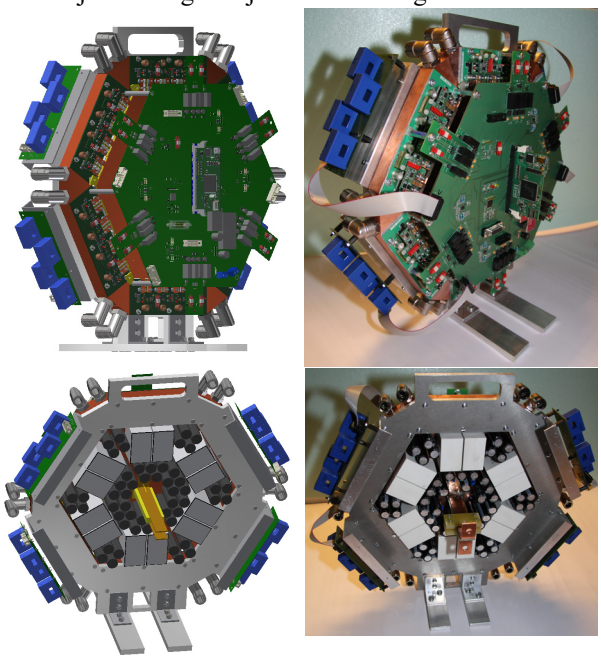
Izdelavi sestavnih podsklopov je sledilo testiranje ustreznosti delovanja posameznih podsklopov, nato pa sestava razsmernika in postopni zagon sistema kot celote (glej sliko 6). Najprej so bili na nosilno ogrodje montirani hladilni bloki s pripadajočimi močnostnimi in gonilniškimi stopnjami. Temu je sledila namestitev enosmernega vmesnega tokokroga, ki je na priključne stebričke vijačen z medeninastimi vijaki. Zatem so bile sistemu dodane še plošče z LEM sondami in končno še krmilna plošča z ustreznimi distančniki.

V prvem koraku je bilo izvedeno testiranje delovanja mikrokrmilnika, preverjena pravilnost vseh krmilnih signalov, nato pa pri nizki napajalni napetosti še pravilnost delovanja tranzistorskih stikal. Sledila je lažja obremenitev in preliminarne meritve napetosti in tokov.

4 Sklepne ugotovitve

Upoštevajoč električno-mehanske zahteve je bil zasnovan in izdelan laboratorijski vzorec 6 – faznega razsmernika z MOSFET stikali. Razsmernik je izdelan kompaktno in razmeroma optimizirano, kljub temu pa omogoča modularnost, nemoteno izvajanje meritev v

laboratorijskem okolju, s čimer predstavlja odlično izhodišče za nadaljnja testiranja in raziskave. Predvideno je nadaljnje delo v sklopu vključitve v postroj z namenskim večfaznim, večsistemskim strojem, kar bo omogočilo oceno ustreznosti različnih principov vodenja, primerjavo uporabe prepletanja vej, različnih močnostnih stikal, itn. Razsmernik lahko namreč zaradi štirih vzporedno vezanih vej v vsaki fazi omogoča delovanje v konfiguraciji do 24 – faznega sistema.



Slika 6. Sprednji (zgoraj) in zadnji (spodaj) pogled na tridimenzionalni model (levo) in končno izvedbo (desno) laboratorijskega vzorca 2 x 3 – faznega razsmernika.

Možnost testiranja bo pomemben korak v smeri spoznavanja in uveljavitve večfaznih pogonov, ki bodo zaradi specifičnih prednosti v prihodnosti lahko imeli velik potencial za uspešno elektrifikacijo glavnih pogonskih sklopov vozil.

Zahvala

Delo sta sofinancirali Republika Slovenija ter Evropska unija v sklopu projekta EVA4green (OP20.00362).

Literatura

[1] https://www.greentechmedia.com/articles/read/how-vehicle-electrification-will-evolve-in-2018#gs.oKSpb2o, dostopano dne 8.6.2018.

[2] A. Lajunen, P. Sainio, L. Laurila, J. Pippuri-Mäkeläinen, K. Tammi: Overview of Powertrain Electrification and Future Scenarios for Non-Road Mobile Machinery. *Energies*, 11(5), str. 1184, 2018.

[3] Y. Yang, K. Arshad-Ali, J. Roeleveld, A. Emadi: State-of-the-art electrified powertrains-hybrid, plug-in, and electric vehicles. *International Journal of Powertrains*, 5(1), str. 1-29, 2016.

[4] F. Barrero, M. J. Duran: Recent advances in the design, modeling, and control of multiphase machines—Part I. *IEEE Transactions on Industrial Electronics*, 63(1), str. 449-458, 2016.

[5] E. Levi: Advances in converter control and innovative exploitation of additional degrees of freedom for multiphase machines. *IEEE Transactions on Industrial Electronics*, 63(1), str. 433-448, 2016.

[6] N. K. Nguyen, F. Meinguet, E. Semail, X. Kestelyn: Fault-tolerant operation of an open-end winding five-phase PMSM drive with short-circuit inverter fault. *IEEE Transactions on Industrial Electronics*, 63(1), str. 595-605, 2016.

[7] S. Bhattacharya, D. Mascarella, G. Joós, J. M. Cyr, J. Xu: A dual three-level T-NPC inverter for high-power traction applications. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 4(2), str. 668-678, 2016.

[8] R. Teichmann, S. Bernet: A comparison of three-level converters versus two-level converters for low-voltage drives, traction, and utility applications. *IEEE Transactions on Industry Applications*, 41(3), str. 855-865, 2005.

[9] M. Schweizer, T. Friedli, J. W. Kolar: Comparative evaluation of advanced three-phase three-level inverter/converter topologies against two-level systems. *IEEE Transactions on industrial electronics*, 60(12), str. 5515-5527, 2013.

[10] L. G. Franquelo, J. Rodriguez, J. I. Leon, S. Kouro, R. Portillo, M. A. Prats: The age of multilevel converters arrives. *IEEE industrial electronics magazine*, 2(2), 2008.

[11] J. Zhang: High performance control of a three-level IGBT inverter fed ac drive. In *Industry Applications Conference, Thirtieth IAS Annual Meeting, IAS'95.*, vol. 1, str. 22-28, Okt. 1995, IEEE.

[12] H. Kim, H. Chen, J. Zhu, D. Maksimović, R. Erickson: Impact of 1.2 kv sic-mosfet EV traction inverter on urban driving. In *Wide Bandgap Power Devices and Applications (WiPDA), 2016 IEEE 4th Workshop on*, str. 78-83, Nov. 2016, IEEE.

[13] E. Gurpinar, A. Castellazzi: Single-phase T-type inverter performance benchmark using Si IGBTs, SiC MOSFETs, and GaN HEMTs. *IEEE Transactions on Power Electronics*, 31(10), str. 7148-7160, 2016.

[14] B. A. Welchko, M. B. de Rossiter Correa, T. A. Lipo: A three-level MOSFET inverter for low-power drives. *IEEE Transactions on Industrial Electronics*, 51(3), str. 669-674, 2004.

[15] H. Kernstock, B. Plassnegger: High efficiency soft switched 3-level MOSFET Inverter for an Electric Vehicle PMSM Drive. In *Power Electronics and Motion Control Conference (EPE/PEMC), 15th International*, str. DS1b-9, Sep. 2012, IEEE.

[16] O. García, P. Zumel, A. De Castro, A. Cobos: Automotive DC-DC bidirectional converter made with many interleaved buck stages. *IEEE Transactions on Power Electronics*, 21(3), str. 578-586, 2006.

[17] A. Rihar, D. Vončina, R. Fišer, H. Lavrič: Primerjava analitičnega in numeričnega izračuna izgubnih moči pri trifaznem razsmerniku z MOSFETi. V: *Proceedings of the Twenty-sixth International Electrotechnical and Computer Science Conference ERK 2017*, str. 295-298, Sep. 2017, IEEE.

[18] B. Prah, A. Rihar, D. Vončina: Simulacije in konceptualna zasnova trinivojskega pretvornika. V: *Proceedings of the Twenty-sixth International Electrotechnical and Computer Science Conference ERK 2017*, str. 291-294, Sep. 2017, IEEE.

[19] M. Nemec, A. Rihar: Načrtovanje programske opreme za vodenje večfaznega pogona. *ERK 2018*, poslano v objavo, 2018.

Načrtovanje, izdelava in preliminarno testiranje eksperimentalnega modela trinivojskega pretvornika

Blaž Prah, Andraž Rihar, Danjel Vončina

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, 1000 Ljubljana
E-pošta: prah.blaz@gmail.com

Design, assembly and preliminary testing of a three-level inverter

Abstract. *Three-level inverters are becoming a strong alternative to the well-established two-level inverters, especially under specific operating conditions. As laboratory-based inverter setups are usually less compact, this paper provides a description of design and assembly of a laboratory-based, but medium power three-level MOSFET inverter of compact design. Results of preliminary testing under small load are provided for passive and active NPC operation. Active control of the DC link's midpoint is presented and to some extent its advantages are verified through measurements. Finally, critical ideas for further work are presented.*

1 Uvod

Zaradi stalnega naraščanja števila avtonomnih električnih pogonov, fotovoltaičnih elektrarn in podobnih postrojev naraščajo potrebe po zmogljivih močnostnih pretvornikih, ki jih morajo odlikovati kompaktnost, nizko popačenje izhodne napetosti (angl. THD – total harmonic distortion), visok izkoristek in sprejemljiva cena izdelave [1, 2].

Nenehno višanje potrebne moči narekuje potrebo po višjih napajalnih napetostih (nad 500 V) [3], hkrati pa morajo slednjim zadostiti tudi uporabljena močnostna polprevodniška stikala [4]. Tu se pokaže ena izmed glavnih prednosti premagovanja nižjih preklopnih strmin stikal [13] v večnivojskih pretvorniških topologijah, katere okvirno delimo na a) pretvornik z ločilnimi diodami (angl. NPC – Neutral Point Clamped), b) pretvornik z ločilnimi kondenzatorji (angl. FC – Flying Capacitor) in c) kaskadni mostični večnivojski pretvornik (angl. Cascaded H-bridge) [4]. Na področju višjih napetosti prevladujejo IGBT stikala nazivnih napetosti 600 V ali 1200 V [3, 5, 6], na področju nizkih napetosti (do 50 V) in nižjih moči pa srečamo uporabo predvsem MOSFET-ov, ki omogočajo delovanje pri višjih stikalnih frekvencah in imajo običajno manjše izgube zaradi manjše upornosti kanala. Pretvornikom, ki uporabljajo MOSFET stikala se poveča močnostna zmogljivost tako, da se vzporedno veže več stikal ali pretvorniških vej [7].

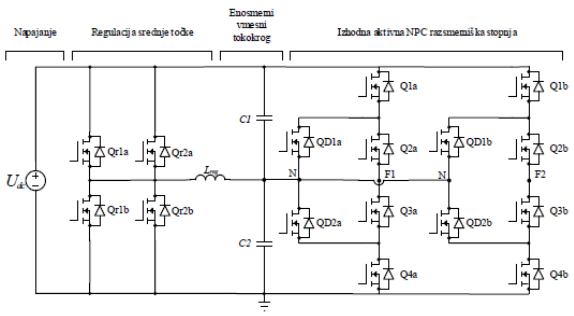
Zaradi novih tehnologij (CoolMOS) [8] in novih materialov (SiC, GaN), ki omogočajo višje izkoristke kljub večjim upornostim kanala v primerjavi s klasičnimi MOSFET-i, se na področju visokih napetosti počasi uveljavljajo tudi visokonapetostna MOSFET stikala, katerih ciljno napetostno področje je nad 600 V, hkrati pa nove tehnologije omogočajo tudi višje stikalne frekvence [9, 10]. Študije kažejo, da nove tehnologije omogočajo nižje izgube, to pa ponuja tudi možnost zmanjšanja površine hladilnih teles ob klasičnem odvajanju toplote s pomočjo hladilnih reber [12, 14]. Termične študije kapljevinskega hlajenja so zaenkrat še pomanjkljive.

Ob pregledu literature je moč zaslediti veliko izdelanih eksperimentalnih modelov večnivojskih pretvornikov, predvsem različnih podtipov, kot so A-NPC ali T-NPC, ki imajo namesto pasivnih elementov stikala [12]. Pri izdelavi eksperimentalnih modelov so avtorji manj pozorni na kompaktnost izvedbe, kar lahko med drugim povzroča parazitne induktivnosti in prenapetosti, tako ne podaja verne slike delovanja sistema, končno pa lahko povzroča tudi EMI ali EMC težave, saj pretvorniki delujejo pri visokih frekvencah [11]. Rešitve se kažejo v kompaktni realizaciji pretvornika, kar zmanjša vpliv parazitnih efektov in izboljša delovanje pretvornika. Manj nihanja napetosti namreč pomeni manj težav pri preklopih z visokimi frekvencami [14].

Glede na to, da se v zadnjem času povečujejo napetostne zmogljivosti baterij in enosmernega vmesnega tokokroga in bi uporaba MOSFET stikal lahko posledično omogočila zvišanje stikalnih frekvenc ter sočasno znižanje izgub, v članku predstavljamo realizacijo eksperimentalnega modela aktivne različice trinivojskega pretvornika z MOSFET-i za eno fazo s paralelnim delovanjem. Predstavljeni so posamezni sklopi modela, kot tudi napake, ki se lahko pojavijo pri nepazljivem načrtovanju. Dodatno so predstavljeni rezultati meritev in ovrednotenje delovanja pretvornika, kot tudi predlogi za morebitne izboljšave in razširitve.

2 Metodologija

Osrednji namen je bila izdelava demonstracijskega pretvorniškega sistema, prikazanega na sliki 1.

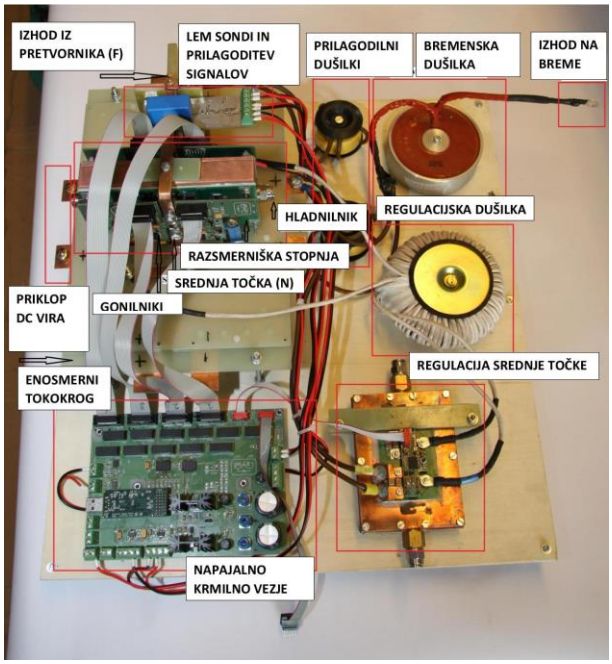


Slika 1. Shematski prikaz močnostnega dela pretvorniškega sistema od napajanja do izhoda s posameznimi stopnjami.

Za napajanje pretvornika je bil uporabljen laboratorijski napajalnik, v primeru elektromotorskega pogona pa bi to bila baterija. Sledi polmostič z dušilko za regulacijo srednje točke N, kjer imamo vzporedno vezavo

tranzistorskih vej za doseganje zahtevane tokovne zmogljivosti tranzistorjev. Nadalje je vključen enosmerni vmesni tokokrog z osrednjim namenom glajenja napetosti in vzpostavitve središčnega potenciala N s kapacitivnim delilnikom.

Na izhodu sledita dve aktivni NPC razsmerniški stopnji, ki delujeta v vzporedni vezavi prek gladilnih dušilk zaradi doseganja potrebne tokovne zmogljivosti.



Slika 2. Fotografija celotnega sistema z opisanimi posameznimi stopnjami.

2.1 Izdelava močnostnega dela

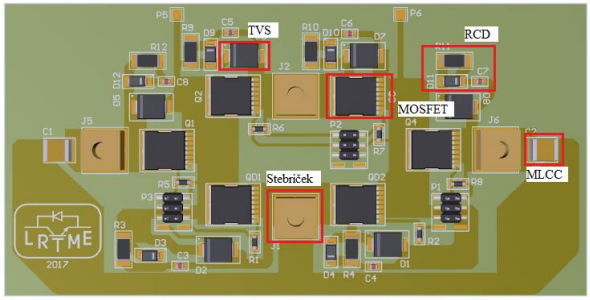
Močnostni del, prikazan na sliki 3, je sestavljen iz dveh vzporedno delujočih polovic, kjer je posamezna močnostna stopnja sestavljena iz šestih tranzistorjev IPT059N15N3 proizvajalca Infineon, h katerim sta vzporedno dodana hitra TVS dioda SMCJ85CA za razbremenitev prenapetosti ter razbremenilno vezje RCD z izračunanimi vrednostmi R = 6,75 Ω in C = 5,8 nF. Posamezna močnostna stopnja je izdelana na aluminijevem IMS substratu za boljše odvajanje toplote. Substrat ima 105 µm debeli nanos bakra. Elementi so po tiskanini razvrščeni tako, da je parazitna induktivnost zaradi dolžine linij čim manjša [15].

Ime	L [nH]	Pozicija	t [mm]	w [cm]	l [cm]
L_1	0,45	QD1 in Stebriček	0,105	0,90	0,30
L_2	0,89	QD1 in Q2	0,105	3,80	0,30
L_3	35,05	MLCC in stebriček	0,105	4,00	6,80
L_4	0,43	J5 in Q1	0,105	1,07	0,30

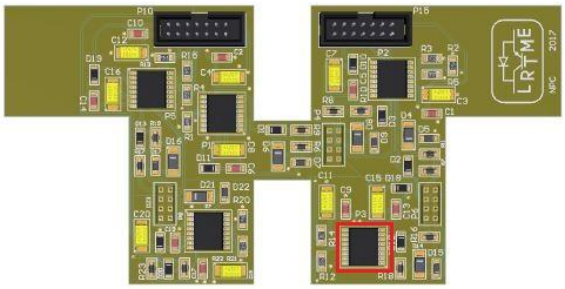
Tabela 1. Izračunane parazitne induktivnosti za posamezne linije med elementi, prikazanimi na sliki 4, kjer je t debelina, w širina in l dolžina bakrene linije.

2.2 Izdelava gonilniških tiskanin

Gonilniški tiskanini (slika 4) sta izdelani tako, da povsem nalegata na močnostni tiskanini za doseganje čim krajših povezav, s tem pa je dosežen učinek zmanjšanja parazitnih induktivnosti. Primer izračuna parazitnih induktivnosti (tabela 1) je opisan v literaturi [15, 16]. Uporabljen je gonilnik IRS2110, ki je v osnovi namenjen proženju polmostične topologije, tu pa je predelan za proženje tranzistorjev trinivojske topologije z dodatkom t.i. (ang. bootstrap) kondenzatorja, ki omogoča proženje vseh tranzistorjev, ki so na ločenih potencialih. Kapacitivnost je določena na minimalno vrednost 9,5 nF, za izračun pa je potrebno imeti podatke o gonilniku, tranzistorju ter stikalni frekvenci [15].



Slika 3. Močnostni del na IMS substratu z opisom posameznih komponent. Razdalje med posameznimi tranzistorji so enake, da je bila dosežena simetričnost tiskanine.



Slika 4. Shema gonilniške tiskanine, ki nalega neposredno na močnostni del. Njena oblika je prilagojena zaradi tokovnih stebričkov (priključnih kontaktov) močnostnega dela.

2.3 Izdelava krmilnega vezja

Krmilna tiskanina, ki je razvidna iz slike 2, vključuje mikrokrmilnik Piccolo proizvajalca Texas Instruments, kot tudi priključne kontakte za povezovanje in komunikacijo ostalih tiskanin z mikrokrmilnikom. Na tiskanini se za galvansko ločitev prožilnih signalov nahajajo tudi DC/DC napajalniki TMA 1515S proizvajalca TRACO Power. Dodani so tako merilni upori za merjenje napetosti enosmernega tokokroga, kot tudi vezja za prilagoditev merjene napetosti na območje A/D pretvornika.

2.4 Enosmerni vmesni tokokrog

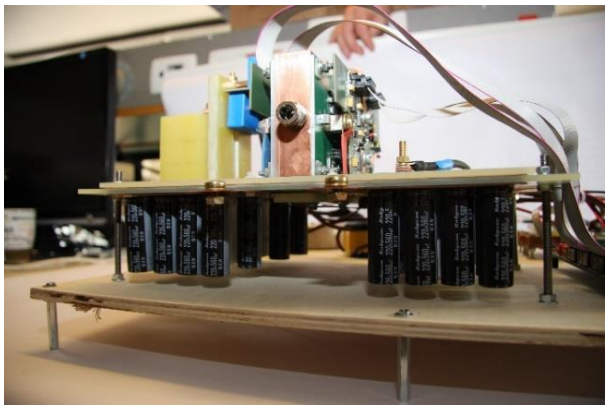
Enosmerni vmesni tokokrog (slika 5) sestavljajo elektrolitski ter folijski kondenzatorji v skupni velikosti 8,4 mF. Doseganje čim bolj simetričnih polovic je bila pri konstruiranju prioriteta, saj je s kapacitivnim delilnikom definirana srednja točka N. Ponovno je bil cilj

doseči čim krajši priklop do močnostnih tiskanin, da se je zmanjšal vpliv parazitnih induktivnosti.

2.5 Konstruiranje dušilk

Za blaženje tokovnih nesimetrij med posameznima stopnjama sta naviti dve manjši dušilki v velikosti 4,7 µH, ki sta vezani na izhod posamezne stopnje.

Za regulacijo srednje točke pretvornika je izdelana regulacijska dušilka z induktivnostjo 104,9 µH. Njena vrednost je bila izračunana iz predvidene 20 % valovitosti toka, stikalne periode in iz podatkov o jedru (T400-30D), ki jih podaja proizvajalec MICRO-METALS v podatkovnem listu. Navitje je izdelano iz dvojne žice za doseganje tokovne zmogljivosti dušilke in tako prenese vsaj 100 A tokovne obremenitve.



Slika 5. Fotografija enosmernega vmesnega tokokroga, kjer sta razvidni dve simetrični polovici ter priključitev na močnostni del.

2.6 LEM tiskanini

Za merjenje toka ter regulacijo obratovalnih stanj pretvornika, se je pojavila potreba po dodatni tiskanini z LEM sondami, na kateri se nahaja še prilagoditveno analogno vezje, ki izmerjeno napetost LEM sonde prilagaja napetostnemu območju A/D pretvornika v mikrokrmilniku.

2.7 Regulacija napetosti enosmernega tokokroga

Ugotovljeno je bilo, da se že pri majhnih obremenitvah pretvornika napetost srednje točke seseda, zato se je pokazala potreba po regulaciji srednje točke. V ta namen so dodani dva podvojena tranzistorja za doseganje tokovne zmogljivosti ter regulacijska dušilka, katere vrednosti so opisane v poglavju 2.5. Pri tem gre za izmenjavo energije med enosmernim vmesnim tokokrogom ter dušilko.

2.8 Programska oprema, preverjanje in zagon pretvornika

Krmilni program za pretvornik je pripravljen na mikrokrmilniku Piccolo proizvajalca Texas Instruments. Program omogoča izbiro delovanja pretvornika, torej tako aktivni kot pasivni režim. Omogoča vključitev ali izključitev regulacije srednje točke, kot tudi nastavitev osnovne in stikalne frekvence, pri čemer je slednja pogojena z izvajanjem prekinitvenega podprograma.

Pretvorniški sistem na sliki 2 je bil postopoma zagnan ter testiran po sklopih. Najprej se je preverilo pravilnost prožilnih signalov, izmerjene so bile morebitne prenapetosti na tranzistorjih, prav tako so bile testirane posamezne tiskanine. Sledile so preliminarne meritve obratovanja pri nižji obremenitvi.

3 Preliminarne meritve

Namen preliminarnih meritev je testiranje posameznih sklopov pretvornika, ovrednotenje slednjih ter testiranje krmilnega programa. Nato je sledil zagon pretvornika v pasivnem in aktivnem režimu delovanja, kjer je bilo preverjeno ali pretvornik deluje v skladu z rezultati simulacij.

3.1 Testiranje pretvornika

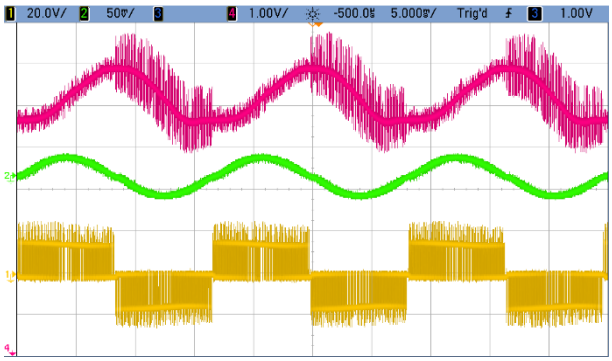
Najprej je bil izveden test pretvornika v pasivnem režimu delovanja, kjer je bilo opazovano dogajanje ob nereguliranem ničlišču, katerega valovitost je prikazana na sliki 6 z vijolično barvo. Opaženo je bilo tako sesedanje srednje točke, kot tudi izhodne napetosti. Slednje je postalo vidno pri obremenitvi pretvornika z RL bremenom, kjer je R znašal 3.3 Ω in L = 135 mH. Pri tej obremenitvi je napetost enosmernega vmesnega tokokroga znašala 30 V, izhodni tok pa 9,2 A.

Pri vključeni regulaciji srednje točke se je valovitost napetosti znižala z 1,5 V na 1 V. Opaziti je bilo izravnavanje valovitosti ničlišča in izhodne napetosti, kar je prikazano na sliki 7 z oscilogramom vijolične barve.

Pri delovanju pretvornika v pasivnem režimu je bilo opaziti padec napetosti na ločilnih diodah. Slednji je bil odpravljen pri aktivni različici delovanja pretvornika, kar je bil hkrati indikator, da tudi aktivna različica deluje pravilno. Padec napetosti na diodah vpliva na izkoristek pretvornika pri velikih tokovnih obremenitvah, saj se povečajo prevodne izgube.

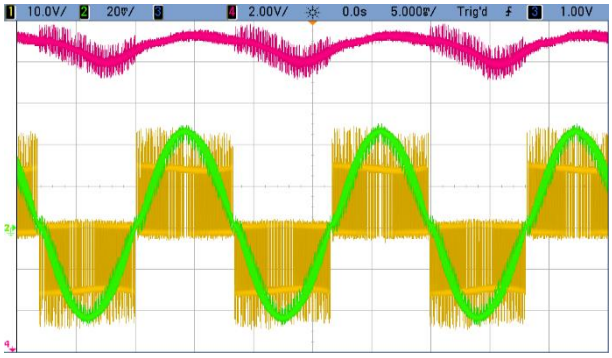
3.2 Omejitve izdelanega sistema ter predlogi za nadaljnje delo

Celoten sistem deluje sicer pravilno, vendar ne najbolje, saj je bilo ugotovljenih nekaj napak pri izdelavi gonilniške tiskanine. Glavna napaka je bila storjena ob predelavi IRS2110 za krmiljenje samo enega tranzistorja, ki se nahaja na ločenem potencialu. Napaka je bila odpravljena s prevezavami, kar pa se je izkazalo kot neoptimalna rešitev, saj so se zaradi dolgih priključnih linij pričele pojavljati prenapetosti, ki so povzročale nenadne vklope tranzistorjev, šum (sliki 6 in 7) ter odpoved nekaterih čipov.



Slika 6. Poteki izhodne napetosti (oker), izhodnega toka (zelena) in valovitosti napetosti ničlišča (vijolična), zaradi katere se seseda izhodna napetost.

V nadaljevanju je najprej potrebno odpraviti napako, ki je bila narejena pri načrtovanju gonilniške tiskanine in slednjo ponovno načrtati, pri tem pa paziti na pozicijo gonilniških čipov tako, da bo povezava med izhodom čipa in vrati tranzistorja še vedno čim krajša. Zanimiva je uporaba nove generacije gonilnikov, ki že merijo napetosti na tranzistorju, s čimer je zmanjšana verjetnost odpovedi.



Slika 7. Prikazan je potek izhodne napetosti (oker barva), izhodnega toka (zelena barva) in izravnana napetost ničlišča (vijolična barva), kar vpliva na zmanjšanje sesedanja izhodne napetosti (poravnani vrhovi oker signala).

Sledi lahko ovrednotenje pretvornika po moči, meritev izkoristkov pri različnih obratovalnih stanjih ter primerjava z rezultati simulacij iz literature [15, 16], kot tudi primerjava z dvonivojsko topologijo. Glede na rezultate testiranja in meritev lahko sledi razširitev na tri faze, kar bo odpravilo potrebo po regulaciji srednje točke in omogočilo realizacijo klasičnega pogona s trifaznim motorjem. Takšen pogonski sklop lahko razširimo na generatorski režim delovanja oziroma na dvosmerni pretok energije. Možno je testiranje pretvornika pri polni moči, ocena izkoristka ter primerjava z dvonivojsko topologijo.

4 Sklepne ugotovitve

V sklopu študije je bila izvedena ena faza trinivojskega pretvornika s po dvema paralelnima vejama za povečanje skupne tokovne zmogljivosti. Predstavljeni so ukrepi pri načrtovanju pretvornika ter pomanjkljivosti, ki so se pokazale med testiranjem. Rezultati meritev kažejo, da realizirani pretvornik deluje zadovoljivo, vseeno pa bi bilo v nadaljevanju smiselno narediti nekaj izboljšav.

Izdelani pretvornik predstavlja dobro izhodišče za preverjanje različnih konceptov vodenja, hkrati pa lahko služi za laboratorijske meritve zaradi kompaktne izdelave, saj so z njo izločene možne motnje, ki bi vplivale na meritve. Ravno zaradi kompaktnosti bo mogoče izvesti primerjavo rezultatov meritev z rezultati predstavljenih numeričnih simulacij [15, 16].

Literatura

[1] https://www.forbes.com/sites/rrapier/2017/02/05/u-s-electric-vehicle-sales-soared-in-2016/#76e559f4217f, 3.7.2018.

[2] https://cleantechnica.com/2015/03/28/ev-demand-growing-global-market-hits-740000-units/, 3.7.2018.

[3] M. Schweizer, J. W. Kolar: Design and implementation of a highly efficient three-level T-type converter for low-voltage applications. *IEEE Transactions on Power Electronics*, *28*(2), 899-907, 2013.

[4] L. G. Franquelo, J. Rodriguez, J. I. Leon, S. Kouro, R. Portillo, M. A. Prats: The age of multilevel converters arrives. *IEEE industrial electronics magazine*, *2*(2), 2008.

[5] A. Salem, M. F. Elsied, J. Druant, F. De Belie, A. Oukaour, H. Gualous, J. Melkebeek: An advanced multilevel converter topology with reduced switching elements. V: *Industrial Electronics Society, IECON 2014-40th Annual Conference of the IEEE* (pp. 1201-1207), 2014.

[6] S. Brueske, B. Benkendorff, R. Kulpe, F. W. Fuchs: Comparison of the power semiconductor design rating of different inverter topologies for the drive inverter of electric vehicles. V: *Energy Conversion Congress and Exposition (ECCE)*, (pp. 3643-3650) IEEE, 2015.

[7] S. Busquets-Monge, R. Maheshwari, J. Nicolas-Apruzzese, E. Lupon, S. Munk-Nielsen, J. Bordonau: Enhanced DC-link capacitor voltage balancing control of DC–AC multilevel multileg converters. *IEEE Transactions on Industrial Electronics*, *62*(5), 2663-2672, 2015.

[8] S. Bhattacharya, D. Mascarella, G. Joós: Modular multilevel inverter: A study for automotive applications. V: *Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on* (pp. 1-6), IEEE, 2013.

[9] J. Biela, M. Schweizer, S. Waffler, J. W. Kolar: SiC versus Si—Evaluation of potentials for performance improvement of inverter and DC–DC converter systems by SiC power semiconductors. *IEEE transactions on industrial electronics*, *58*(7), 2872-2882, 2011.

[10] T. Morita, S. Tamura, Y. Anda, M. Ishida, Y. Uemoto, T. Ueda, D. Ueda: 99.3% efficiency of three-phase inverter for motor drive using GaN-based gate injection transistors. V: *Applied Power Electronics Conference and Exposition (APEC), 2011 Twenty-Sixth Annual IEEE* (pp. 481-484), 2011.

[11] B. A. Welchko, M. B. Corrêa, T. A. Lipo: A Three-Level MOSFET Inverter for Low-Power Drives. *IEEE Transactions on Industrial Electronics*, 51(3), 669-674, 2004.

[12] A. Anthon, Z. Zhang, M. A. E. Andersen, D. G. Holmes, B. McGrath, C. A. Teixeira: Comparative evaluation of the loss and thermal performance of advanced three-level inverter topologies. *IEEE Transactions on Industry Applications*, 53(2), 1381-1389, 2017.

[13] A. Alešnik: Trinivojski razsmernik. *Diplomsko delo*. Fakulteta za elektrotehniko, Univerza v Ljubljani, 2013.

[14] E. Gurpinar, A. Castellazzi: Single-Phase T-type inverter performance benchmark using Si IGBT's, SiC MOSFET's and GaN HEMT's. *IEEE Transactions on Power Electronics*, 31(10), 7148-7160, 2016.

[15] B. Prah: Simulacijski in eksperimentalni model trinivojskega razsmernika. *Magistrsko delo*. Fakulteta za elektrotehniko, Univerza v Ljubljani, 2018.

[16] B. Prah, A. Rihar, D. Vončina: Simulacije in konceptualna zasnova trinivojskega pretvornika. V *Zborniku 26. konference ERK*, str. 291-294, Sep. 2017, IEEE.

Simulation and Analysis of Solar PV modules under shadow effect

Nedzmija Demirovic[1], Alma Huremovic[2], Izudin Softic[1]

1Faculty of Electrical Engineering Tuzla
[2] Border Police of Bosnia and Herzegovina
E-pošta: nedzmija.demirovic@gmail.com

Abstract. *The performance of solar photovoltaic (PV) arrays depends on exposure to solar irradiation and on variations of external temperature. In many cases, the photovoltaic array works precisely in conditions when it is not fully exposed to solar irradiation, precisely because of the effect of the shadows which may be due to cloudy weather or the existence of a shadow object. The operating point of the solar PV arrays corresponds to the points of maximum power. Finding this point can be time-consuming in situations where a string works in uneven working conditions, such as the presence of shadows or changes in the outside temperature.*
In this paper the performance of a solar PV array when it is exposed to the effects of unbalanced solar irradiation will be presented. It will be shown that the reduction in the external temperature affects the output values of the solar PV array.

1 Introduction

To get the required value of the output voltage, photovoltaic (PV) cells are connected in series. In this way module is created. If it is necessary to get a higher output voltage than those it can be obtained from one module, the modules are connected into arrays. Photovoltaic arrays are often exposed to the shadow effect often due to very close neighbouring objects [1].
The most important characteristic of the PV arrays is the I-V characteristic. It is an indication of how much power can be obtained from the shadowed PV arrays. If the string is exposed to uniform solar irradiation, it is easy to find working point of maximum power.

In the case of partial shadows, the maximum power point is located taking into account the value of the actual voltage on the module or array. Accordingly, the I-V characteristic of the PV cell is under the influence of two parameters: solar irradiation and outside temperature [2]. To make the photovoltaic modul work with maximum power and accordingly have the best efficiency, it is necessary to have an appropriate method that will regulate the voltage in real time. In addition, the outside temperature has a significant effect on the I-V characteristic of the PV cell [3],[4].

2 Basic components of the PV model

The Matlab / Simulink-based block allows the modeling of PV modules, with its own parameters, as well as the choice of ready-made models [2]. For the PV modeling, five parameters are required: photo- current I_L, current of saturation of the diode I_0, the ideal factor of diode nI, serial R_S resistance and parallel resistance R_{sh} taking into account the irradiation and the temperature dependence of the I-V characteristic.

An example of equivalent diagram of PV cell with one diode is shown (Figure 1). The I-V and P-V characteristics for a single module are given by the following terms (Figure 2):

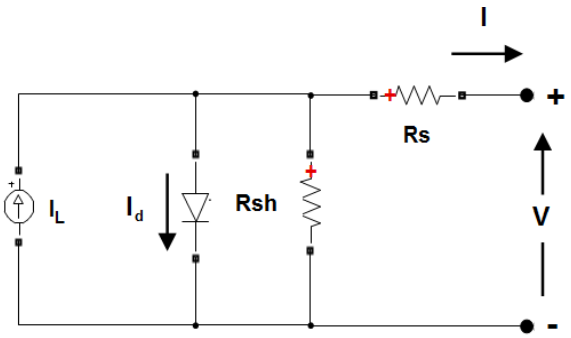


Figure 1. One diode equivalent diagram of PV cell

$$I_d = I_0\left(e^{\left(\frac{V_d}{V_T}\right)} - 1\right) \quad (1)$$

$$V_T = \frac{k \cdot T}{q} \cdot nI \cdot N_{cell} \quad (2)$$

where are: V_d - voltage diode, V_T - thermal voltage, N_{cell}- number of cells, k- Boltzman's constant.

At the output terminal from each PV module, it is possible to measure the voltage and current, based on obtained I-V and P-V characteristics. Each module has 20 serially connected cells which are antiparallel connected with bypass diodes. In the simulation and analysis 6 such modules were used.

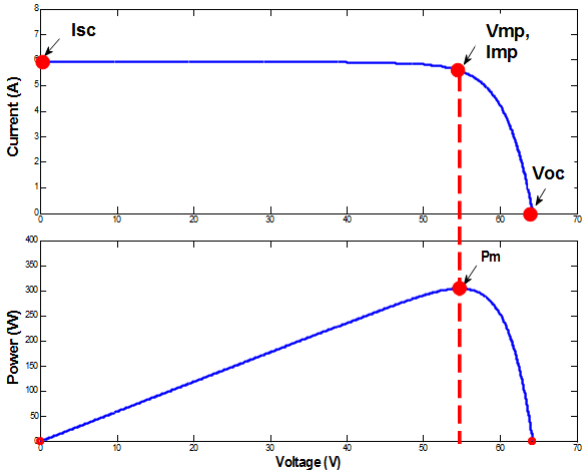


Figure 2. I-V and P-V characteristics of a single module

The temperature coefficient of voltage presents the dependence of the voltage on the temperature given by the term:

$$V_{0CT} = V_{0C}(1 + \beta_{V_0C}(T - 25)) \quad (3)$$

where are: V_{0C} open voltage circuit at temperature 25 (°C), β_V_{0C} is the temperature coefficient (in %/°C), and T is the temperature in (°C).
Temperature current coefficient presents dependence of current on the temperature given by the term:

$$I_{SCT} = I_{SC}(1 + \alpha_{ISC}(T - 25)) \quad (4)$$

where are I_{SC} short circuit current at temperature (25°) C, I_{SCT} is short circuit current at temperature T, α_I_{SC} is the temperature coefficient (in %/°C) and T is the temperature in (°C).

The maximum power of the PV system is obtained at 1000 (W/m^2) irradiation, and the ambient temperature is 25 (°C).

2.1 Role and model of bypass diode

The light intensity on each cell affects the operation of the entire module. Namely, if any cell is in the shadow due to weather conditions eg. or, the existence of a shadow object, then the output values of the module are reduced. Since cells in a single module are connected in series, only one weaker cell brightness leads to moving the point of maximum power and overheating cells. In order to avoid losses, the bypass diode connects antiparallel to the each PV cells.

In normal circumstance, the positive voltage on it acts as blocking, and the bypass diode does not work. As soon as the PV cell finds itself in shadow, it becomes bridged due to the bypass diode being processed. Overheating occurs when other non-shadowed cells try to "push" electricity through a shadow cell [5]. In this paper, the bypass diode is modeled with a resistor, inductance and a DC source connected in a series with a switch. The operation of the switch is controlled by the voltage V_{ak} and the current I_{ak} (Figure 3).

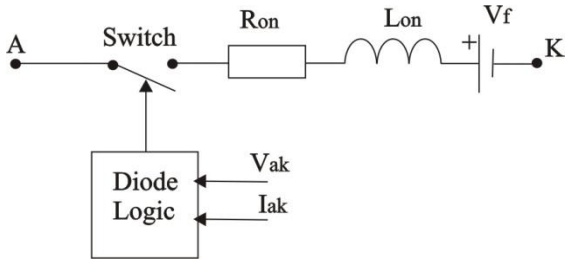


Figure 3. Model of bypass diode

Resistance R_{on} represents the internal resistance of the diodes in (Ω); the L_{on} is inductance; V_f is the control voltage of the diode. The starting current of the diode I_0 is set to zero so the simulation starts when the diode is blocked. This model does not take into account the geometry of the device or the complex physical processes that occur during the state change of the diodes.

3 Simulation and analysis of the results

Partial shading of a 250-W PV module consisting of 60 cells connected in series (Figure 4).

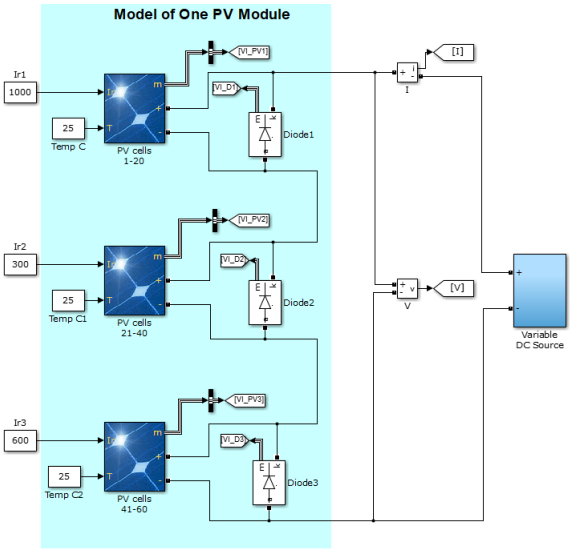


Figure 4. Model of PV partial shading module

In this paper the following cases are analyzed:

Case 1: Number of parallel strings =1, number series connected modules per string =1. Values of irradiation per modules respectively:
I_{r1}=1000 (W/m^2); I_{r2}=300 (W/m^2); I_{r3}=600 (W/m^2).

At initial moment in the case normal irradiation of 1000 (W/m^2), the voltage on the first module V_PV1 = 8.6 (V). The voltage at the bypass diode D1 is equal to zero because the diode does not lead (Figure 5). In the

same time, the voltage on the second module is V_PV2 = 2.6 (V), the voltage of the bypass diodes is 6 V and gradually decreases.

In time t = 0.81 (sec), voltage of first module dropped at V_PV1= 5.4 (V); voltage of bypass diode is decreased at 2.62 (V), and it is equal to zero in t=1.84 (sec) In t = 1.84 (sec), voltage at V_PV1 = 2.6 (V).

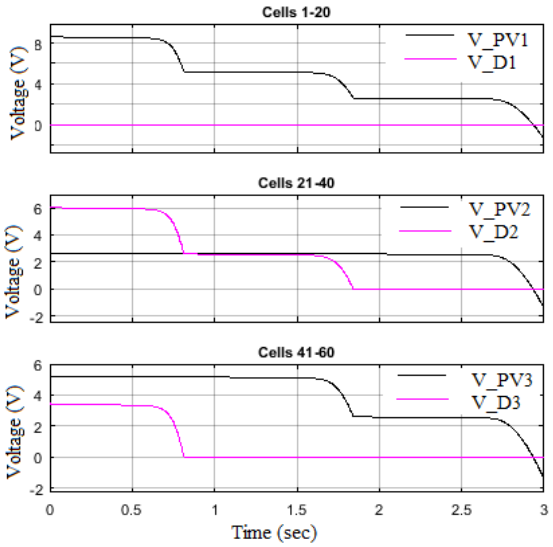


Figure 5. Voltages of PV modules and bypass diodes

The voltage on the third module, which is better illuminated than the second but less than the first, is V_PV3 = 5.18 (V) in initial moment. Voltage of bypass diode is 3.4 (V) and gradually decreaces to zero. So, whenever the illumination is smaller, the diodes are activated. The smaller the illumination, the higher the voltage on the diode. For example, the voltage of the bypass diode of the module 2 is 6.04 (V), and the voltage of the bypass diode of the module 3 is 3.15 (V). The PV module's currents have also a similar shape (Figure 6). The total current of the circle is presented by the red line as well as current of the first module.

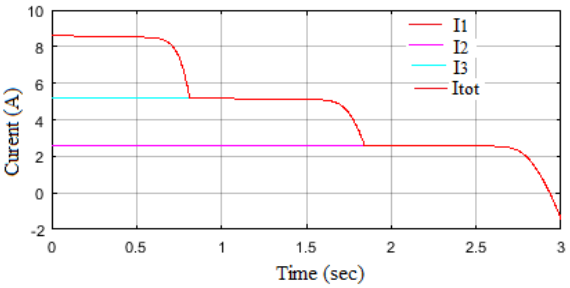


Figure 6. Currents per modules and total curent of all modules

The maximum power is 104 (W) and the current corresponding to this value is 5 A (Figure 7).

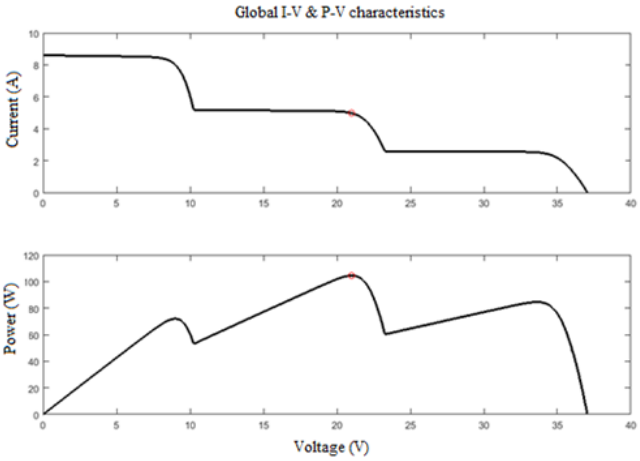


Figure 7. I-V and P-V characterstics of PV modules in the case of different irradiation per modules

Case 2: Number of parallel strings =1, number series connected modules per string =1. Values of irradiation per modules respectively:
I_{r1}=1000 (W/m^2); I_{r2}=300 (W/m^2) ;I_{r3}=300 (W/m^2).

The total power of all modules is 79.7 (W), therefore less than the nominal power of one module. This is due to the less power of the other two modules (Figure 8).

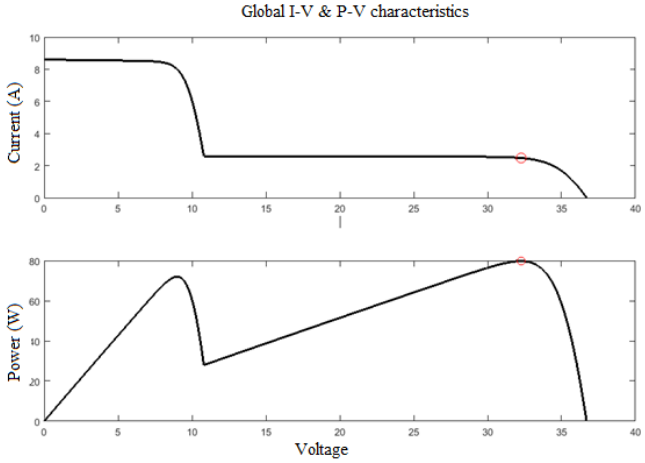


Figure 8. I-V and P-Vcharacteristics in the case of decrease of irradiation at modules 2 and 3

Case 3: The operating temperature of the modules respectively: T_1=25 (°C), T_2=15 (°C), T_3=25(°C). Values of irradiation per modules respectively:
I_{r1}=1000W/m^2; I_{r2}=1000W/m^2;I_{r3}=1000 (W/m^2).

Maximum power of modules is 249.5 (W) which could be expected (Figure 9).

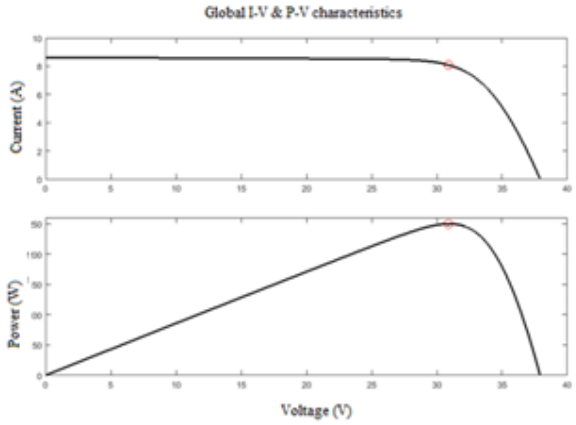


Figure 9. I-V and P-V characterstics of PV modules in the case of uniform irradiation per modules

Comparison of total currents of modules under different outdoor temperature is given (Figure 10). The blue line presents the current for which the system achieves maximum power in the case irradiation of 1000 (W/m^2) on all modules and the same outdoor temperature of 25 (°C). A dark red line corresponds to equal irradiation on all modules and the outdoor temperature of third module of 15 (°C).

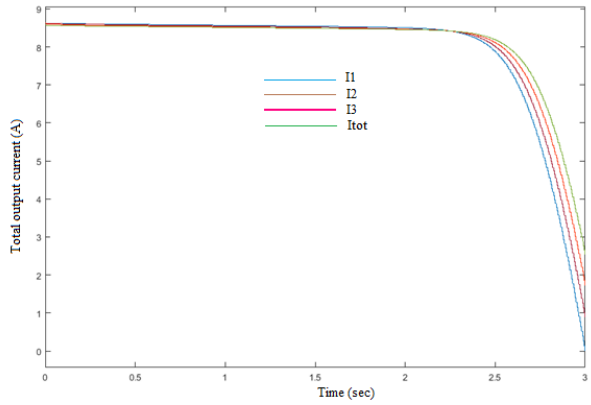


Figure 10. Current per modules in the case 3 of simulation

In the next simulation, the irradiation was also kept at a maximum value and the reduced temperature of the second and third modules at 15 (°C) (red line). Finally, the temperature of all three modules at irradiation of 1000 (W/m^2) was reduced (green line). In addition to this, the comparison of currents per modules is presented when the irradiation is as in case 1 and the outdoor temperature changes from 25 (°C) to 15 (°C) (Figure 11).

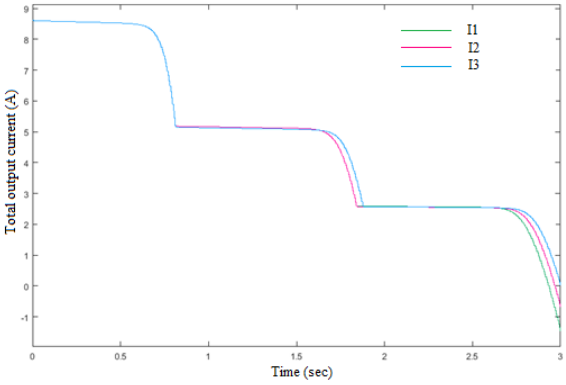


Figure 11. Comparison of currents in the case of different outdoor temperatures at modules

Green line: irradiation is 1000 (W/m^2), 300 (W/m^2), 600 (W/m^2), outdoor temperatures for all three modules are 25 (°C).

Red line: the same irradiation as previous, outdoor temperature of module 2 is 15 (°C).

Blue line: the same irradiation as previous, outdoor temperature of modules 2 and 3 is 15 (°C).

Conclusion

Partial shading of a 250-W PV module consisting of 60 cells connected in series are presented. It has been shown that power reduction is mostly influenced by the brightest lighted PV module. Thus, in the case1, the smallest current and voltage was in module 2. The global I-V characteristic is determined by the weakest module, so the output power is reduced. When the outside temperature decreases, there is an increase in current through the modules, which implies that there is overheating. The maximum power point moves slightly to the right.

Literature

[1] S. Sumathi, L. Ashok Kumar, P. Surekha: Solar PV and Wind Energy Conversion Systems An Introduction to Theory, Modeling with MATLAB/SIMULINK, and the Role of Soft Computing Techniques, Springer International Publishing Switzerland 2015

[2] Xuan H. Nguyen: Matlab/Simulink Based Modeling to Study Effect of Partial Shadow on Solar Photovoltaic Array 2015

[3] Matlab/Simulink User Guide

[4] H. Mohamed Diab, Maximum Power Point Tracking of Grid Connected Photovoltaic Using Artificial Neural Networks, Master Thesis 2012

[5] C. Salvadores, J.Francisco: Shadowing Effect on The Performance in Solar PV Cells, Master Thesis 2015

Comaparative Anaylisis of Maximum Power Point Tracking Algorithms of PV Array with Three-level NPC Power Converter with Selectable Topologies and Power Switching Device

Nedzmija Demirovic[1], IzudinSoftic[1]

[1]Universisty of Tuzla Faculty of Electrical Engineering
E-pošta: nedzmija.demirovic@gmail.com

Abstract. *The efficiency of the photovoltaic system is based on the efficiency of each individual component as well as the installation way of the photovoltaic system. Each photovoltaic (PV) system has an inverter that converts the DC output voltage of the PV system to the AC voltage of a network with a certain degree of efficiency. The conversion efficiency depends on the speed and precision of finding the maximum power point. This paper provides a comparative performance analysis of a 100 kW photovoltaic system connected to the 25-kV network via a DC-DC boost converter and a three-phase three level Voltage Source Converter (VSC) by two modes of finding the maximum power point. In addition, the analysis of the overall harmonic distortion of the output power of the photovoltaic system and the line voltage behind the three level bridge for the two modes of finding the maximum power point are also given.*

1 Introduction

The photovoltaic (PV) system consists of a variaty of individual components that are serially connected and in parallel by making arrays and modules in order to obtain the appropriate output voltage and current, and thus the output power.

The maximum power point (MPP) depends on the current solar irrradiation and the outside temperature. These parameters determine the output voltage of the PV array. Maximum power is always determined by the optimal voltage value [1]. In this purpose the PV system has a DC-DC converter and a DC-AC converter. The role of DC-DC converter is to regulate the output voltage of the PV array by changing the duty cycle of the converter. A very important component of the PV system is a DC-AC inverter that converts DC voltage to the appropriate AC voltage. Conversion efficiency depends on the speed and precision of finding the maximum power point [2]. As the DC-DC converter changes the duty cycle and thus the output voltage, the DC-AC converter fixes this DC voltage and maintains it constant and converts to the AC voltage.

2 The role of power converters inside of PV system

2.1 Voltage booster

The DC-DC converter has the role of a voltage booster whose frequency is 5 kHz. This converter raises the input voltage of 273 V to 500V at the maximum power. The function of converting the output voltage (V_{out}) to the input voltage (V_{in}) is given by the formula:

$$\frac{V_{out}}{V_{in}} = \frac{1}{1-D} \tag{1}$$

where D is duty cycle.

The interrupt interval or duty cycle is optimized by the maximum power point controller based on the Incremental Conductance (IC) and the Perturb & Observe (P&O) algorithms. By using these techniques, the maximum power point tracker (MPPT) system automatically changes the duty factor to generate the required voltage needed to gain maximum power.

2.2 Three phase voltage regulator VSC

The three-phase voltage VSC consists of a measuring unit, a current and voltage regulator, an AC voltage generating unit, and three level PWM (pulse width modulation) generator. The VSC three-phase voltage controller converts the voltage from 500 V DC to 260 V AC, while retaining the unit power factor. VSC has two regulating loops. In the outside regulating loop DC voltage is controlled at +/- 250 V, while in the inside regulating loop active and reactive components of network currents I_d and I_q are regulated. Reference value I_d is the output value from an external voltage controller. The output voltages of the current controller V_d and V_q are converted into three modulated U_{abc_ref} signals used by the PWM generator. The scheme of VSC regulator is realized in Matlab/ Simulink (Figure 1).

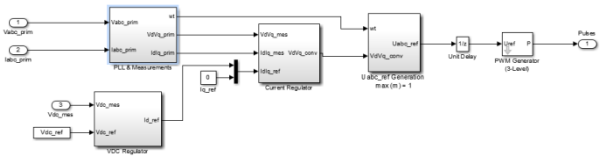


Figure 1 Scheme of the VSC regulator

The Three-Level Bridge block implements a three-level power converter that consists of one, two, or three arms of power switching devices (Figure 2). Each arm consists of four switching devices along with their antiparallel diodes and two neutral clamping diodes (NPC).

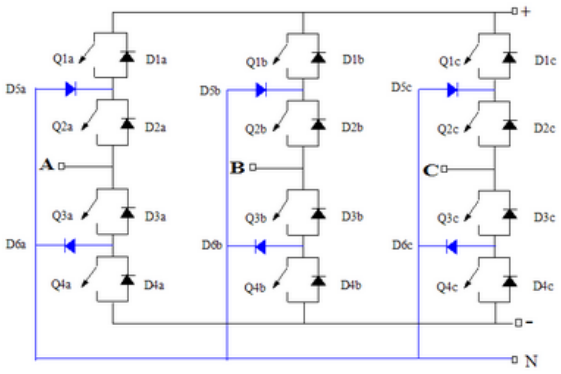


Figure 2 Scheme of the Three Level Bridge

3 Maximum power point tracking by using incremental conductance method

The Incremental Conductance (IC) method works under the assumption that the relationship between internal conductivity is equal to negative value of the output conductance [3],[4]. That means:

$$dI/dV = -I/V \tag{2}$$

where are:

dI, dV = fundamental components of I and V ripples measured with a sliding time window T_MPPT.

I , *V* = mean values of *V* and *I* measured with a sliding time window T_MPPT.

The MPPT regulates the PWM control signal of the DC-DC boost converter until the condition (3) is satisfied:

$$(\partial I/\partial V) + (I/V) = 0 \tag{3}$$

Scheme of MPPT by using Incremental Conductance algorithm is modeled and realized in Matlab/Simulink (Figure 3).

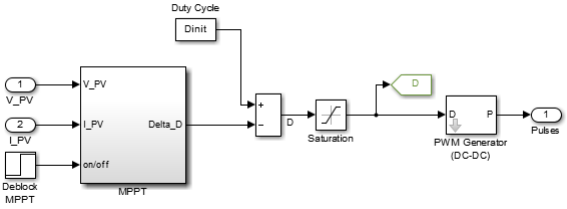


Figure 3 Simulink scheme of MPPT by using Incremental Conductance algorithm

4 Maximum power point tracking by using P&O algorithm

Algorithm Perturb and Observe (P&O) works in such way to determine current power, P_{old} first [3].

$$P_{old} = V \cdot I \tag{4}$$

After that duty cycle increases.

$$D = D + \Delta D \tag{5}$$

If the new measured power is greater than the old value ($P_{new} > P_{old}$) then the direction of maximum power point location is correct, and the duty factor can be furtherly increased. If this condition is not fulfilled then duty factor must be reduced (6), after what the amount of power is checked again.

$$D = D - \Delta D \tag{6}$$

MMPT scheme using P&O algorithm is modeled and realized in Matlab/Simulink (Figure 4).

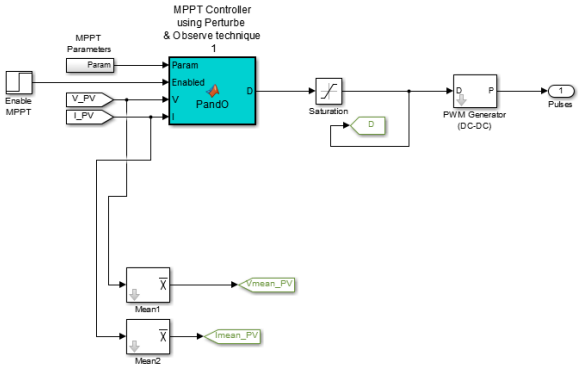


Figure 4 Simulink of MPPT scheme using P&O algorithm

5 System description

For the purpose of the simulation a 100-kW SunPower SPR-305E-WHT-D PV array was used. It consists of 5 serially-connected modules. Each of them has 66 parallel strings connected to a 25-kV grid via a DC-DC boost converter and a three-phase three-level Voltage Source Converter (VSC) (Figure 5). Maximum Power Point Tracker (MPPT) is implemented in the boost converter by means of a Simulink model using the IC and P&O algorithms. The capacitor bank filtering harmonics produced by VSC. I-V and P-V characteristics of PV array are preseneted (Figure 6).

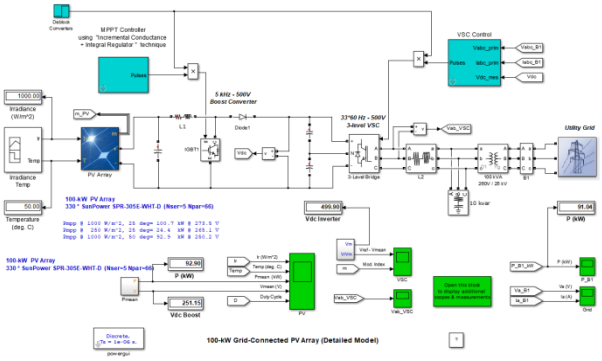


Figure 5 Simulink block scheme of PV array by using two different algorithm for MPPT

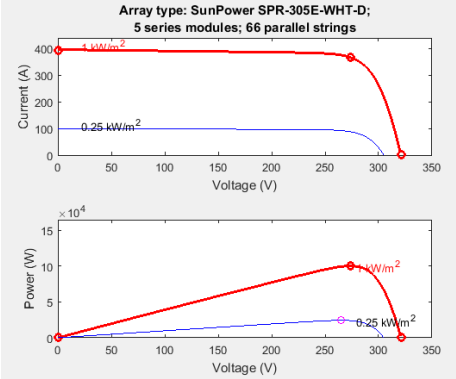


Figure 6 I-V and P-V characteristics of PV array at different radiations

6 Comparison of simulation results

Two cases are simulated:

1st case: simulation of PV array in case the maximum power point is found by using the incremental conductance algorithm.

2 nd case: simulation of PV array in case the maximum power point is found by using Perturb and Observe algorithm.

In the case of application P&O algorithm, duty cycle changes from 0.38 to 0.58.and cames on PWM DC-DC impulse generator of parallel IGBT1 diode. Start value of duty cycle is 0.5. The incremental value used to increase and decrease ΔD is 3x10-4. The output voltage of the PV array corresponds to the voltage of 321 (V). (N_{ser} ·V_{oc}=5·64.2=321 (V). The three-level bridge operates as a diode rectifier and DC link capacitors are charged above 500 (V). At the input of the PV array radiations in normal circumstances is 1000 (W/m^2) and the outdoor temperature is 25 (°C). The effect of the irradiation change on the output power of the PV system is simulated by variable irradiance and temperature (Figure 7).

The amount of output power of PV array, output voltage of PV array, duty cycle, active power being injected into the grid, modulation index, alternating voltage V_{ab} behind three-level bridge are observed and compared (Figure 8 to Figure 13). In addition, the FFT analysis of active power, and AC voltage V_{ab} behind the three-level bridge was performed. Comparison of total harmonic distortion is observed on two selected signals (Figure 14 , Figure 15).

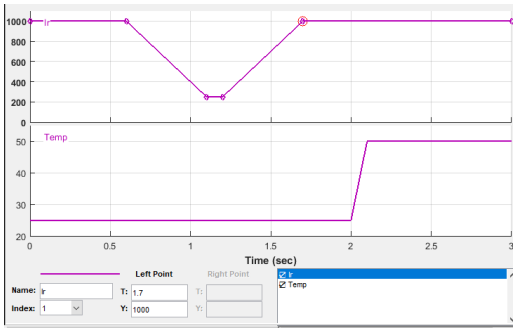
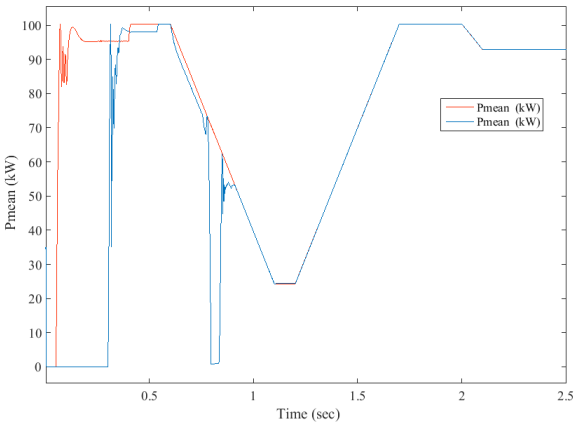


Figure 7 Changes of irradiation and outdoor temperature



8 Comparison of output power of PV array using Incremental conductance (red line) and P&O (blue line) algorithms MPPT

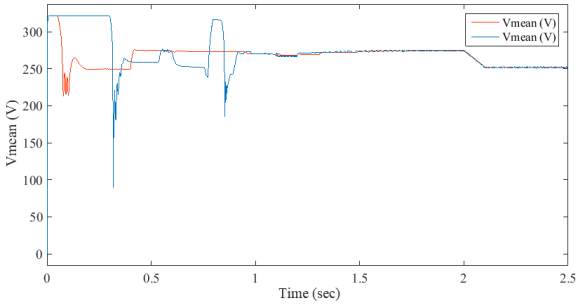


Figure 9 Comparison of output voltages of PV array using Incremental conductance (red line) and P&O (blue line) algorithms MPPT

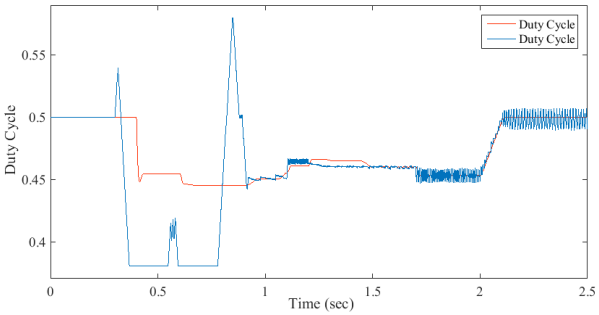


Figure 10 Comparison of duty cycles using Incremental conductance (red line) and P&O (blue line) algorithms MPPT

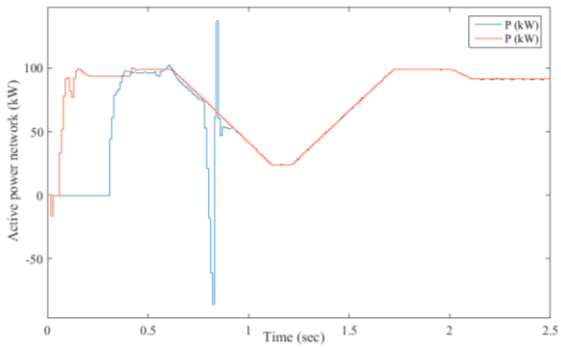


Figure 11 Comparison of the injected network active power using Incremental conductance (red line) and P&O (blue line) algorithms MPPT

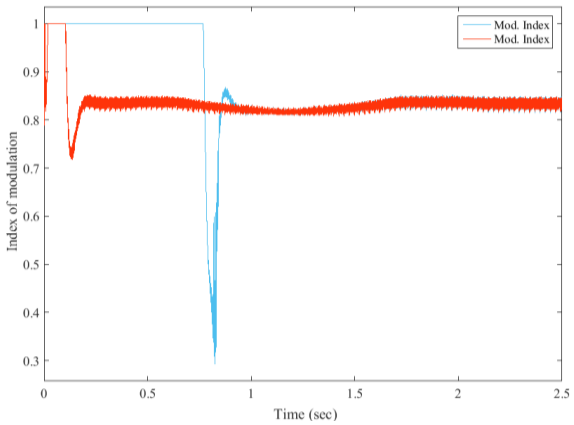


Figure 12 Comparison of indexes modulation using Incremental conductance (red line) and P&O (blue line) algorithms MPPT

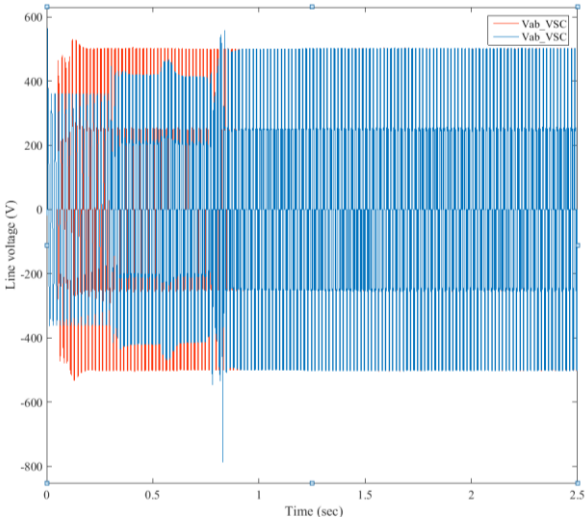


Figure 13 Comparison of AC line voltages using Incremental conductance (red line) and P&O (blue line) algorithms MPPT

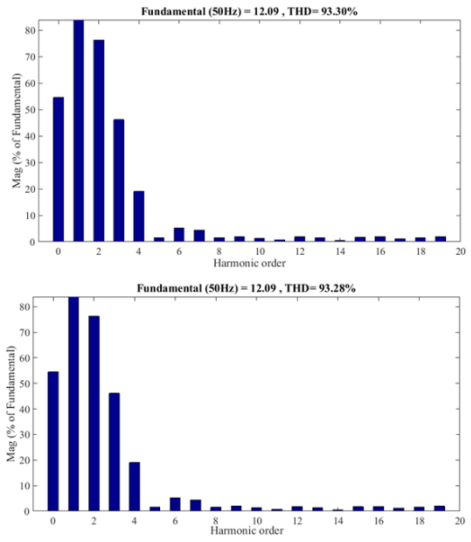


Figure 14 Comparison THD of active output power of PV array (up:Incremental Conductance algorithm, down: O&P algorithm)

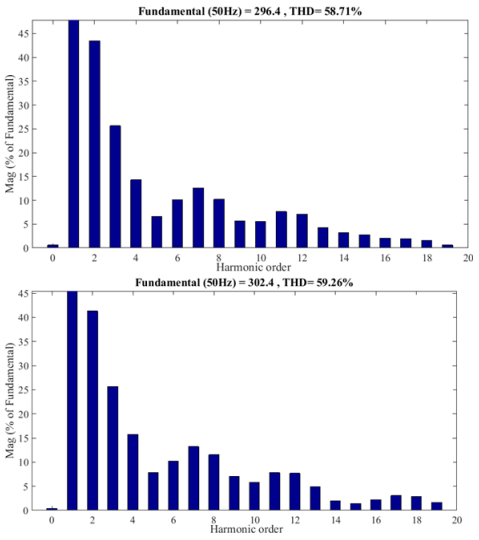


Figure 15 Comparison of THD line voltage V_{ab}_VSC (up: Incremental Conductance algorithm, down: O&P algorithm)

7 Conclusion

This paper analyzes the detailed model photovoltaic system of 100 kW connected to the distribution network. For the purpose of fast and precise finding the maximum power point, two different algorithms were used. The output values of the PV arrays were compared. There is a great influence of impulse generator on the output signals as well as VSC regulator control. The application of two different algorithms contributes to the different modes of generating the signal where the difference always exists in the starting point of the selected signals. The differences in the total output power, DC link voltage, the network power are negligible. The contribution to generating higher harmonics is slightly smaller when using the incremental conductance algorithm.

Literatura

[1] M.Lokanadham, K.Vijaya Bhaskarr: Incremental Conductance Based Maximum Power Point Tracking (MPPT) for Photovoltaic System, (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.1420-1424.

[2] S. Sumathi, L. Ashok Kumar, P. Surekha: Solar PV and Wind Energy Conversion Systems An Introduction to Theory, Modeling with Matlab/Simulink, and the Role of Soft Computing Techniques, Springer International Publishing Switzerland 2015.

[3] R.Ramachandran,P.S.Manoharan:Comparative Analysis of Various MPPT Techniques for Photovoltaic System with SEPIC Converter, EICC, IJTRD, 2016.

[4] M. Muthuramalingam1, P. S. Manoharan: Energy comparative analysis of MPPT techniques for PV system using interleaved soft-switching boost converter, World Journal of Modelling and Simulation Vol. 11 (2015) No. 2, pp. 83-93, 2014.

Prenos udarne napetosti iz visokonapetostnega na terciarno navitje močnostnega transformatorja

Mislav Trbušić[1], Anton Hamler[1], Konrad Lenasi[2] (upok.)

[1]*Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, koroška cesta 46, 2000 Maribor, Slovenija*

[2]*Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija*

E-pošta: [1]mislav.trbusic@um.si [1]anton.hamler@um.si [2]konrad.lenasi@fe.uni-lj.si

Surge voltage transfer from a high-voltage winding to a tertiary winding in a power transformer

Abstract. *In the paper the influence of transferred surge wave from a high voltage winding to a tertiary winding of a power transformer is investigated numerically. When testing the high-voltage winding of a power transformer on the lighting impulse, some voltage is transferred to the non-tested windings. While the transformer windings are coupled mutually capacitively and magnetically, the oscillations in the transferred voltage to non-tested windings are to be expected. The aim of the article is to determine the spatial distribution of the oscillations` amplitudes along a tertiary winding with short-circuited and earthed ends. For this purpose, a transient model of transformer windings is represented by an equivalent electric circuit, for which the system of ordinary differential equations of nodal voltages and inductive currents is set and solved using Octave software. The numerical results obtained by calculations show approximately a sinusoidal distribution of the amplitudes in self-oscillations along the tertiary winding, which is in good agreement with the analytical predictions based on the oscillation theory.*

1 Uvod

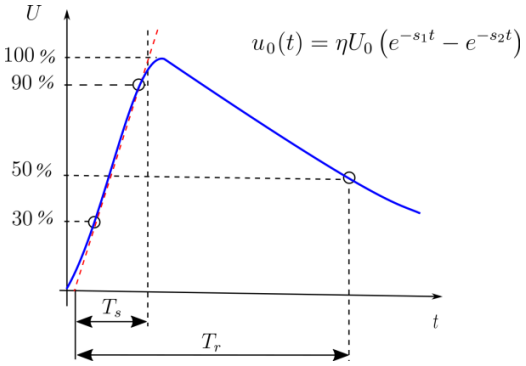
Pri preskušanju visokonapetostnega navitja (VN) močnostnega transformatorja z udarnim napetostnim valom, ki ustreza obliki aperiodičnega napetostnega sunka, ki se pojavi v elektroenergetskem omrežju pri atmosferski razelektritvi, se zaradi kapacitivne in induktivne povezave med navitji del udarnega vala prenese na ostala navitja [1][2][3][4]. Predvsem je kritičen prenos udarnega vala na regulacijsko navitje, kjer se lahko pojavijo napetostna nihanja v navitju katerih amplitude dosežejo tudi do 50% maksimalne vrednosti udarnega vala, kar lahko v določenih primerih povzroči poškodbo izolacije navitja in posledično okvaro transformatorja [1][2]. Manj kritično, vendar prav tako pomembno s stališča koordinacije izolacije v transformatorju je prenos udarne napetosti iz visokonapetostnega na nizkonapetostno (NN) oziroma terciarno navitje (TN), ki se praviloma nahaja tik ob stebru [2]. Problematika prenosa udarne napetosti na NN navitje transformatorja pride do izraza pri posebnih konstrukcijskih izvedbah NN navitja, kjer lahko nastale prenapetosti povzročijo škodo na izolaciji navitja.

Cilj članka je raziskati napetostne razmere znotraj TN navitja med preskušanjem VN navitja z udarnim napetostnim valom in izračunati krajevno razporeditev amplitud vzdolž TN navitja. Z ozirom, da TN navitja praviloma ne preizkušamo z udarno napetostjo, so tovrstni izračuni lahko v pomoč pri načrtovanju izolacije navitja.

2 Matematični model

2.1 Matematični model udarne preizkusne napetosti (udarnega vala)

Kot je bilo omenjeno v uvodu, je namen preizkušanja transformatorskih navitij z udarnim valom preverjanje ustreznosti izolacije navitij, ki so med obratovanjem lahko izpostavljena podobnim napetostnim sunkom zaradi atmosferskih razelektritev. Karakteristično za obliko udarnega vala je strmo čelo, kjer doseže udarni val maksimalno vrednost v rangu mikrosekunde in položen hrbet, kjer vrednost udarnega vala pade na polovico amplitudne vrednosti. Brez superponiranih oscilacij nastopi takšno stanje nekaj deset mikrosekund po sprožitvi vala [1]. S standardom IEC je oblika udarnega vala natančno predpisana z vršno vrednostjo udarnega vala U_0, ki je enaka nazivni vrednosti preizkusne napetosti, z amplitudnim faktorjem η, časom čela T_s in časom, ko pade vrednost udarne napetosti na polovico maksimalne vrednosti (časom hrbta) T_r, pri tem T_s določa strmino čela, T_r pa položnost hrbta (Slika 1)[5][6].



Slika 1: Oblika in definicija časovnih konstant pri standardnem udarnem valu po IEC 60076-4.

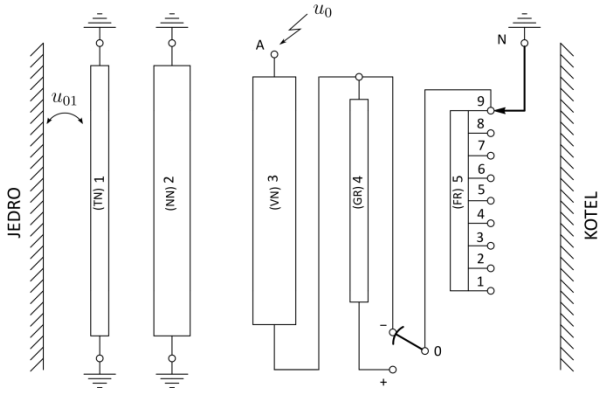
$$u_0(t) = \eta U_0 \left(e^{-s_1 t} - e^{-s_2 t} \right) \quad (1)$$

Vrednosti parametrov standardnega IEC udarnega vala $1{,}2/50$ µs s tolerancami:

- η … amplitudni koeficient 1,036
- U_0 … nazivna vrednost preizkusne napetosti
- s_1 … časovna konstanta hrbta $15000\ s^{-1}$
- s_2 … časovna konstanta čela $2470000\ s^{-1}$
- T_s … čas čela $T_s = 1{,}2\ \mu s \pm 30\ \%$
- T_r … čas, ko pade napetost na polovico maksimalne vrednosti
- $T_r = 50\ \mu s \pm 20\ \%$

2.2 Model transformatorja za tranzientne izračune

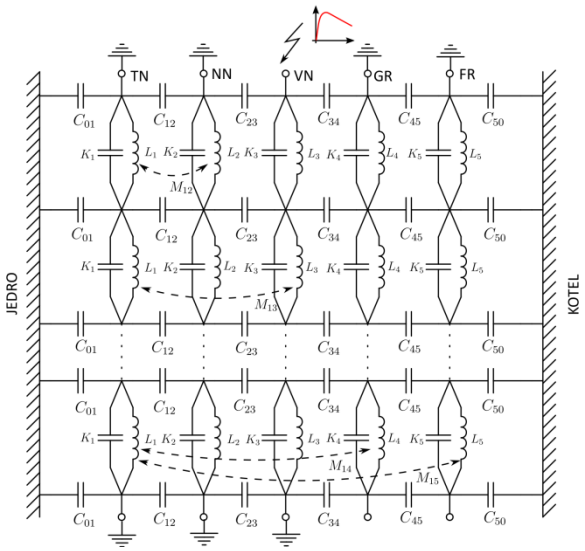
Model transformatorja za namen preučevanja prenosa udarne napetosti med navitji smo naredili za primer transformatorske enote 31,5 MVA - 110/21/(10,5) kV YNyn6+(d5), podobne kot jo uporabljajo elektrodistribucijska podjetja v Sloveniji. Slika 2 kaže razporeditev navitij v oknu transformatorja, pri tem je regulacija napetosti na visokonapetostni strani v obsegu +/- 16 % (8×+/-2%) izvedena z grobo (GR) in fino (FR) stopnjo regulacije. Oznaka napetosti u_{01} na sliki pomeni napetost lastnih nihanj TN navitja do zemlje in je krajevno-časovna funkcija $u_{01} = u_{01}(x,t)$ [1].



Slika 2: Razporeditev navitij v transformatorskem oknu 31,5 MVA-110/20 kV transformatorja v YNyn6+d5 vezavi.

Navitja v transformatorju se lahko za potrebe hitrih tranzientnih izračunov obravnavajo kot kapacitivno - induktivno vezje [2][7][8][9]. Za natančno in praktično uporabo takšnega modela je potrebno razdeliti navitja na sekcije, pri tem je treba pravilno upoštevati kapacitivne in induktivne povezave med sekcijami (Slika 3) [7]. Na sliki so zaradi preglednosti oznake ki se nanašajo na medsebojne induktivnosti podane le za prvo (TN) navitje, podobno velja za ostala navitja.

V modelu smo izračun lastnih in medsebojnih kapacitivnosti izvedli na osnovi analitičnih izrazov, ki jih bralec lahko najde v referencah [1][2][9][10]. Računsko bolj zahtevna je določitev lastnih in medsebojnih induktivnosti med posameznimi sekcijami, za kar je bil razvit lasten računalniški program, ki računa lastne in medsebojne induktivnosti med cilindričnimi tuljavami pravokotnega prereza v poljubnem medsebojnem položaju.



Slika 3: Modelno vezje ene faze tansformatorja za preučevanje hitrih tranzientnih pojavov v transformatorskih navitjih.

Na podlagi nadomestnega vezja transformatorja kot ga kaže Slika 3, lahko na podlagi 1. in 2. Kirchhoffovega zakona zapišemo sistem diferencialnih enačb, za vozliščne napetosti u in induktivne toke i (2) [7][9][10].

$$\begin{bmatrix} \frac{du}{dt} \\ \frac{di}{dt} \end{bmatrix} = \begin{bmatrix} 0 & C^{-1}D \\ -L^{-1}D^T & 0 \end{bmatrix} \begin{bmatrix} u \\ i \end{bmatrix} + \begin{bmatrix} C^{-1}E\ \frac{du_0}{dt} \\ L^{-1}F\ u_0 \end{bmatrix} \quad (2)$$

V sistemu (2) so opuščene ohmske upornosti navitij, kjer veliki delež predstavljajo inducirani vrtinčni toki, ki zaradi frekvenčne odvisnosti vnašajo v sistem (2) nelinearnost, s tem pa tudi otežijo in podaljšajo izračun [8]. Posledično bodo izračunana nihanja v navitjih nekoliko bolj izražena, saj upornost vnaša dušenje, kar s praktičnega vidika predstavlja nek dodaten varnostni faktor, še posebej, če upoštevamo, da računske vrednosti kapacitivnosti in induktivnosti lahko odstopajo od dejanskih tudi do 30% [8][9][10].

Pri tem je pomen oznak v (2):

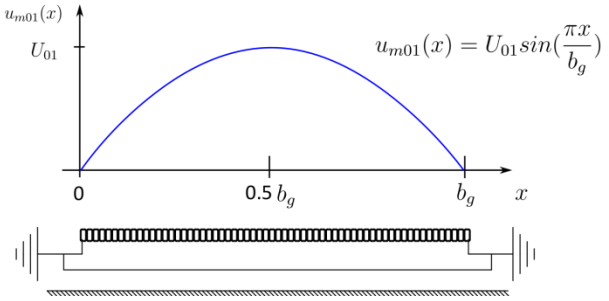
- C … kapacitivna matrika
- L … induktivna matrika
- D … incidenčna matrika

- E … vektor kapacitivnih povezav med udarnim vozliščem in ostalimi vozlišči
- F … vektor, ki določa katero vozlišče je udarno (na katero vozlišče deluje udarni val u_0)
- u_0 … udarni napetostni val
- u … vektor vozliščnih napetosti
- i … vektor induktivnih tokov

Rešitev sistema diferencialnih enačb (2) ob upoštevanju začetnih in robnih pogojev da časovni potek induktivnih tokov $i(t)$ in vozliščnih napetosti $u(t)$, kjer so v luči napetostnih dogajanj med navitji zanimive predvsem slednje.

S teoretičnega stališča, predstavlja prenos udarne napetosti med navitji transformatorja problem robnih vrednosti, kjer pri prostorski porazdelitvi prenesene napetosti igra ključno vlogo način ozemljitve priključnih sponk navitij [1][2]. Pri preskušanju VN navitja so priključne sponke ostalih navitij, ki niso galvansko povezana s preskušanim navitje kratko sklenjene in ozemljene, prav tako je ozemljeno tudi ničlišče VN navitja [5][6]. Z upoštevanjem ozemljenih koncev navitij po Slika 4 je TN navitju moči predpisati Dirichletov tip robnega pogoja, kjer je vrednost potenciala na robu (priključnih sponkah) znana. Največkrat lahko privzamemo, da je referenčna vrednost potenciala ozemljenih delov transformatorja 0 V. Na podlagi teorije nihanj je zato pričakovati, da bodo maksimalne vrednosti napetostnih nihanj vzdolž TN navitja razporejene približno po sinusnem zakonu (3), kjer U_{01} pomeni maksimalno vrednost napetosti TN navitja do zemlje, b_g pa geometrijsko višino TN navitja (Slika 4) [1][2][11].

$$u_{m01}(x) = U_{01}\sin(\tfrac{\pi x}{b_g}) \qquad (3)$$



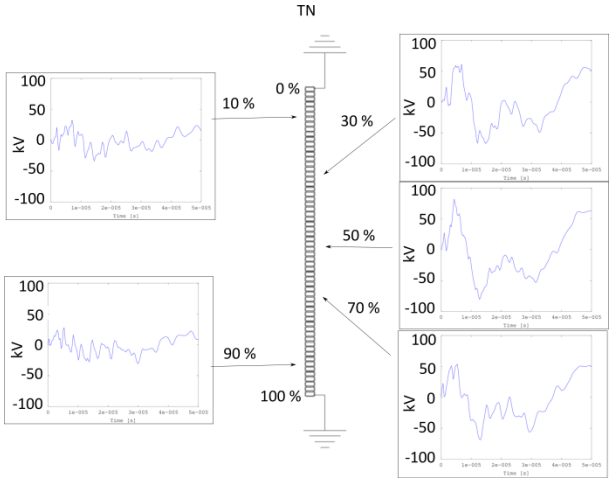
Slika 4: Prikaz krajevne razporeditve amplitudnih vrednosti lastnih nihanj kratko-sklenjenega in ozemljenega enoplastnega plastnega navitja.

Praviloma je TN navitje načrtovano za tretinjsko moč transformatorja za kar velikokrat zadostuje že izvedba navitja s profilno žico v enoplastni izvedbi (Slika 4).

3 Rezultati

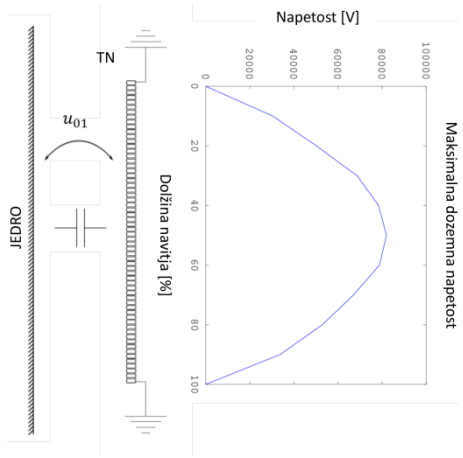
Izračuni prenosa udarne napetosti iz VN na TN navitje so bili opravljeni na tranzientnem enofaznem modelu transformatorske enote 31,5 MVA-110/21/(10,5) kV YNyn6+(d5) v programskem paketu Octave [12]. Pri simulacijah smo upoštevali vezavo po Slika 2, ki ustreza preizkušanju VN navitja brez vključenih regulacijskih ovojev, kjer lahko pričakujemo največje amplitude napetostnih nihanj v TN navitju, saj je razmerje med številom ovojev TN in VN navitja v tem primeru največje [1][2]. V izračunih smo upoštevali standardno obliko udarnega vala (1,2/50) (1) z amplitudo U_0=550 kV, ki se uporablja za preizkušanje 110 kV izolacijskega nivoja [5][6].

Slika 5 prikazuje časovni potek prenesene udarne napetosti u_{01} vzdolž TN navitja iz katere je opazno, da se amplitude lastnih nihanj povečujejo s približevanjem sredini navitja. Prav tako je moč opaziti, da so amplitude višjih harmonskih komponente lastnih napetostnih nihanj bolj izražene na začetku in koncu navitja ter se njihova vsebnost proti sredini navitja zmanjšuje. V časovnih potekih napetosti na Slika 5 so nihaja prikazano v časovnem intervalu od $0 - 50$ µs.



Slika 5: Izračuni časovnih potekov lastnih napetostnih nihanj u_{01} vzdolž TN navitja, ki se pojavijo med preskušanjem VN navitja z udarno napetostjo.

Krajevna razporeditev maksimalnih vrednosti dozemnih napetosti vzdolž TN navitja u_{m01} je prikazana na Slika 6, kjer je razvidno, da je krajevna razporeditev amplitud lastnih nihanj vzdolž navitja približno sinusne oblike z amplitudo $U_{01} \approx 80$ kV in predstavlja stojni val, ki zaniha okoli referenčne vrednosti napetosti. Razmerje med amplitudo prenesene in udarne napetosti je $\frac{U_{01}}{U_0} \approx 0{,}15$, kar je v rangu razmerja ovojev TN in VN navitja. Potrebno je poudariti da na fazo lastnih nihanj vpliva smer navijanja navitij, kjer so lastna nihanja v VN in TN navitju sofazna, kadar se navijalni smeri razlikujeta in protifazna, ko sta navitji naviti v isto smer. Upoštevanje smeri navijanja navitji pri prenosu udarnega vala je še posebej pomembno pri posebnih konstrukcijskih izvedbah navitij, kjer se dve navitji med seboj prepletata.



Slika 6: Računske vrednosti krajevne razporeditve amplitud lastnih napetostnih nihanj vzdolž TN navitja.

4 Sklep

Na podlagi računskih rezultatov razporeditve amplitud lastnih napetostnih nihanj vzdolž kratko-sklenjenega in ozemljenega TN navitja lahko sklepamo, da se pri prenosu udarne napetosti iz VN na TN navitje ta razporeditev spreminja po sinusni funkciji vzdolž navitja, pri tem bodo lastna napetostna nihanja v navitju največja približno na polovici navitja, medtem, ko bo medovojna izolacija najbolj obremenjena na začetku in koncu navitja. Približno lahko privzamemo, da je razmerje med maksimalno vrednostjo prenesene napetosti $|U_{01}|$ in amplitudo udarne napetosti $|U_0|$ enako razmerju števila ovojev TN in VN navitja, pri tem je treba upoštevati najbolj neugoden primer, ko je položaj regulacijskega stikala na skrajno negativnem odcepu regulacije. Napetostni gradient, ki ima največjo vrednost na začetku in koncu navitja, ne predstavlja nevarnosti za medovojno izolacijo enoplastnega TN navitja, ker je le ta zaradi tehnoloških razlogov pri izdelavi navitja praviloma predimenzionirana za napetostni nivo terciarja.

Literatura

[1] B. Heller, A. Veverka, Surge Phenomena in Electrical Machines, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1968.

[2] Karsai K., Kerényi D., Kiss L., Large Power Transformers, Budapest: Académia Kiadó, 1987

[3] Popov M., Van der Sluis L., Smeets R. P. P., "Evaluation of surge-transferred overvoltages in distribution transformers", Elec. Pow. Syst. Res., 78, (2008), no. 3, 441-449

[4] Borghetti A., Morched A., Napolitano F., Nucci C. A., Paolone M., "Lightning-induced overvoltages transferred through distribution power transformers", IEEE Trans. on Pow. Deliv., 24, 2009, 1, 360-372

[5] IEC 60076-1, Power Transformers, Part 1: General, IEC Standard, 2000.

[6] Power Transformers. Part 4: Guide to the lighting Impulse and Switching Impulse Testing – Power Transformers and Reactors, 1st edition, IEC 60076-4, 2002

[7] Lenasi K., Analiza udarnih napetosti v navitjih transformatorja z matematičnim modelom, magistrsko delo, Ljubljana, 1974.

[8] K. Lenasi, Udarne prenapetosti v navitjih močnostnih transformatorjev z vpeljavo dušenja induktivnih tokov, doktorska disertacija, Ljubljana 1975.

[9] B. Ćućić: Raspodjela udarnog napona po preloženom namotu transformatora, Magistrsko delo, Fakultet elektrotehnike i računarstva, Zagreb, 2001.

[10] Trbušić M., Čepin M., "Razporeditev udarne napetosti vzdolž zvrnjenega navitja močnostnega transformatorja", Elektrotehniški vestnik, 78, (2011), no. 3, 106-111

[11] Inman D.J., Engineering vibration, Pearson Prentice Hall, Upper Saddler River, New Jersey, 2009.

[12] Octave software, www.gnu.org/software/octave/ (zadnjič obiskano 5.7.2018)

Improve forecasting of daily load

Boris Bizjak

Faculty of Electrical Engineering and Computer Science
University of Maribor
Koroška cesta 46, 2000 Maribor, Slovenija
E-pošta: boris.bizjak@um.si

Improve forecasting of the daily load

Abstract. *In this paper, we discuss the daily hourly load forecasting of load for an industrial complex using ARIMA methodology with one predictor. For predicting the daily load, we need to make a forecast of 24 steps in the case of the 1-hour prediction interval. The story of predicting in the article is due to trading in electricity. The basis for the forecasting is the time series of energy consumption data, which is a time series from 15 min values.*

The presented load forecasting model needs at the learning phase two input historic time series: Energy consumption and real production data. We get production data (predictor) from the Production Planning Department. We wanted to use a simple and robust predictor, which not only has a theoretical value for forecasting, but also a practical utility to be used in practice. So, the predictor has a value of 1 when the furnace is running and 0 when the furnace is not in operation. For predictor type 1, the time resolution at the predictor is 1 hour, and for predictor type 2 the time resolution at the predictor is 15 min. The predictor also follows production with different dynamics. Type 1 has a constant value of 1 for all the duration interval of the arc furnace working, and predictor type 2 follows the arc furnace cycle more dynamically with fluctuation between 1 and 0.

To find the optimal predictor we used a similar spectral density method of two time series. We called it predictor type 2. In the case of the 1-hour prediction step with predictor type 2, by linear regression, the independent variable of production explains 84% of the variance in the load, MAPE is 23% and RMSE 3389. Additionally, energy consumption at the industrial plant had a seasonal day, hours feature. Therefore, we improved forecasting performance - statistical indicators, using ARIMA methodology, by approximately 15%.

Finally, the ARIMA model, with one predictor type 2, explains 95% (R^2 0.95) of the variance in the load, with MAPE 9% and RMSE 1982.

1 Introduction

Our example of predicting the daily power flow for the industrial complex has a specific large electric arc furnace (UHP). In addition, other companies within the industrial complex operate according to the seasonal model. Therefore, three prediction methods should be used in order to make good predictions: ARIMA methodology, seasonal model and predictor of the operation of the arc furnace. In the article, we first show how to determine a predictor that illustrates the operation of an arc furnace. The final result is a forecast of the daily power flow using the selected predictor, ARIMA and the seasonal model.

2 Industrial complex and steel production

The beginnings of the ironworks in Koroška go back to the year 1620. Today, there are several companies in the area of the former Ravne Ironworks, employing approximately 3,000 people. The largest company on the site is Metal Ravne with approximately 900 employees. Metal Ravne is the largest, and the largest consumer of electricity. The company consists of a steelwork, a rolling mill and electro-smelting under slag. In the steel plant, the basic unit is a 45-tonne electric UHP oven and a vacuum refill kiln for castings of classical ingots. In the Electro-under-slip section under the slag, 36-tonne and 3-tonne ESR devices are in use. At present, it operates on the scale of an industrial complex of 20 companies.

The melting process [1] is always carried out in an arc furnace with reduced voltage, since the conditions for burning the arc are poor in a cold cartridge; the ignition of the arc is carried out in such a way that the graphite electrode is lowered to the cartridge until it touches it, and until contact is reached with the other electrodes with the cartridge. At the discharge of the electrode, then an electric arc is triggered - like the firing of the arc during manual arc welding. Because of this, the current size changes from the short-circuit current through the rated power to the zero current at the end of the arc. We say that the arc furnace is operating restlessly at the beginning of melting. Due to the formation of the first melt at the bottom of the furnace, the conditions for burning the arc are improved due to good ionization conditions, so we increase the voltage of the arc gradually and the power of melting to full power: This is always the largest when melting the cartridge when there is already a melt on the bottom of the furnace. We say that we are melting with a hidden arc, which radiates at full power in the crater, which the boulder has drilled into the plunged insert of old iron. In the further heating of the melt, or in maintaining its temperature, the power of the furnace is

significantly lower. The characteristics of the electric arc must be different in this situation, since the arc can now freeze to the walls and the furnace vane.

3 Forecasting daily power consumption per hour interval

Due to the way of trading, the forecast of the daily flow of electricity flows requires a forecast for at least 24 hours, even better in 48 steps (48 hours). We want to predict power flow on the common energy supply of the industrial complex Ravne. The structure of total energy consumption for the Ravne industrial complex is the sum of 20 different companies. The biggest, 75% of the energy consumers, is the Metal Ravne Steel Mill. The dominant energy consumer in the steel industry is the UHP electric arc furnace, the second largest consumer is the oven for overheating LF + VD, and then the other technological line of the steel mill, which follows the melting process of old iron. The remaining 25% of consumption, on a common energy connection, is represented by the other companies, with a typical spending profile of 7 days a week, of which 5 are working days. Change of 25% of consumption over a day is relatively low, and therefore it can be predicted very well.

The steel plant does not operate statistically on a random basis, but over the day according to the orders and the price of electricity. Since its consumption is 75% of the entire energy of the industrial complex, it has the greatest influence on the consumption profile on the total energy supply of the industrial complex. The issue is the production time; real production will be moved to the late afternoon, next to the evening (cheap electricity), over midnight, until 9:00 AM the next day. Occasionally, when there is a large volume of orders, the steel plant operates for 24 hours continuously, regardless of the daily price of the electric power unit over a working day. Due to cheap electricity, Saturdays and Sundays also operate for 24 hours. Once a year, the steel works carry out a repair that lasts about a month. Of course, the electric arc furnace does not work then. Power flow in a UHP arc furnace during steel production is more or less constant. The difference between individual loads at the arc furnace is only in what happens after we have finished melting in the arc furnace. Sometimes overheating occurs and sometimes it does not, which means a short leap in the consumption of electricity to the maximum.

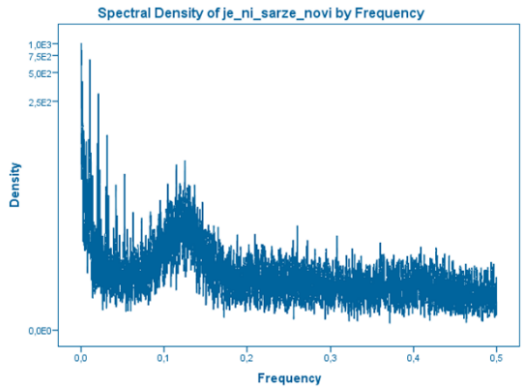


Figure 1: Spectral density for 15 minutes' power flow.

The spectral density of the for 15 minutes' power flow for one year indicates additional periodicity of the signal at frequency 0.12 (Figure 1). Given the spectral density fluctuation for 1-hour average power flow at one year, we have had monotonic falls from low frequencies to the highest frequencies. In other words: 1-hour time series have less frequency content.

The prediction was started with predictor type 1. The predictor has a value of "1" if the hourly average power consumption is > 4500 and "0" if the hourly average power consumption is < 4500. The predictor's time is coincident with the progress of the blue step line in Figure 2. We reached: The ARIMA model, with one predictor, explains 85% (R^2 0.85) of the variance in the load, with MAPE 13%. The results of predicting with the type 1 predictor were not satisfied.

Therefore, we have introduced a new type 2 predictor with only two amplitude values: 0 and 1. Predictor type 2 is more time-divergent than type 1, so it is better to follow the operation of the electric arc furnace and other technological lines of the steel plant. The predictor has a value of "1" if the 15 minutes' power consumption is > 6500 and "0" if the 15 minutes' power is < 6500.

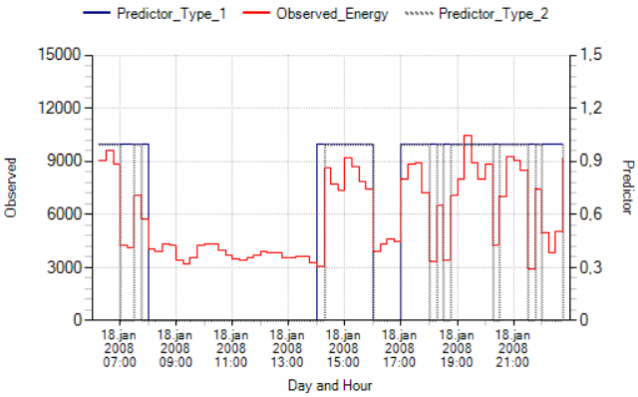


Figure 2: Predictor type 1 and predictor type 2.

To find the optimal predictor we used a similar spectral density method of two time series. Predictor type 2 is an optimal predictor. This is confirmed by the similarity of the density of the frequency spectrum of the time series of electricity consumption and the frequency of the type predictor type 2 for the same time period (Figure 1,3). Also, there were excellent results of predictions using the type 2 predictor.

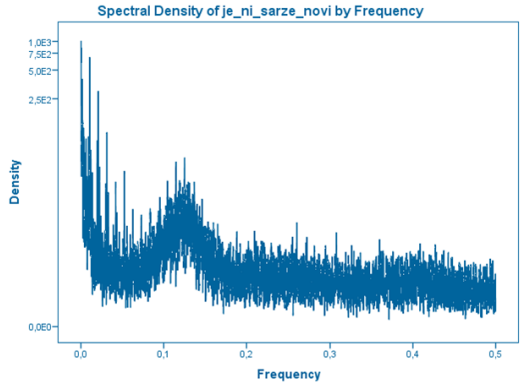


Figure 3: Spectral density for predictor type 2.

In the time series of electricity consumption, the outliers are also noticed; these are values when the consumption of electricity escapes from the standard Gaussian distribution and is close to zero. Such values are expressed on May 1 and December 31. These two cases were not addressed specifically in the forecast itself, but, in any case, they have a negative impact on forecasting performance statistics.

Various statistics are used to evaluate the performance of forecast models. We decided to use in the article only MAE, MAPE, RMSE and R^2 to comparison between individual solutions.

The Mean Absolute Percentage Error (MAPE), is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses accuracy as a percentage, and is defined by the formula:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Ai-Fi}{Ai}\right|$$

where Ai is the actual value and Fi is the forecast value. Problems can occur when calculating the MAPE value with a series of small denominators. A singularity problem of the form 'one divided by zero' and/or the creation of very large changes in the Absolute Percentage Error can occur, caused by a small deviation in error.

The Root-Mean-Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values observed. The RMSE represents the sample Standard Deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation and are called prediction errors when computed out-of-sample. RMSE is a measure of accuracy, to compare forecasting errors of different models for data and not between datasets, as it is scale-dependent. RMSE is sensitive to outliers.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Ai-Fi)^2}{n}}$$

R squared in statistics, the coefficient of determination, denoted R^2, is the proportion of the variance in the predictable variable Fi that is from the actual value Ai:

$$SS_{tot} = \sum_{i=1}^{n}(Ai-\overline{Ai})^2$$

$$SS_{res} = \sum_{i=1}^{n}(Ai-Fi)^2 = \sum_{i=1}^{n}e_i^2$$

$$R^2 = 1-\frac{SS_{res}}{SS_{tot}}$$

For the described type of industry complex we started forecasting continuously with linear regression [2]. This is a robust method that has a weakness in the limited level of confidence in the forecast. We calculate the 1-hour step predictor type 2 from the 15-minute time series with aggregate functions. The predictor's time is coincident with the progress of the grey step line in Figure 2. In the case of the 1-hour prediction step with linear regression, the independent variable of production data at the arc furnace now explains 84% of the variance in the load, which is highly significant, and the F-test says we can trust β_0 and β_1> 99.9 % , MAPE 23% and R square 0.84.

Table 1: Daily load forecasting at linear regression.

R square	MAE	MAPE	RMSE	Number of predictions
0,84	2700,62	22.84 %	3388,97	8543

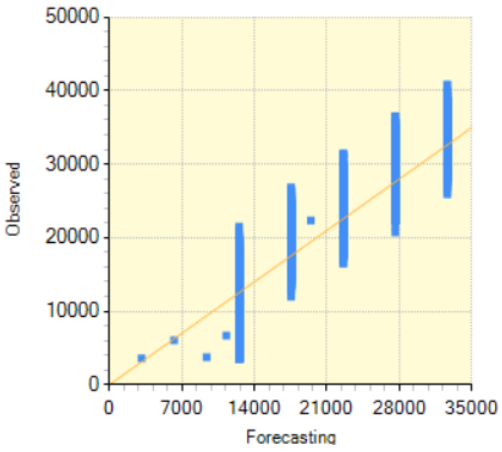


Figure 4: Daily load forecasting at linear regression.

Figure 4 show a typical chart of a prediction with linear regression. A forecast with linear regression could be improved by changing the predictor's timing or amplitude, which means that the on / off principle would be adapted more often to actual production or that the amplitude would follow a production value with not only 0 or 1, but, for example, "0.9", "1" and "1.1". We achieved this by proceeding with the forecast in 15-minute steps, and then we aggregate the 15 min prediction to a 1 hour prediction step. We do this because we believe that such a predictor can be expected from the Production Planning Department.

Table 2: Forecasting daily power consumption per hour at ARIMA methodology.

Hour +	R square	MAE	MAPE	RMSE	Number of predictions
1	0,95	1386,93	6.81 %	1965,98	504
2	0,94	1455,65	7.33 %	2060,92	504
3	0,94	1534,3	7.8 %	2106,36	504
4	0,94	1616,43	8.26 %	2149,73	504
5	0,93	1675,69	8.78 %	2224,52	504
6	0,93	1711,12	9.08 %	2238,19	504
7	0,93	1748,47	9.28 %	2268,13	504

8	0,93	1756,52	9.51 %	2280,33	504
9	0,93	1759,68	9.47 %	2280,7	504
10	0,93	1789,77	9.76 %	2317,68	504
11	0,93	1828,01	9.83 %	2360,4	504
12	0,93	1837,03	9.91 %	2364,44	504
13	0,93	1824,9	9.89 %	2330,4	504
14	0,92	1855,75	10.02 %	2387,83	504
15	0,93	1816,1	9.87 %	2330,17	504
16	0,93	1845,7	10.03 %	2352,68	504
17	0,93	1795,34	9.73 %	2276,53	504
18	0,93	1794,97	9.77 %	2286,87	504
19	0,93	1813,72	9.91 %	2295,4	504
20	0,93	1814,17	9.77 %	2322,71	504
21	0,93	1758,78	9.7 %	2262,13	504
22	0,93	1781,22	9.81 %	2309,08	504
23	0,93	1780,48	9.86 %	2296,16	504
24	0,93	1786,35	9.99 %	2298,05	504

From linear regression results will be improved with ARIMA metrology [3] and one predictor. Energy consumption at the industrial plant had a seasonal day, hour (7/24) feature and one dominant predictor. The forecasting results in Table 2 show much better: For "24 hour - 24 steps" forecasting the MAPE statistic was 9% and R^2 0.95.

Linear regression has an RMSE constant value of 3388 for all predictions. With ARIMA methodology, for the same historical time series, RMSE ranges between 1386 and 1859, which is a significant improvement. So, the model ARIMA plus predictor is a better prediction model, because the lower RMSE value means a better forecast model. Looking in Table 2 at the 24-hour forecasts, we notice that better forecasts are for the values near 1+ and 24+ hours, which shows the 24-hour periodicity of our time series.

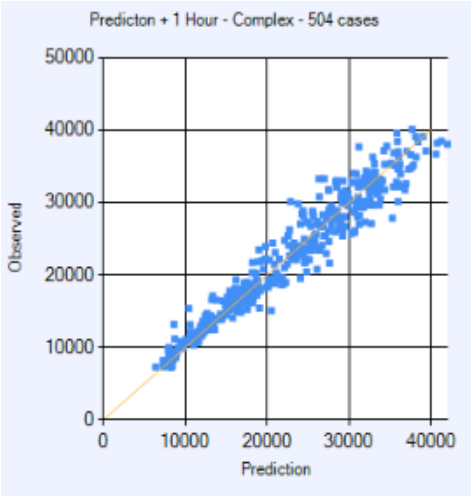


Figure 5: Forecasting daily power consumption per hour with ARIMA methodology, + 1 hour.

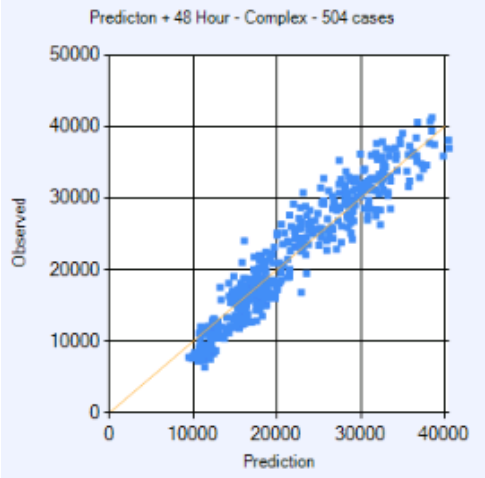


Figure 6: Forecasting daily power consumption per hour with ARIMA methodology, + 48 hour.

Figure 5 and Figure 6 shows that we have improved the forecast for low power consumption dramatically. There is also an improvement in high consumption, but it's not as good as we would like.

3 Conclusion

The test forecasting was carried out over 3 weeks from December 1 to December 21. We predict 48 times for each hour, which means first 48 hours ahead ... and the last announcement is 1 hour ahead. The results are shown in Figure 5,6 (+1 hour, +48 hours) and in Table 2 for each hour plus separately. The advantage of the presented forecast system is its small forecast error. Finally, forecasts for the Ravne industrial complex are implemented with ARIMA methodology, seasonal models and extended with one predictor. The solution is based on modified IBM SPSS and Microsoft software products, and the theory from [2] and special [3]. Models of forecasts are learning dynamically. The prediction models are, thus, optimized constantly according to the current time series. To learn the model, we use the time series of the common energy meter of the Ravne industrial complex and the optional energy meter before the steelmaker Metal Ravne. The prediction database is a SQL server. All images and Tables are part of an existing WEB site application.

References

[1] Janez Bratina, Elektrobločna peč, Ravne na Koroškem, 1994.

[2] Douglas C. Montgomery, George C. Runger, Applied statistic and probability for engineers, Wiley, 2003.

[3] George E. P. Box Gwilym M. Jenkins Gregory C. Reinsel. Time series analysis forecasting and control, Wiley, 2014.

Generic Test Network Model for Scalability and Replicability Analysis on the DeCAS Project

Leopold Herman, Mitja Antončič, Boštjan Blažič

Faculty of Electrical Engineerign, University of Ljubljana
E-pošta: leopold.herman@fe.uni-lj.si

Abstract. *The main aim of this paper is to present test network model to be used within European project DeCAS by individual pilot sites to demonstrate replicability and scalability of proposed solutions and results. The main objective of the DeCAS project is to research and analyse the coordination of ancillary services such as, aggregated demand response, individual voltage control and reactive power management concepts over traditional boundaries from high voltage (HV), medium voltage (MV) to low voltage (LV).*

1 Simulation Model Description

The simulation model is one of the results of an European project DeCAS (Demonstration of Coordinated Ancillary Services covering different Voltage Levels and the Integration in Future Markets) [1, 2].

The simulated test network covers all elements between the transmission (220 kV) level, distribution (110 kV and 20 kV) level and the LV connection points. A 110 kV looped grid is supplied from two sides, with two three-winding 220/110 kV transformers.

At the MV level, two feeders (rural and urban) were modelled in details, representing the generic model of two typical feeders. The consumption of other MV feeders was determined by appropriate scaled equivalent loads.

At the LV level, several LV (radial) feeders were modelled in details to represents the situation in rural and urban LV networks. The consumption of other LV networks that were not modelled in details was determined by scaled equivalent LV loads.

1.1 HV Network

Modelled HV network is supplied with a slack bus. In order to observe the variables in the HV net-work, one 110 kV loop has been modelled. Underlying MV networks are supplied through this HV loop. Topology of the proposed HV network is presented in Figure 1.

1.2 MV Network

To illustrate a typical situation in the MV network, the transformer station was modelled with two generic radial feeders:

– Rural (OH) feeder and
– Urban (cable) feeder.

The following line parameters have been considered:

Rural line	Urban line
Type: Al/Fe 70/12	Type: NA2XS(F)2Y 1x150RM
U_r = 20 kV	U_N = 12/20 kV
I_r = 290 A	I_N = 309 A
r = 0,413 Ω/km	r = 0,210 Ω/km
x = 0,362 Ω/km	x = 0,122 Ω/km
	b = 79,796 S/km

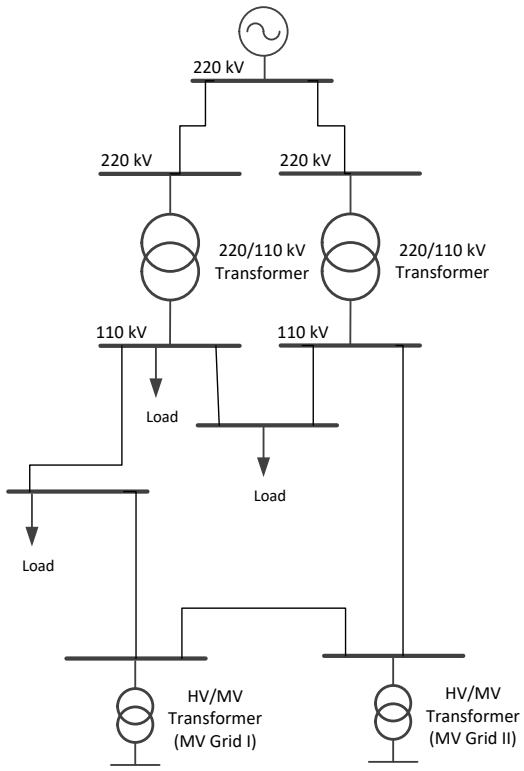


Figure 1: Modelled HV network single-line diagram.

U_r being the nominal voltage, Ir the nominal current, r series resistance per km, x series reactance per km and b parallel susceptance.

The length of the rural feeder is 8 km, which represents a typical length of a rural feeder, where larger (400 kVA) transformers (5 in total) are uniformly distributed along the beginning of the feeder, followed by 13 transformers of 250 kVA and 4 transformer with a nominal power of 160 kVA. Transformers towards the end of the feeder are supplying remote small settlements, with low consumption.

The length of the urban feeder is 2 km, with 400 kVA transformers (10 in total) uniformly distributed along the feeder.

Rural feeder mainly supplies household consumers, while a portion of industrial or business consumers is attached to the urban feeder (approximately 30% of total consumption).

The MV network modelled in DIgSILENT PowerFactory v.2017 software is shown in Figure 2.

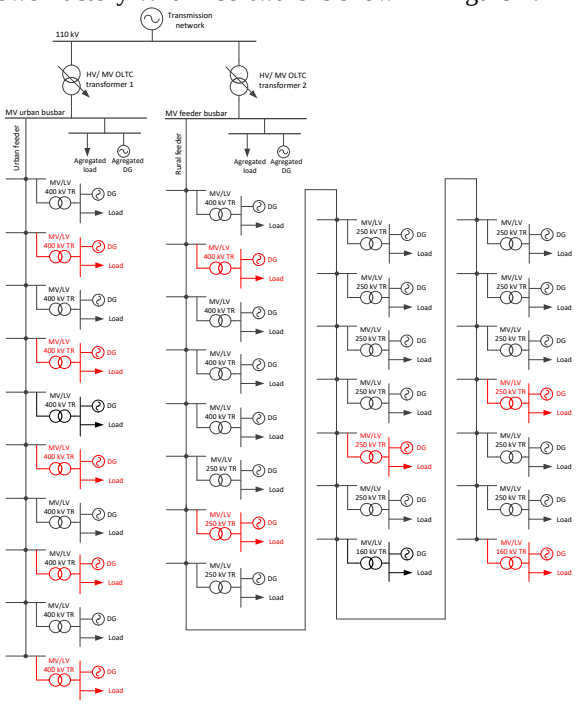


Figure 2: Single-line diagram of a MV distribution network with two feeders. At the point, where the transformer stations are reddish, LV networks are modelled in details.

1.3 LV Rural Network

Individual LV radial feeders are supplying 70 customers and are connected to the rural MV network at different points along the feeder.

The main conductors have a cross-section of 70 mm^2, more distant conductors are 35 mm^2 and 16 mm^2. The total lengths of conductors of all feeders are as follows:

- 3.000 m conductors with a cross-section of 70 mm2,
- 1.000 m conductors with a cross-section of 35 mm2,
- 1.500 m conductors with a cross-section of 16 mm2.

The following line parameters have been considered:

X00/0-A 70 (NFA2X)	X00/0-A 35 (NFA2X)	X00/0-A 4x16 (NFA2X)
U_N = 0,6/1 kV	U_N = 0,6/1 kV	U_N = 0,6/1 kV
I_N = 223 A	I_N = 142 A	I_N = 91 A
r = 0,496 Ω/km	r = 0,972 Ω/km	r = 2,139 Ω/km
x = 0,100 Ω/km	x = 0,100 Ω/km	x = 0,100 Ω/km

In contrast to the MV network, the loading of lines vary considerably over time. The rural network supplies only household consumers.

1.4 LV Urban Network

Individual radial feeders of urban LV (cable) network are connected to the MV urban feeder at different points along the feeder. Consumption is more concentrated than in the rural case, and the feeders are shorter and amount to about 400 m. The following data has been used:

Type: PP 00-A 4x70 (NAYY)

U_r = 0,6/1 kV

I_r = 175 A

r = 0,444 Ω/km

x = 0,075 Ω/km

b = 254,5 S/km

The network supplies both, household and business customers.

In both cases, with the rural and urban LV networks, the transformers tap positions are set so that the nominal voltage at the LV busbar in the transformer station is always achieved, thus allowing a 10 % voltage drop in the LV network.

2 Generation of Load and DG Profiles

For modelling the loads, real measured 15-minute load data of LV customers were used. The measurement data contains 1 year load profiles of 3000 consumers. For each transformer station, the required number of consumers is set in such a way that in winter times, when a peak in demand is reached, the transformer is loaded with approximately 75 %.

The aggregated household consumption differs from the business or industrial load especially in a clear eveningpeak (see Figure 3). In the case of other types of loads, the peak demand occurs during working hours. As mentioned, in the urban network, 70 % of the consumption is modelled as household loads and the rest 30 % as a combination of business and industrial loads.

In terms of voltage dependence, loads are modelled as constant power type of load and operate with cosφ = 0,95.

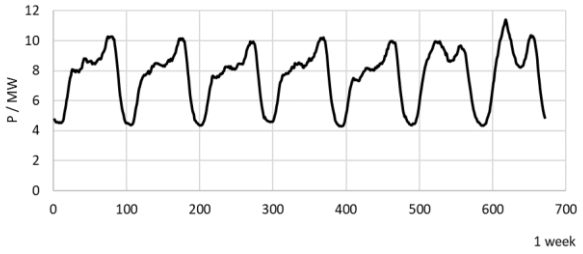


Figure 3: Week diagram of 3000 household loads.

Simulations can be performed for different days of the week (or for several days) and for different seasons. Household consumption is higher during the weekends, whereas industrial consumption is generally lower during the weekends. There are also some other differences.

The time of the year for which we are preforming the simulations is also important. In winter, consumption is high and production from solar power plants (PV system) is low, and in the summer it is the opposite, consumption is lower, and production from DG is higher. An example of one-week production of a 10 kW

solar power plant is shown in Figure 4, where we can see how production can vary within one week.
Similarly to load diagrams, measured data was also used to model the PV sources. Diagrams for the production of solar power plants are based on yearly measured production diagrams with appropriate scaling of power. In the analysis, only PV sources were taken into account, since they represent the majority of DG in distribution networks, taking into account that the rated power of each PV source is 10 kW.

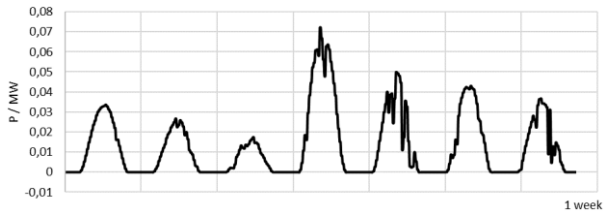


Figure 4: Example of a one-week production diagram from a solar power plant

3 Simulation Procedure

Load flow simulations can be performed on the network model presented in previous chapters. Several loads are connected to the MV network, which represent the consumption of individual transformer stations. In order to analyse the situation on the LV level, at different locations along the MV feeder, several LV feeders (urban and rural) were modelled in details. One- or several-days simulations of all networks (MV and LV) are carried out simultaneously.
The process of calculating the LF simulations roughly follows the following steps:
1. A network model is developed in a simulation program.
2. Load consumption and PV production diagrams for one or several days with a resolution of 15 min are generated with Matlab file LoadPVGenerator.m.
a. Randomly, but with certain limitations, the consumption of loads is determined and the PV units are placed in the network.
3. The load flow is calculated for all the operation points.
4. The simulation results (voltages, power flow etc.) for every simulation time step are saved into an excel file.
Simulations can be made for different days of the week and for different seasons. The results below represent the situation for a winter day, as the most severe season.

4 One-Day Simulation Results

4.1 MV feeders

The maximum voltage drop that occurs on the MV rural feeder is about 4 % and maximal power consumption of the feeder about 3 MVA. Both number represent a realistic situation on the MV level.
The maximum voltage drop occurring on the MV urban feeder is about 0,8%, with a maximal power flow of 2 MVA.
Despite the fact that permitted voltage drop in the MV network is 7.5 %, in reality, such high values normally do not occur. In reality, 5% drop in voltage is already a very high value.
In the MV transformer station there is also a generic load and generation connected to the MV busbars. The power of the generic load on the rural feeder is about 10,5 MVA, and on the urban feeder 14,5 MVA.
The 20 kV network is connected to the 110 kV level via a 31,5 MVA OLTC transformer, with ± 12 taps of 1,33% U_n each. The reference voltage on the secondary side of the OLTC transformer is 1,03 p.u.

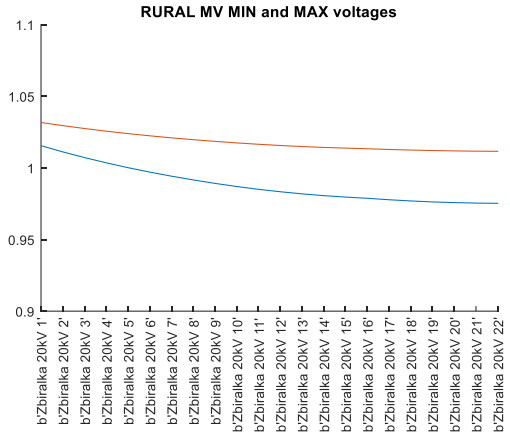


Figure 5: MIN and MAX voltages of individual MV busbars – rural feeder.

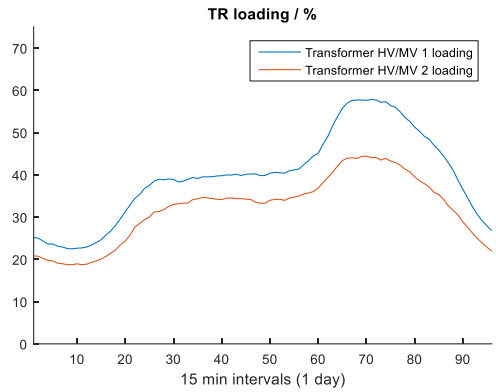


Figure 6: Loading of HV/MV transformers.

4.2 LV rural feeders

Below in Fig. 7 are power flow results for transformer station no. 19 on the rural MV feeder. The maximum power flowing through the transformer is about 125 kVA, which means 78 % of the peak load. The highest voltage drop is app. 6,9 % (Fig. 3.9).

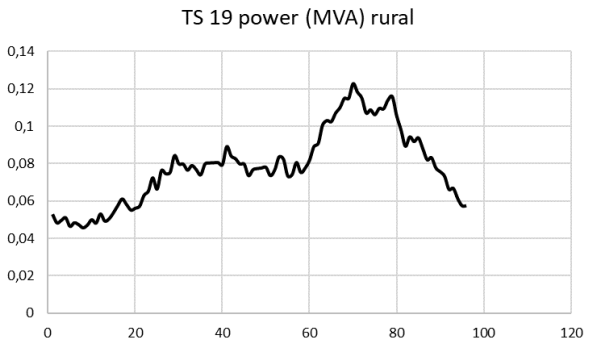


Figure 7: MV/LV transformer no. 19 (rural feeder) load flow data.

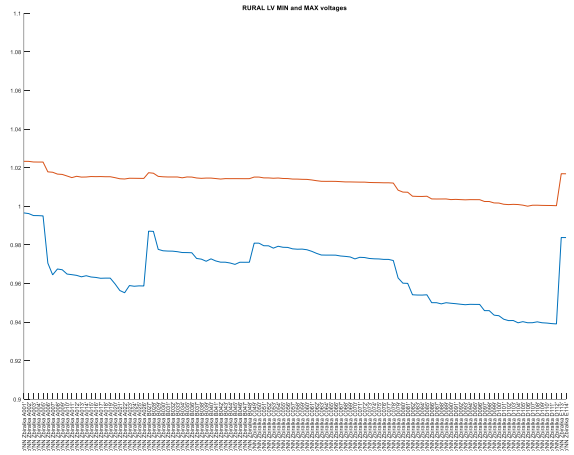


Figure 8: MIN and MAX voltages of individual LV busbars – rural feeder.

4.3 LV urban feeders

Below in Fig. 9 are power flow results for transformer station no. 10 on the urban MV feeder. The maximum power flowing through the transformer is about 226 kVA, which means 57 % of the peak load. The highest voltage drop is app. 1,8 %.

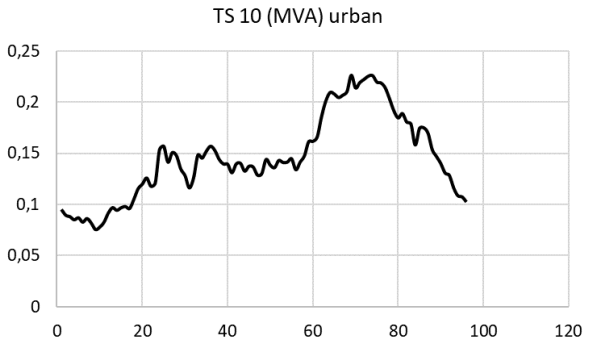


Figure 9: MV/LV transformer no. 10 (urban feeder) load flow data.

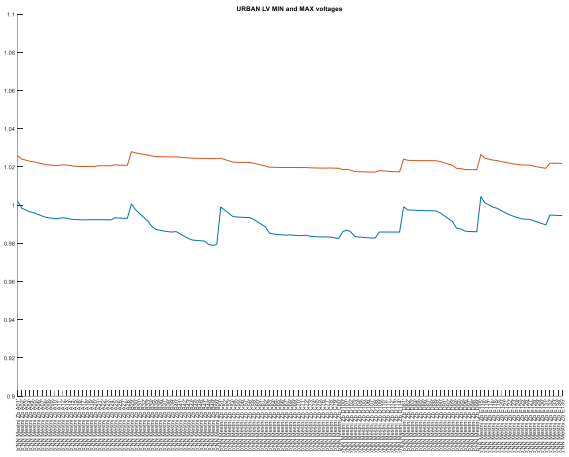


Figure 3.10: MIN and MAX voltages of individual LV busbars – urban feeder.

5 Conclusions

In this paper a generic test network model was presented, which was used within the DeCAS project for scalability and replicability analysis of individual pilot action's outputs.

Literature

[1] Fingrid, 'Supply of Reactive Power and Maintenance of Reactive Power'. .

[2] 'DeCAS project'. [Online]. Available: decas-project.eu. [Accessed: 05-Dec-2017].

Modelling of Equivalent Low-Voltage Nodes

Leopold Herman, Mitja Antončič, Boštjan Blažič

Faculty of Electrical Engineerign, University of Ljubljana
E-pošta: leopold.herman@fe.uni-lj.si

Abstract. *This paper presents modelling procedure and simulation results of equivalent low-voltage nodes models. Models were used in a DeCAS project, to define the reactive power of the LV grid as a function of the load, generation and voltage and especially to simplify the modelling procedure of large networks.*

1 Introduction

This paper demonstrates results of the DeCAS project (Demonstration of Coordinated Ancillary Services covering different VoltageLevels and the Integration in Future Markets) [1, 2]. The main objective of the DeCAS project is to research and analyse the coordination of ancillary services such as, aggregated demand response, individual voltage control and reactive power management concepts over traditional boundaries from high voltage (HV), medium voltage (MV) to low voltage (LV) and develop approaches and concepts for a co-ordinated control approach considering the different objective functions of individual voltage levels. It included the integration related monitoring and controls in process control systems as well as to flexibility markets. LV grids are usually not automated yet and there are hardly any measurements available. Thus, the project evaluates promising concepts for LV grid operation tools, processes and how they can interface with MV/HV SCADA DMS [3, 4].

A Passive LV network model development approach that was developed within the DeCAS project is presented in this paper. Whole LV network is to be represented as black box with its inputs and outputs being explained hereafter. For the sake of transparency, we will refer to the "black box LV model" as a "model" through the whole paper.

2 Passive LV Nodes – Black Box Model

2.1 Model description

The purpose of the developed model is to define the reactive power of the LV grid as a function of the load, generation and voltage. It can be written as:

$$Q_{LV} = f(P_{load}, P_{PV}, U) \quad (1)$$

Conceptual scheme of the developed model is shown in Figure 1.

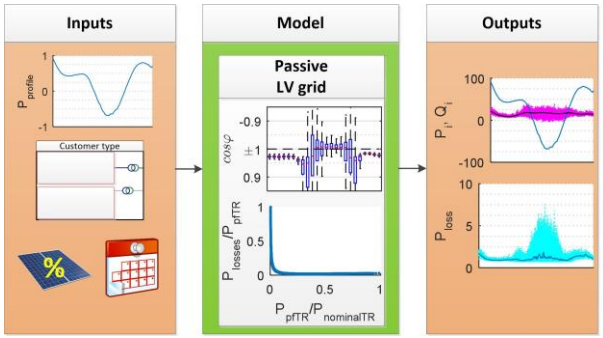


Figure 1: LV network model concept

Input for the model is a normalized active power injection profile of particular LV network, together with the time of the year, the type of the network customers (e.g. residential, commercial, urban, rural…) and the PV penetration as a rate of transformer nominal power.

Outputs from the developed model are active and reactive power injections that can be used for definition of network loading within simulation program (e.g. Neplan). Active power injection definition is based on appropriately scaled normalized load profile. Reactive power injection is calculated using previously defined active power injection and $\cos(\varphi_{MODEL})$ database. Mentioned database is defined with quantiles in a probabilistic manner. It allows for both deterministic *Q* calculation using median value of $\cos(\varphi)$ and also for stochastic network simulation using provided probability distribution function for iterative *Q* definition.

Additionally also network losses are available outputs from the model. They are calculated using provided normalized $P_{loss}(P_{pfTR})$ function.

Input load profiles are distinguished by the season (winter, summer, between) and type of day (weekday, Saturday, Sunday). It consequently leads to 9 different load profiles.

Consequently, also $\cos(\varphi_{MODEL})$ database has to be developed for the same periods of year as input load profiles are. However, based on our observations, the difference between Saturday and Sunday $\cos(\varphi)$ distribution is negligible, therefore we have merged these two days into a single subgroup.

2.2 Data source

Two sources of data were used to generate the $\cos(\varphi_{MODEL})$ database. Actually, we have created two different subgroups of the $\cos(\varphi_{MODEL})$: $\cos(\varphi_{LOAD})$ (without PVs) and $\cos(\varphi_{PV})$ (with PVs) database. Sources of data used for their definition are measurements and simulations respectively.

2.3 Measurements

The cos(φ_{LOAD}) is based on the field measurements of the real MV/LV transformers supplying the real LV grid. Three different subgroups were defined, based on the type of supplied customers, connected to LV grid: residential, commercial and mixed (mixture of both). Additionaly, also a distinction between three different seasons and two type of days was made, which leads to 18 different hourly cos(φ_{LOAD}) databases.

2.4 Simulations

In order to generate realistic cos(φ_{PV}) database and to properly describe the losses for the built model, we did as follows:

I. We have modeled Köstendorf LV network in detail from the provided Neplan data. Load allocation factor for each consumer in this network was also extracted from the Neplan data.
II. For load active power definition we rely on provided customer and transformer measurements from Köstendorf network. Firstly, we have subtracted measured aggregated PV generation and EV consumption of the whole network from the real power injection measurements of the MV/LV Köstendorf transformer. Obtained values were then used for probabilistic transformer active power injection database generation. LV load reactive power is defined using the previously mentioned cos(φ_{LOAD}) database
III. From the PV measurements we have generated PV generation database. Since the reactive power injection of PVs is voltage dependent, we were also provided with the $Q_{PV}(U)$ characteristic. Both databases are further segmented in the same manner as previously mentioned cos(φ_{LOAD}) database.
IV. We implemented Monte Carlo simulations of the modeled network using the databases defined in the previous steps. This step is further explained later in the Simulation details subchapter.
V. Based on the calculated values of interest probabilistic cos(φ_{LOAD}) database was obtained. From the calculated network losses also network losses are defined stochastically, on a per-hour basis, as a rate of consumption of the network loads. It is later used for the model losses calculation.
VI. Now the injection of active and reactive power together with the losses of the network can be defined based on the network power profile and time of the year using cos(φ_{PV}) and losses databases from previous step.

2.5 Simulation details

Following is detailed explanation of the Monte Carlo (MC) simulations from the step IV nad.
Altogether 100 MC iterations of a daily load flow simulation in a 5-minute resolution was performed. Presented are the results for the summer, when the PV generation reaches its peak. It consequently results in a voltage rise in the LV network. Aggregated nominal power of all installed PVs in the network equals 100% of the transformer nominal power.
MV/LV transformer with rated power of 250 kVA is supplied from the grid which provides constant supply voltage to its MV winding at 1.03 pu.
The voltages of the network buses during the simulation are presented in Figure 2. There is a significant voltage rise during the peak PV generation.

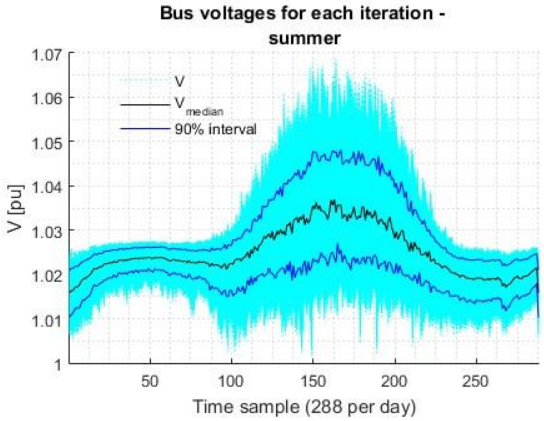


Figure 2: Network voltages

Provided are the results of workday simulation in the summer season. The transformer active power injection, cos(φ_{LOAD}) and PV generation database is chosen in accordance with the chosen simulation period. Three different cos(φ_{LOAD}) databases were defined separately, taking into consideration the type of supplied consumers, for each season and day type. These groups are: households, commercial and mix of both of them.
Aggregated active power of all the LV network loads profile for each iteration $P_{\sum loads\text{-}i}$ is defined from the database. It is also presented in Figure 3, together with its median profile and upper and lower 90% boundaries. Positive power represents power flow into the supplied LV network.

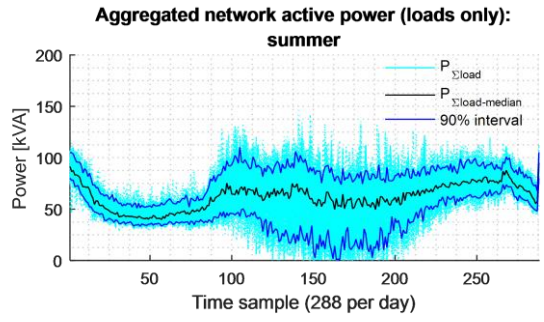


Figure 3: Aggregated active power of LV network loads

Reactive power injection of each PV is voltage dependent. Its dependency is shown in Figure 4.

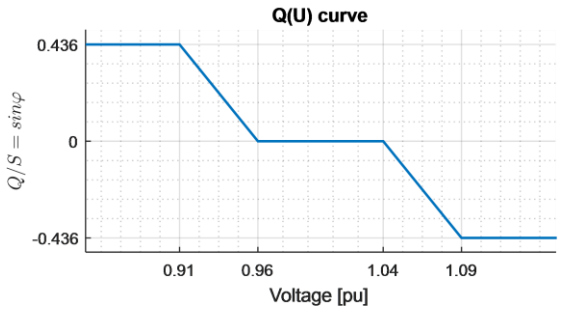


Figure 4: Q(U) characteristic

2.6 Simulation outputs

At the end of the last MC iteration, the power flow through the MV/LV transformer is available for the whole simulation (all MC simulations). Active and reactive power flow of the network is presented in Figure 5, together with its median and upper/lower 90% boundaries. Positive values mean power flowing into the LV network, whilst the negative ones represent the power flowing out from the network, into the MV distribution network.

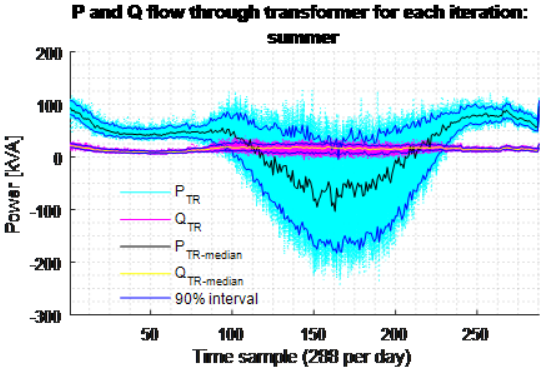


Figure 5: Power flow through MV/LV transformer

Definition of the cos(φPV) database is of main goal of these simulations, as also mentioned before. This database is derived from the simulated P_{pfTR} and Q_{pfTR} values presented in Figure 5. Due to the high penetration of the PVs in the simulated LV network there is a change in the P_{pfTR} during the day in terms of its direction. On the other hand, the flow of the Q_{pfTR} retains its direction through the whole day. Consequently, when the P_{pfTR} is opposite to the Q_{pfTR} the cos(φPV) turns out to be a negative value. However the direction, cos(φPV) is close to the 1 in one case and close to -1 in the other. Therefore we decided to present the cos(φPV) with a bit unusual approach as can be seen in Figure 6. Presented is a boxplot of values, where values bellow 1 represent the positive cos(φPV), while the values above 1 represent the negative cos(φPV).

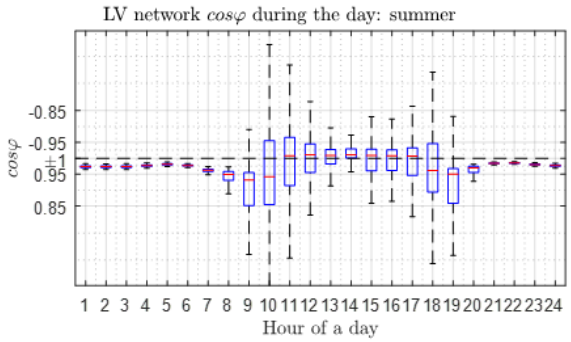


Figure 6: Boxplot of cos(φPV)

Besides the MV/LV transformer power flow data, the network losses measurements are also of interest in our case. These measurements serve as a basis for developed model loses calculation. Network losses, for each iteration are presented in Figure 7

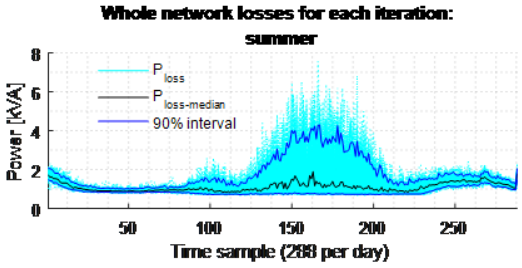


Figure 7: LV network losses (including the transformer losses)

These measurements are then used for investigation of relation between the power flow through the transformer and the losses in the network. The relation in absolute values is shown in Figure 8.

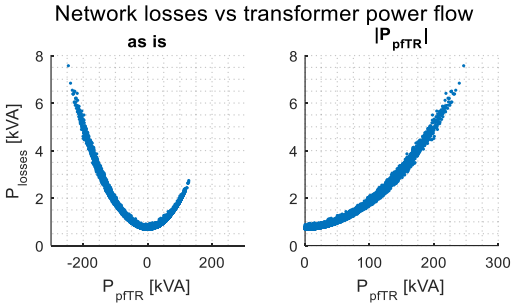


Figure 8: $P_{loss}(P_{pfTR})$ presented with original measurements (left) and with absolute P_{pfTR} values (right)

In order to generalize the relation to different LV networks, the values have to be normalized. Losses are normalized with the aggregated consumption of active power of all network loads only (PV are not taken into account) at each iteration. The variance of values is a bit higher during the PV generation period as can be seen in the Figure 9.

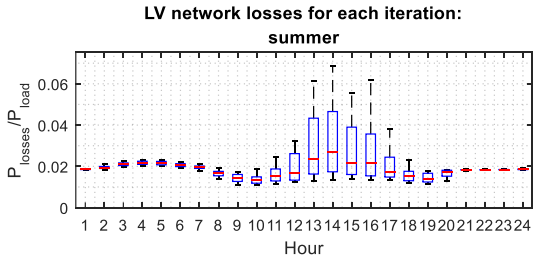


Figure 9: Normalized LV network losses boxplot

For validation of the used method for the losses calculation, histograms of the measured and the calculated network losses are presented in Figure 10. Similarity of the shown histograms is obvious. There is however some discrepancy, which is a consequence of relating the losses only to the consumption of the network (without PVs). The discrepancy can be observed also from the Figure 11, where the daily network losses profile comparisons are shown. Losses calculation is based solely on the network load consumption, disregarding the PVs generation. However, one can see, that mean values fit perfectly.

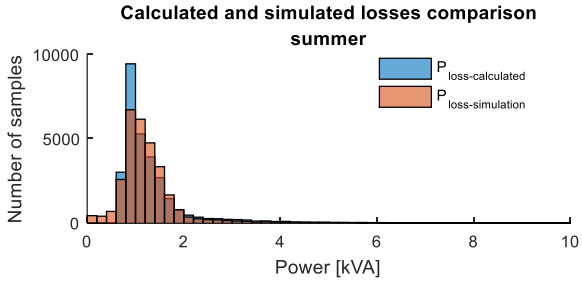


Figure 10: Network losses: histogram of calculated and simulated values

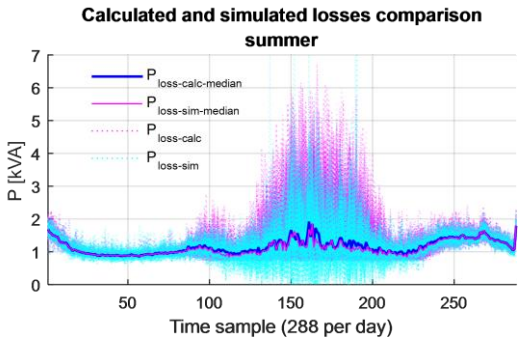


Figure 11 Network losses: daily plot of calculated and simulated values

3 Results

Following is the graphical presentation of the cos(φ_{MODEL}) for the summer season. Due to the reasons of straightforwardness only cos(φ_{MODEL}) of 3 different consumption types are presented (households-no PV, commercial-no PV, mixed-100% PV) each for both weekday and weekend.

Figure 12 shows a per-hour boxplot of cos(φ_{MODEL}) for the LV network supplying a mix of households and commercial users, for the summer season: weekday and weekend. PV penetration rate in terms of MV/LV transformer nominal power equals 100%. This is actually a cos(φ_{PV}) database, which is a result of the network simulation described above. One can see, that cos(φ_{MODEL}) variance is higher, during the PV generation period of the day. Otherwise, there is no significant difference between the two distributions.

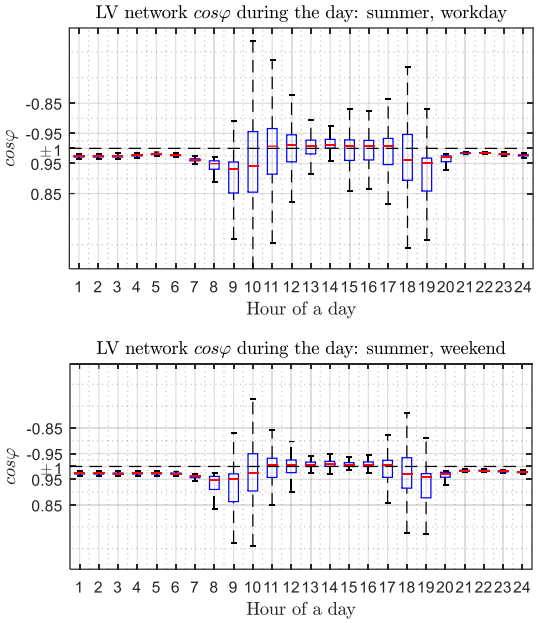


Figure 12: Boxplot of cos(φ_{MODEL}) for the summer season, workday (top) and weekend (bottom)

Literature

[1] Fingrid, 'Supply of Reactive Power and Maintenance of Reactive Power'.

[2] 'DeCAS project'. [Online]. Available: decas-project.eu. [Accessed: 05-Dec-2017].

[3] EC, network code on requirements for grid connection of generators. 2016.

[4] J. Xie, C. Liang, and Y. Xiao, 'Reactive Power Optimization for Distribution Network Based on Distributed Random Gradient-Free Algorithm', Energies, vol. 11, no. 3, p. 534, Mar. 2018.

Artificial neural networks based short-term forecasting of daily electrical energy consumption: an approach using consumption data subsets

Domen Kodrič, Gašper Podobnik, Martin Žukovec, Jan Gašpar

University of Ljubljana, Faculty of Electrical Engineering
E-mail: dk1824@student.uni-lj.si

Abstract. *The purpose of this paper is to present a possible approach towards artificial neural networks based short-term predictions of daily electrical energy consumption (EEC) in Slovenia and its practical application. Since power consumption depends on complex non-linear relationships between several influential variables, artificial neural networks are commonly used in creating EEC forecast models. A characteristic of power consumption is a superposition of multiple time-based patterns, which can be recognised, and through classification of past EEC data, data conforming to patterns can be grouped, and specific submodels can be designed. We show that the chosen constructed data subsets contain less noise, and that by using submodels built upon these subsets, we produce predictions that are more accurate than those made using a single dataset approach.*

1 Introduction

Electrical energy consumption (EEC) is not constant. It changes gradually through years, differs between seasons of the year, fluctuates from day to day (workdays, weekends), and can steeply increase and decrease from hour to hour. Power system operators, electrical energy traders, as well as electrical energy producers all benefit from accurate future EEC prediction.

In this paper, we describe an artificial neural networks (ANN) based model for producing short-term (1-3 days) forecasts of daily electrical energy consumption in Slovenia. We use past data on electrical energy consumption, past seasonal data, and past meteorological data. A characteristic of the power demand through time is a superposition of multiple patterns that depend on classifications such as day of the week, day of the year, holiday etc. An overview of the existing literature on (ANN based) short-term EEC forecast considering these factors offers two basic approaches to the problem. Approach A (example: [4]) involves assigning values for each variable (for example 1-7 for days of the week), and creating a single ANN. Approach B (by Lee already in 1992 [2]) employs classifying the data based on the aforementioned factors, and designing multiple ANN models, each assigned to produce forecasts for a specific type of day. It appears that in similar problems on short-term EEC forecast, approach A prevails [1, 3, 4, 5]. In the paper, we employ approach B: we dissect the data into subsets, and show improvement in prediction accuracy over a model built with a single dataset. Using two resulting specific models (considering only January workday-days/weekend-days), we produce predictions of daily consumption for 18 days of January 2018, with results comparable to publically available predictions.

2 The data

Data on the energy flow from the transmission network in hourly intervals from Jan 1st 2016 onwards are available on the website of the Slovenian electricity transmission system operator (ELES) [7]. The company has, upon request, generously provided us the data for years 2010-2017 by e-mail [8] (69600 hourly EEC values).

The data represent the electrical energy flow from the transmission network into distribution networks and to directly connected clients; this includes losses in the transmission network [8]. We therefore forecast the electrical energy flow from the transmission network, and not the final electrical energy consumption. The two quantities are, however, closely related, and the term electrical energy consumption (EEC) will be used when describing correlations between variables for intuitive reasons.

Although the research primarily focused on the effect of date variables, we have considered the impact of meteorological variables in order to produce more accurate forecasts. We acquired meteorological data from the web servers of the Slovenian Environmental Agency [9]. These encompassed the average, minimal and maximal values of air pressure, temperature and relative humidity, average values of diffuse and total solar irradiation, and total precipitation, in hourly intervals. We acquired the data from the automatic station Ljubljana Bežigrad, because we estimated that due to its location (in the centre of the country) and a large population living in its vicinity would provide the most relevant meteorological data.

Empty data cells were filled using linear interpolation.

3 Data analysis

We avoided using too many meteorological variables, and only considered the most influential ones. In order to get a rough estimate of the importance of meteorological variables, we trained with Matlab R2016b a simple regression tree, which offers a straightforward insight into predictor importance. We trained regression trees first using hourly and then daily values of all meteorological variables as inputs, while the output of the regression tree was the value of EEC.

The air temperature consistently proved itself as the variable with the most influence on EEC. With significantly lower importance, total radiation and relative humidity followed; however, both of these are already related to the air temperature themselves, and we therefore did not include them. Instead, we added a variable *dT*, the change in temperature from the previous hour/day, to observe the effect it has on the change in EEC relative to the previous hour/day, as in human perception, the *change* of temperature also plays a role.

We then analysed the whole EEC dataset visually over time periods of different magnitude. We discerned a superposition of multiple patterns that the power consumption follows through time. Plotted daily consumption values against the days of a year (DoY) offer an insight into the difference between seasons. Figure 1 shows the difference in consumption between seasons of the year.

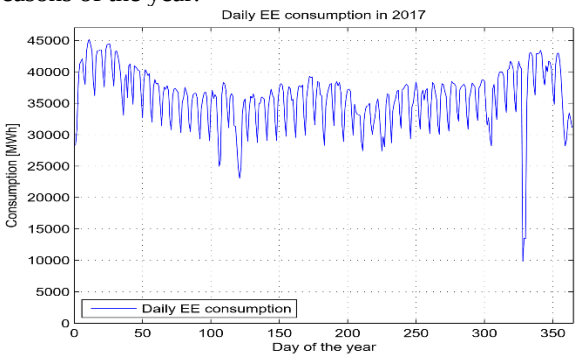


Figure 1. Daily EEC in 2017.

We can observe greater consumption values in the winter months. The second, smaller peak is seen around DoY 150-190 (the month of June and beginning of July). Noticeable deviations can be noted in the days around day 105 (Easter holidays), day 120 (May Day), and the substantial decrease in consumption around day 325 (a consequence of a weather induced power outage). These observations are referenced in Chapters 3.1 and 4.

The following Figure 2 represents the daily EEC through seven consecutive weeks in Oct and Nov 2017.

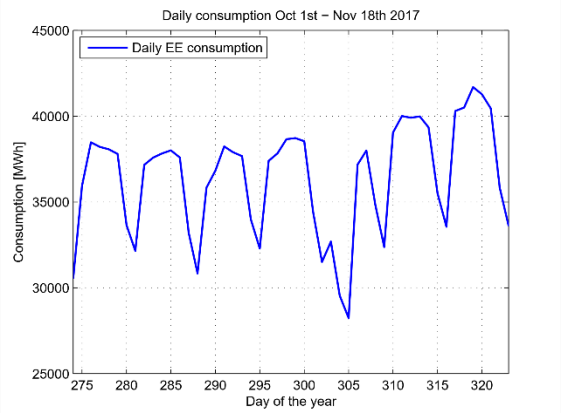


Figure 2. Daily EEC through 7 weeks of 2017.

We notice the periodic weekly pattern of EEC. Highs represent middle-of-the-week workdays (Tuesday - Thursday), while lows represent Sundays. A deviation between day 301 and 304 can be attributed to two bank holidays and school holidays in that period.

Another pattern can be observed in the recurring shape of the hourly EEC curve with two distinct peaks in the morning and evening hours. Figure 3 shows the hourly values of EEC on Thursday, Oct 12th 2017.

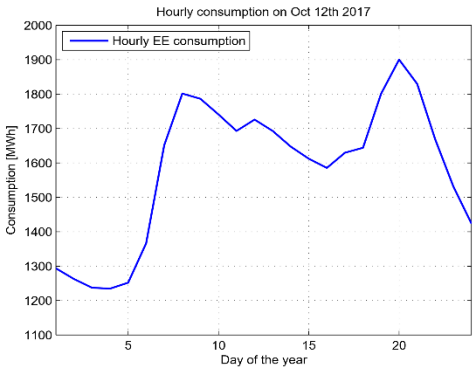


Figure 3. Hourly EEC on Oct 12th 2017.

3.1 Data classification

Considering the recurring patterns recognised in the previous chapter, we deduced that date variables (such as DoY, day of the week (DoW), and bank holidays) would need to be taken into consideration when creating a prediction model. In the following paragraphs, we describe an analysis of the effect that the day of the week has on the EEC pattern.

We used the hourly EEC data from Jan 1st 2010 to Dec 11th 2017. To determine the (dis)similarity of the pattern of the hourly EEC on different DoW, we calculated the standard deviation of the EEC for every hour of the same DoW, for all days of the week. To get an insight into normalized dispersion of the hourly values of EEC in different DoW, we calculated the coefficient of variation (CV) of the hourly EEC. (CV = σ/μ; σ represents standard deviation and μ represents the mean value of the EEC in a specific hour.)

Figure 4 shows the CV of hourly values of EEC on all Thursdays in the dataset.

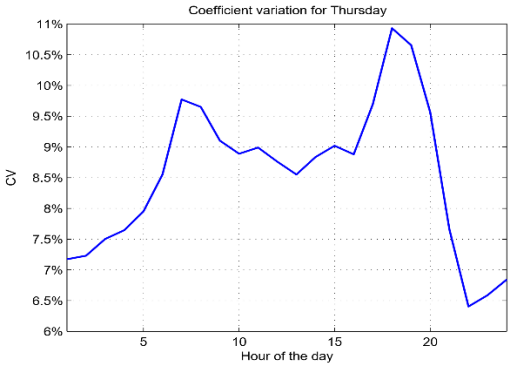


Figure 4. CV through an average Thursday.

A lower CV means a lower relative dispersion of hourly values, and demonstrates the degree of similarity of a day's consumption pattern. The peaks of CV in the

morning and evening show that the relative variance in EEC is greatest between the hours of 6-9 and 16-19 – periods that coincide with peaks in the average hourly EEC.

To evaluate the similarity of the consumption pattern between different days of the week, four charts are shown together in the following Figure 5. The blue, triangle-point curve represents the hourly coefficient of variation (CV) on Thursdays (same as Figure 4), the red, continuous curve represents the hourly CV on Saturdays, and the green, dashed curve represents the hourly CV of a combined set of Thursday and Saturday data (which gives one combined average around which data for both days are dispersed). Similarly, the black, continuous/dotted curve represents the hourly CV of a combined set of Thursday and Friday data.

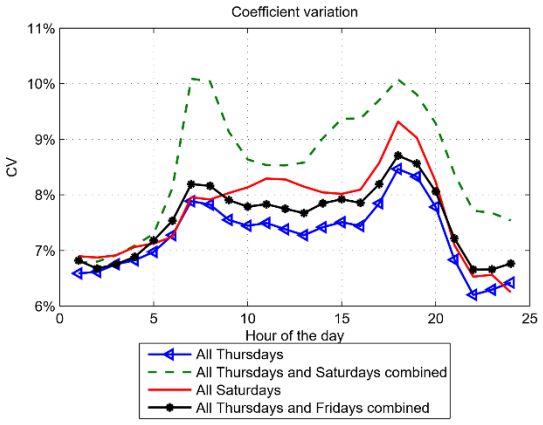


Figure 5. Comparison of CV's of different datasets.

Comparing the curves, the difference between CV's of the Thursday and the Thursday+Saturday datasets is the most obvious. While the CV of the Thursday dataset never exceeds 11 % (max 10.93 %, average 8.54 %), and the CV of the Saturday dataset peaks at 12.64 % (averages 9.39 %), the CV of the combined Thu+Sat is noticeably higher and peaks at 14.18 % (averages 11.04 %). CV of the Thu+Fri dataset, (max 11.41 %, average 9.13 %) presents a smaller increase from the Thu value.

The differences in CV of the datasets show that the consumption curve pattern is less predictable, when the datasets of two different days of the week are combined. Furthermore, the similarity between patterns of a workday and a weekend day (Thursday+Saturday) is substantially lower than the similarity of two workdays (Thursday+Friday). Since combining datasets of different days of the week into one dataset evidently produces a dataset with more noise, it follows that it may be meaningful to produce different prediction submodels, trained on certain subsets of similar data, at the cost of less data in a subset.

Figure 6 is a visualisation of the similarity between all weekday datasets. The area of the circle represents the RMS value (in MWh) of the difference between respective hourly EEC values of two different DoW, for all 414 weeks (RMSd). Dissimilarity is greatest between Thu and Sun (RMSd = 319.0 MWh), and smallest between Wed and Thu (RMSd = 69.9 MWh).

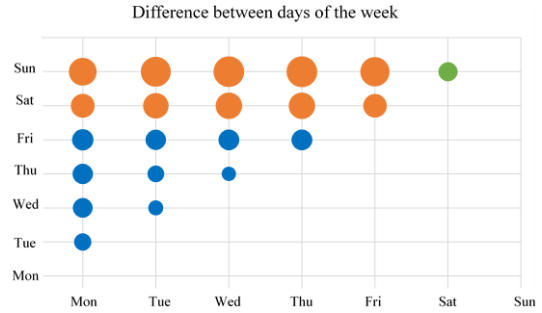


Figure 6. Comparison of (dis)similarity between days of the week.

Using a similar-days approach, Mandal [6] splits the days of the week into 4 classes (Monday, Saturday, Sunday, and weekdays comprising Tue-Fri). Regarding our dataset, we propose that if one were to account for Sat-Sun distinction (RMSd Sat-Sun = 123.3 MWh), it would follow to also account for at least the distinction between Tue and Fri (RMSd = 140.2 MWh), if not to take each DoW as an individual dataset (RMSd Tue-Mon = 102.4 MWh, RMSd Tue-Wed = 77.6 MWh, RMSd Tue-Thu = 95.6 MWh).

4 Devising a model for daily forecast

Mandal [6] uses a similar-days approach to produce 1-6 hours-ahead forecasts with a moving dataset, using data 65 days before the forecast day, and 65 days before and after the forecast day in the previous year. In our approach, we pre-determined the similar days, and our dataset (considering the analysis in Chapter 3) included the variable *day of the year* to account for seasonal similarity. Furthermore, we looked for similarity in all 8 years of data, and included *bank holidays* as an influential variable.

We first constructed a model using the entire dataset (daily EEC values from Jan 1st 2010 to Dec 11th 2017). Since the relation between the input variables and the output EEC value is non-linear, and the form of non-linearity is unknown, we have devised a model using artificial neural networks (ANN), an artificial intelligence method. We separated the data into a 'learning set' and a 'test set', and used the learning set to train in RapidMiner an ANN, which would produce the best dependency algorithm between the input (daily meteorological and date) variables and the output values of daily EEC. Input data of the 'test set' were then applied to the constructed ANN, and we evaluated the accuracy of the resulting predictions by comparing them to actual daily EEC values. Multiple combinations of the number of hidden layers and the number of neurons were tested, and minimization of the mean APE (absolute percentage error) was then used as the criterion in selecting the optimal values. The final ANN consisted of an input layer (6 neurons; one for each input variable), an output layer (one neuron for the output variable), and two hidden layers with 15 neurons each. Using the whole dataset, the mean absolute percentage error (APE) between predictions and actual daily EEC values was 3.6%. The maximum APE was 9.7 %.

4.1 Improvement of predictions using data subsets

We then trained multiple ANN's with different data subsets. We first trained an ANN using workday-day data only – we eliminated weekends and bank holidays. We also eliminated prediction-irrelevant datasets with atypically low EEC. The results showed an improvement in prediction accuracy compared to the general model – mean APE was 3.3 %, and maximum APE was 7.0 %. A model using only weekend data (without bank holidays) gave a mean APE of 3.7 %, maximum APE was 7.3 %.

Substantial improvements in prediction accuracy were reached when even stricter filters were applied, and forecasting models became more and more specific. A model using only workdays from March until May (days of the year 60 - 150) returned a mean APE of 2.5 %, maximum APE was 5.9 %. A model encompassing workdays from May to July (DoY 120 - 210) performed with mean APE of 1.9 %, maximum APE was 3.4 %.

5 A practical application of the model

We trained an ANN using only January workday-days, with no bank holidays and no prediction-irrelevant data. We produced a forecast (temperature predicted with [10]) of the EEC for the following day. Forecasts were made at midnight, with the previous day's actual EEC known. An analogous model was made for January weekends.

Table 1 shows our daily prediction along with the actual daily EEC. The prediction on the website of the Slovenian electricity transmission system operator (TSO) serves as a reference. Absolute percentage error (APE) describes the accuracy of predictions; a green background marks the lower of the errors. For three of the days in the time span, predictions were not made.

Table 1: Real predictions in January

Date (Jan)	Actual EEC [MWh]	Our forecast [MWh]	APE of our forecast	Forecast at the TSO's website [MWh] [7]	APE of the TSO's forecast[7]
8	37948	37695	0,67%	40434	6,55%
9	39543	38513	2,60%	39754	0,53%
10	39813	38725	2,73%	39646	0,42%
11	39304	39415	0,28%	39568	0,67%
12	39793	40226	1,09%	39889	0,24%
13	35653	/	/	35280	1,05%
14	33732	/	/	33663	0,20%
15	40701	40151	1,35%	39680	2,51%
16	41472	41359	0,27%	41106	0,88%
17	41612	40597	2,44%	42071	1,10%
18	41258	40931	0,79%	41257	0,00%
19	41701	40307	3,34%	41694	0,02%
20	37414	36960	1,21%	36358	2,82%
21	33933	35108	3,46%	34179	0,72%
22	40265	40432	0,41%	39819	1,11%
23	40274	42368	5,20%	40941	1,66%
24	40760	42491	4,25%	41635	2,15%
25	41422	41786	0,88%	41271	0,36%
26	40771	40355	1,02%	40792	0,05%
27	36838	/	/	36665	0,47%
28	34123	33882	0,71%	33596	1,54%

6 Conclusion

Through the analysis of past EEC data and through practical examples, we showed that when using artificial neural networks, the creation of multiple submodels with appropriate data subsets results in an improvement of prediction accuracy. We devised such a submodel for a specific timeframe (the month of January) and produced practical predictions of the daily EEC.

When predicting hourly EEC, the selection of subsets could be made similarly, and additional classification could be introduced by periods of the day (e.g. peaks of CV during hours of high demand), or by days with similar meteorological conditions. Further specializing the submodels reduces the quantity of learning data in the submodel's dataset – an optimum has to be found between increasing the number of submodels and the resulting decrease in learning data per submodel.

Acknowledgement

The authors kindly thank the Slovenian TSO (ELES, Ltd.) for providing us past EEC data.

References

[1] I. Chernykh, D. Chechushkov, and T. Panikovskaya. "The prediction of electric energy consumption using an artificial neural network." *Energy Production and Management in the 21st Century: The Quest for Sustainable Energy* 190 (2014): 1109.

[2] K. Y Lee, Y. T. Cha, and J. H. Park. "Short-term load forecasting using an artificial neural network." *IEEE Transactions on Power Systems* 7.1 (1992): 124-132.

[3] M. Buhari, and S. S. Adamu. "Short-term load forecasting using artificial neural network." *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 2012.

[4] L. Hernández et al. "Artificial neural network for short-term load forecasting in distribution systems." *Energies* 7.3 (2014): 1576-1598.

[5] K. Gajowniczek, R. Nafkha, and T. Ząbkowski. "Electricity peak demand classification with artificial neural networks." *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017.

[6] P. Mandal et al. "A neural network based several-hour-ahead electric load forecasting using similar days approach." *International Journal of Electrical Power & Energy Systems* 28.6 (2006): 367-373.

[7] ELES d.o.o. »Prevzem in proizvodnja,« eles.si [online] Available: https://www.eles.si/prevzem-in-proizvodnja [Accessed 30. 1. 2018].

[8] Info ELES, »Poizvedba po podatkih za raziskovalne namene.« Personal email (Dec 21th 2017).

[9] ARSO METEO, »Uradna vremenska napoved za Slovenijo – Vreme podrobneje,« arso.gov.si [Online] Available: http://meteo.arso.gov.si/met/sl/app/webmet/ [Accessed Dec 21th 2018].

[10] The Weather Channel, »Vremenska napoved in pogoji« weather.com, [Online] Available: https://weather.com/sl-SI/vreme/danes/l/SIXX0002:1:SI [Accessed Jan 27th 2018].

Asinhronski generator

Jaša Vid Meh Peer, Simon Strmšnik, Nataša Meh Peer, Matjaž Žerak

Elektro in računalniška šola, Šolski center Velenje, Trg mladosti 3, Velenje
E-pošta: jasa.meh.peer@gmail.com, simon.strmsnik@gmail.com, natasa.mehpeer@scv.si, matjaz.zerak@scv.si

Abstract. When something goes wrong and the power goes out, we are lost in darkness. The aim of our research was to create a cheaper induction generator, which would be user friendly and made out of a motor we have at our homes.

In the research part the induction generator was stirred with capacitors, some load was added to see how the current will affect the voltage level and the frequency that it generates. Adding more load to the generator causes more current to flow, that affects the voltage to drop, if we don't add any capacitors. By adding capacitors we can regulate the voltage level of the strained generator to our liking. We determined the borders of the regulated voltage level and the stage of capacitors for adjusting it.

We used a system that can measure the voltage level and regulate it by adding or subtracting capacitors. It consists of a voltmeter connected to an Arduino board with specified outputs. We prepared 5 stages of capacitors connected in a triangle. For adding and subtracting capacitors, we used 5 contactors and designed the voltmeter from a transformer, AC-DC converter and a voltage divider made specifically for the Arduino analogue input. The programme that ran on the Arduino board functioned on the principle that used the analogue input signal to measure the voltage level of the generator and regulate the number of capacitors added with Arduino's output pins connected to the contactors. Our system lets us successfully generate electrical energy and adding loads that strain the generator, without letting the voltage level drop.

1 Uvod

Večino pretvorb iz mehanske v električno energijo opravljajo sinhronski generatorji (99 %), tudi velike elektrarne uporabljajo sinhronske generatorje, ki dosegajo moči do nekaj GVA (GW).

Asinhronski generatorji se zaradi slabšega izkoristka redko uporabljajo. Zaradi nižje cene in poenostavljenih regulacijskih in zaščitnih naprav, pa se uporabljajo v malih hidroelektrarnah do moči nekaj deset kilovatov, zato so primerni tudi za napajanje manjših zasebnih porabnikov.

Predstavili bomo uporabo asinhronskega stroja v otočnem režimu obratovanja in njegov način izdelave iz komponent, ki so jih imamo pri roki (asinhronski motorji kot generatorji in traktor kot pogonski stroj), kar je pomembno ob izpadu električne energije ob vremenskih ujmah za kmetije.

2 Asinhronski stroj kot generator

Generatorsko delovanje asinhronskega motorja dosežemo z izpolnitvijo dveh pogojev: rotor asinhronskega stroja moramo zavrteti in statorskim tuljavam zagotoviti magnetilni tok. Ker asinhronski stroj kot generator (ASG) sam ne more proizvajati magnetilnega toka in induktivne komponente bremenskega toka, mu to priskrbimo s kondenzatorji, ki so priključeni na statorske sponke. [4]

2.1 Samostojno obratovanje ASG

Kadar je asinhronski stroj ločen od ostalih delov električnega omrežja, in deluje samostojno, jalovo energijo pridobimo iz kondenzatorskih baterij, ki so vezane na statorske sponke generatorja. Pogonski stroj pa zagotavlja vrtilni moment na osi generatorja.

Samostojno obratovanje dosežemo tako, da generator zavrtimo, na statorske sponke pa priključimo trifazno kapacitivno breme – vzbujalne kondenzatorje. [5]

Statorsko trifazno navitje povežemo v vezavo zvezda, saj se tako izognemo močnim tokovom trojne frekvence, ki krožijo v sklenjenem trikotnem navitju, kamor vežemo vzbujalne kondenzatorje. Kondenzatorje ustrezne kapacitivnosti C povežemo v vezavo trikot in tak generator prenese tudi nesimetrične obremenitve. [1]

2.2 Karakteristika napetosti in toka (*UI*) generatorja

UI karakteristika statorskega električnega kroga nam pokaže odvisnost statorskega toka prostega teka od priključne napetosti. Izmerimo jo lahko pri motorskem obratovanju. Pri nižji napetosti je tok majhen. Statorski tok se nesorazmerno poveča, je nelinearen, ko napetost preide koleno *UI* karakteristike.

Vsi električni stroji delujejo pod kolenom *UI* karakteristike, v področju visoke impedance, ko železno jedro še ni prešlo v zasičenje. K *UI* karakteristiki generatorja vrišemo še *UI* karakteristiko kondenzatorjev, ki je linearna, saj kapacitivna upornost ni odvisna od napetosti ali toka, temveč od frekvence in kapacitivnosti kondenzatorjev, kar zapišemo z naslednjim izrazom:

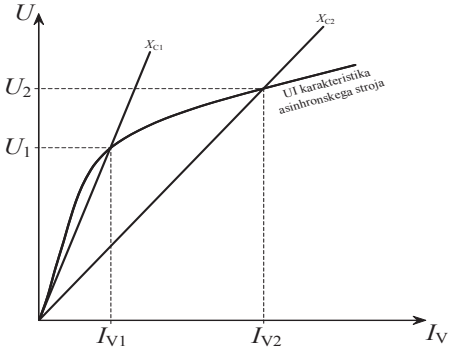
$$X_C = \frac{1}{2 \cdot \pi \cdot f \cdot C}$$

Področje, kjer je magnetilna karakteristika višja od kondenzatorjeve, je področje samovzbujanja (slika 1). Ko vzporedno na sponke asinhronskega generatorja priključimo kondenzatorje, bo zaradi remanentnega magnetizma prišlo do postopka samovzbujanja. Nizka napetost, ki je posledica remanentnega magnetizma, bo čez kondenzatorje in statorsko navitje pognala majhen tok, ki bo povečal skupno magnetno polje, kar vodi k še višji indukciji.

Generator bo, pri konstantnem številu vrtljajev in konstantni kapacitivnosti, obratoval v točki presečišča

magnetilne in *UI* karakteristike kondenzatorjev s konstantno napetostjo, frekvenco in tokom.

S spremembo kapacitivnosti kondenzatorjev se spremeni naklon *UI* karakteristike kondenzatorja in s tem točka presečišča z magnetilno karakteristiko motorja. S spreminjanjem kapacitivnosti spreminjamo napetost in kapacitivni vzbujalni tok I_V čez stator in kondenzatorje. Pri napetostih višjih od kolenske bo tekel nesorazmerno večji vzbujalni tok. Pri večji kapacitivnosti bo stekel večji vzbujalni tok in inducirana napetost se bo povišala po *UI* karakteristiki motorja. Zmanjšanje kapacitivnosti kondenzatorjev ima obraten učinek − zmanjšanje vzbujalnega toka in inducirane napetosti in s tem napetosti na sponkah generatorja. [1]



Slika 1. Samovzbujanje ASG [1]

2.3 Obremenitev ASG pri samostojnem obratovanju

Asinhronski generator obremenimo tako, da na statorske priključne sponke s kondenzatorji vzbujenega generatorja, vzporedno k vzbujalnim kondenzatorjem priključimo električni porabnik – breme.

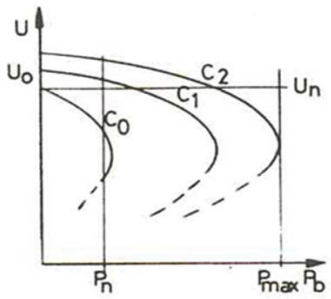
Po statorskem navitju stečeta, poleg kapacitivne komponente izmeničnega toka, še tok porabnika, ki se vektorsko prišteje k vzbujalnemu toku kondenzatorjev. V primeru delovne obremenitve je statorski tok vsota kapacitivne in delovne komponente, kar ponazarja naslednji izraz:

$$I = \sqrt{I_d^2 + I_c^2}$$

Povečanje statorskega toka zaradi obremenitve generatorja, povzroči znižanje napetosti generatorja. Pri konstantni vzbujalni kapacitivnosti in toku se pri višanju obremenitve napetost generatorja niža. Manjšanje napetosti je odvisno tudi od faktorja delavnosti porabnikov, saj je padanje napetosti večje, čim bolj je breme induktivno. [1]

Karakteristika (slika 2) kaže tudi na to, da generatorja ne moremo obremeniti preko maksimalne moči, pri kateri izgubi sposobnost samovzbujanja in se razmagneti. [1]

Z zmanjšanjem hitrosti vrtenja rotorja se manjša inducirana napetost, saj se zaradi delovne obremenitve in posledic zavornega navora generatorja zmanjša vrtilna hitrost pogonskega stroja (turbine), kar vodi v znižanje inducirane napetosti generatorja.



Slika 2: Odvisnost napetosti od moči [1]

2.4 Frekvenca napetosti ASG

V prostem teku je slip skoraj nič in frekvenca napetosti je določena z naslednjim izrazom:

$$f_0 = p \cdot n$$

Pri obremenitvi generatorja se slip poveča in frekvenca napetosti zniža, kar prikazuje naslednji izraz:

$$f = f_0/(1-s)$$

Rotor prehiteva magnetno polje. Vrtilna hitrost rotorja je večja kot sinhronsko vrtilna hitrost magnetnega polja in frekvenca v statorskem navitju inducirane napetosti (napetost generatorja) se zniža. Če se vrtljaji pri delovni obremenitvi generatorja zmanjšajo, se frekvenca dodatno zniža.

3 Materiali in metode dela

Raziskavo smo razdelili na raziskovalni in projektni del.

V raziskovalnem delu smo preizkusili samostojno obratovanje asinhronskega generatorja v prostem teku in v obremenjenem stanju.

Uporabili smo asinhronski motor iz šolskega laboratorija, ki je bil izdelan za motorsko obratovanje in ima nazivne podatke: P_N = 1,1 kW, U_N = 380 V (zvezda), f = 50 Hz, I_N = 3,7 A, $cos\varphi$ = 0,67, n_N = 1360 vrt/min.

Stroj ima dva para polov in sinhronsko vrtilno hitrost magnetnega polja (1500 vrt/min). Poganjali ga bomo s sinhronskim motorjem, ki bo zagotovil konstantno vrtilno hitrost neodvisno od obremenitve. Tudi sinhronski motor je štiripolni in se pri priklopu na omrežno napetost frekvence 50 Hz vrti s sinhronskim številom vrtljajev (1500 vrt/min). Merilna oprema in materiali:

Pogonski stroj:	U_N = 400 V
Sinhronski motor:	I_N = 2,4 A
	P_N = 1 kW
	$cos\varphi$ = 0,8
	n = 1500 vrt/min
Dve ohmski bremeni:	3x1000 Ω, 0,57 A
	3x3300 Ω, 0,31 A
Merilni instrumenti:	Finest 707 true RMS multimeter. VOLTCRAFT VC 830

Victor VC890C+

Kondenzatorji: 40 µF ,20 µF, 10 µF, 5 µF, 2,5 µF (400 V)

Iz dobljenih rezultatov smo v projektnem delu načrtovali in izdelali regulacijski sistem napetosti generatorja, ki bi zagotavljal napajanje enostavnih porabnikov pri konstantnem številu vrtljajev.

4 Rezultati raziskovalnega in projektnega dela

V raziskovalnem delu smo opravili preizkus samovzbujanja generatorja, njegove obremenitve v samostojnem obratovanju in preizkus obremenitve ASG z napetostjo, ki smo jo regulirali in nazadnje še določali stopnjo kapacitivnosti za regulacijo napetosti. Ugotovili smo, da laboratorijski stroj ne bo primeren za napajanje porabnikov pri nazivni napetosti 400 V, saj bi bil takrat že preobremenjen z jalovo komponento toka. In sprejeli kompromis, da bomo regulirali fazno napetost na 180 V.

Na podlagi rezultatov pridobljenih iz raziskovalnega dela naloge smo se lotili izdelave regulatorja napetosti asinhronskega generatorja pri samostojnem obratovanju.

4.1 Regulacijski sistem in njegove komponente

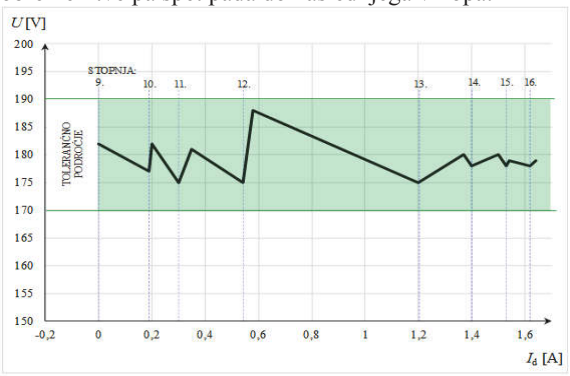
Spreminjanje kapacitivnosti je bilo potrebno avtomatizirati. Izvedli smo avtomatsko meritev napetosti in z mikrokrmilniško ploščico Arduino informacijo obdelali in nato vplivali na vklop/izklop kapacitivnosti (slika 3).

Naš generator smo izpostavili meritvam napetosti in izpisali program, ki bo reguliral kapacitivnost in vklapljal kontaktorje. Za izvršne člene smo uporabili releje in kontaktorje (aktuatorje).

4.2 Delovanje agregata

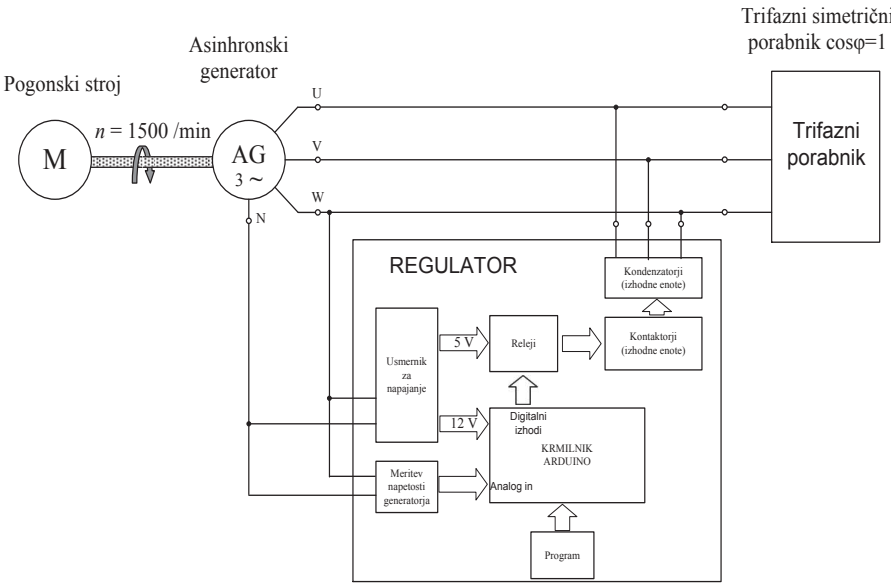
Delovanje agregata smo ocenjevali z rezultati naslednjih preizkusov:

- Preizkus samovzbujanja agregata smo opravili pri vsakem zagonu oziroma višanju števila vrtljajev. Agregat se je pri vseh poskusih uspešno vzbudil. Razlog je dovolj visoka skupna kapacitivnost kondenzatorjev (31. stopnje) v trenutku zagona.
- Vzbujeni generator smo obremenili z delovnim porabnikom in opazovali spreminjanje napetosti na njegovih sponkah ter preklapljanje stopenj kapacitivnosti. Rezultati meritve (slika 4) kažejo, da napetost ostane znotraj tolerančnega področja, kljub večanju delovnega toka. Pri vsakem vklopu naslednje stopnje kapacitivnosti se napetost zviša, z višanjem obremenitve pa spet pada do naslednjega vklopa.



Slika 4: Preizkus delovanja regulacijskega sistema

- Preizkus stopnične obremenitve samovzbudnega ASG je pokazal dobro odzivnost regulatorja. Napetost generatorja se je pri stopnični obremenitvi za kratek čas znižala pod napetost tolerančnega območja, vendar se je zaradi odziva regulatorja v kratkem času vrnila v tolerančno območje.
- Velik zagonski tok je po priklopu asinhronskega motorja na sponke ASG povzročil velik padec napetosti generatorja. Vendar je naš regulator pravočasno povečal kapacitivnost in stabiliziral napetost generatorja na regulirano vrednost. Neobremenjen asinhronski motor je kot RL porabnik obremenil ASG z delovno močjo 107



Slika 3: Blok shema regulacijskega sistema

W in jalovo močjo 635 VAr. Ker je skupna kapacitivnost regulacijskih kondenzatorjev, ki smo jih vezali v vezavo zvezda, zaradi enostavnejše izvedbe meritev in spreminjanja kapacitivnosti, dovolj velika, je pri obremenitvi generatorja z asinhronskim motorjem regulator s kapacitivno močjo kondenzatorjev napajal ASG in breme – asinhronski motor. ASG pa je zagotovil bremenu delovno moč.

- Opravili smo tudi preizkus pri enofazni obremenitvi, kjer smo ugotavljali, kako povečanje obremenitve ene faze vpliva na fazne napetosti generatorja. Obremenili smo samo eno fazo ASG z ohmskim porabnikom in merili vse tri fazne napetosti med priključnimi sponkami in nevtralno točko statorskih navitij. Napetost obremenjene faze se je znižala glede na fazni napetosti neobremenjenih faz.

5 Razprava in sklep

Izdelali smo regulator za regulacijo napetosti ASG.

Uvodoma izražena želja je bila možnost uporabe regulatorja na domačih asinhronskih motorjih gnanih z domačimi pogoni (kardani), ki bi omogočili enostavno predelavo motorjev v ASG in bi bili sposobni napajati električne porabnike pri stabilnih napetostnih pogojih (omejeno nihanje napetosti).

Izdelan regulator napetosti asinhronskega generatorja je podrejen kompromisom, ki smo jih sprejeli zaradi specifičnih lastnosti laboratorijskega stroja, ki smo ga uporabili.

Določili smo napetostne meje samostojnega generatorskega obratovanja obravnavanega stroja:

Spodnja napetostna meja obratovanja je določena s sposobnostjo samovzbujanja. Ugotovili smo, da se ASG ni sposoben vzbuditi pod mejno napetost v točki samovzbujanja na *UI* karakteristiki (slika 1). Pod to napetostjo generator ni sposoben generirati napetosti v samostojnem obratovanju.

Zgornja napetostna meja je določena z obremenljivostjo stroja oziroma njegovim nazivnim tokom. Pri napetostih, ko teče vzbujalni tok enake ali večje vrednosti od nazivnega toka navitja stroja, generator ne more več napajati porabnikov z delovnim tokom, kajti skupni tok bi bil večji od nazivnega.

S spreminjanjem kapacitivnosti je možno regulirati napetost obremenjenega ASG pri samostojnem obratovanju s konstantnim številom vrtljajev.

Spoznanje, da lahko ASG pri nižji regulirani napetosti napaja porabnike z večjo delovno močjo, je svojevrsten paradoks. Elektrotehniška teorija pravi, da je moč sorazmerna z napetostjo in tokom, torej bi pri višji napetosti moral biti generator sposoben napajati porabnike z večjo močjo kot pri nižji napetosti.

Generator je samovzbuden takrat, ko rotor zavrtimo in ne potrebuje zunanjega vira napetosti za začetek delovanja. Preizkus delovanja regulatorja na realnem asinhronskem generatorju je pokazal odlične sposobnosti avtonomnega obratovanja.

Zaradi dovolj visoke kapacitivnosti se je generator pri povišanju števila vrtljajev vedno, brez izjem, samovzbudil. Hkrati je napetost generatorja začela preko napajalnega usmernika napajati regulator in nanj priključene releje, zato je regulator začel preko relejev in kontaktorjev izklapljati previsoko kapacitivnost in napetost se je stabilizirala znotraj določenega tolerančnega področja $U \pm 10$ V.

Dosežki in ugotovitve raziskovanja so:

- Napetost obremenjenega ASG lahko uspešno reguliramo s spremembo kapacitivnosti.
- Nazivna napetost asinhronskega generatorja je manjša od nazivne napetosti pri motorskem obratovanju.
- ASG ima pri nižjih reguliranih napetostih večjo moč kot pri višjih napetostih.
- Izdelali smo regulator napetosti ASG s stopenjskim spreminjanjem kapacitivnosti.

Izboljšave izdelanega regulatorja so možne v stabilnosti delovanja, zaradi zmanjšanja števila preklopov kapacitivnih stopenj in izboljšanja programske kode, kjer bi bila sprememba kapacitivne stopnje odvisna od velikosti odstopanja napetosti od željene vrednosti. To bo tema naše raziskave v prihodnje.

Literatura

[1] Avčin, F., Jereb, P.: Preizkušanje električnih strojev in njihove lastnosti. Tehniška založba Slovenije, 1983 Ljubljana

[2] Brechmann, G., Dzieia, W., Hörnemann Heiden, E., Hübscher Lüneburg, H., Jagla, D., Klaue Roxheim, J.: Elektrotehniški priročnik. Viharnik d.o.o., 1994, Ljubljana

[3] Bartenschlager, J.: Mehatronika. Pasadena, 2009, Ljubljana

[4] Miljavec, D., Jereb, P.: Električni stroji. Temeljna znanja. Tiskarna Formatisk, 2005, Ljubljana

[5] Uršič, L.: Sklop enosmernega motorja in asinhronskega samovzbudnega generatorja z regulacijo vrtilne hitrosti. UL, FE, Laboratorij za električne stroje, 2011, Ljubljana

[6] Zagradišnik, I., Slemnik, B.: Električni rotacijski stroji. UM, FERI, 2001, Maribor

[7] Žalar, Z.: Osnove elektrotehnike 1. Tehniška založba Slovenije, 1991, Ljubljana

[8] Žalar, Z.: Osnove elektrotehnike 2. Tehniška založba Slovenije, 2010, Ljubljana

[9] The Definitive Guide To Electric Motors, https://www.beatson.co.uk/history-electric-motors/

[10] Uvod v Arduino strojno platformo, https://zmaga.com/content.php?id=4394

[11] Arduino Uno Rev3, https://store.arduino.cc/arduino-uno-rev3

Izjava

Tako Jaša Vid Meh Peer kot Simon Strmšnik sta prispevala enako k raziskavi in razvoju, zato si delita prvo avtorstvo.

Elektromagnetno-mehanska sklopljena analiza na primeru sinhronskega motorja s trajnim magnetom

Urban Rupnik, Martin Mavrič, Damijan Miljavec

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

Coupled electromagnetic-mechanical analysis based on synchronous motor with permanent magnet

Abstract. *The paper addresses problems of electro mechanics forces in synchronous motor with permanent magnet. These forces are the source of vibration and air pressure oscillation in surroundings. To define forces, we need to analyse electromagnetic conditions in the motor. This analysis is the base to define mechanical transient response of motor. With experimental measurements on motor response, we indicated also some further problems of magnetic and mechanical conditions appearing in the motor, and their consequences.*

Keywords: *coupled analysis, electromagnetic analysis, mechanical analysis, mechanical response, vibrations*

1 Uvod

Sveta brez elektromotorjev si že vrsto let ne moremo predstavljati. Za namen izboljšanja karakteristik teh motorjev potekajo optimizacije in raziskave njihovih elektromagnetnih in termičnih stanj. Največkrat te analize potekajo ločeno, odvisno od fizikalnega problema, ki je v danem trenutku obravnavan. Na tak način nikoli povsem natančno ne zajamemo celotnega sistema. Na primer, ko opazujemo samo termična stanja električnega motorja, zanemarimo povratno zanko, v kateri je od temperature odvisna električna prevodnost bakra ali pa temperaturno razmagnetenje magneta. Vse bolj aktualne so tudi motnje, povezane s hrupom. Ljudje so z višanjem standardov pripravljeni plačati več za tišje izdelke. Ko imamo v mislih vibracije, je pomembno tudi to, da izdelek konstrukcijsko zdrži vsa obratovalna stanja. Za doseganje zgoraj opisanih ciljev so se podjetja pričela aktivno ukvarjati z različnimi sklopitvami fizikalnih problemov. Razvoj informacijske tehnologije, s katerim so nastala različna programska orodja, te sklopitve omogoča (Ansys, Comsol, JMAG). Težava takih programov je v tem, da so ob večjih modelih časovno zelo potratni, za uporabo v podjetjih pa precej dragi. Cilj tega članka je zato povezava magnetostatičnih stanj v motorju z mehanskim odzivom rotorja. Na tem preprostem primeru bomo pokazali, kako obsežno nastane delo, če motor poskušamo modelirati analitično ali numerično tako elektromagnetno kot tudi mehansko. Izračune, dobljene s programskim orodjem Ansys, bomo primerjali z eksperimentalnimi rezultati.

V prvem delu je predstavljen elektromotor BG19 podjetja ebm-papst Slovenija.

V drugem delu je nato predstavljena analiza elektromagnetnih stanj v motorju, katere cilj je izračun navora in sil, ki delujejo med statorjem in rotorjem.

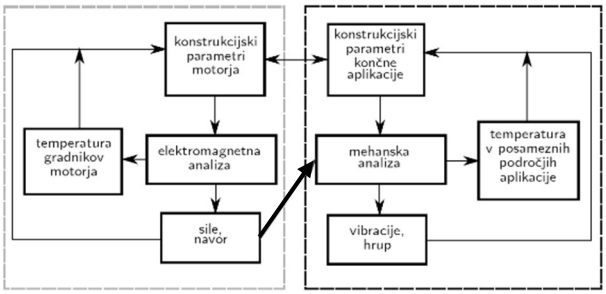
Dobljene sile bomo nato v tretjem delu uporabili za izračun mehanskega odziva sistema rotor-stator.

Na koncu so komentirane posamezne metode izračuna, možne izboljšave ter možnosti za nadaljnje delo.

2 Motivacija

Klasičen potek elektromagnetne in mehanske analize je predstavljen sliki 1 (brez poudarjene puščice). Na levi strani je del, kjer so obravnavana elektromagnetna stanja motorja, na desni pa mehanska stanja. Analiza poteka tako, da se najprej izračuna osnovne dimenzije motorja za želeno navorno karakteristiko in moč. Nato se z iteracijami opazuje termične spremembe v motorju in posledične spremembe elektromagnetnih stanj. Po optimizaciji se motor vgradi v aplikacijo, kjer se zmanjšuje vibracije in hrup. S trajnimi teki se opazujejo spremembe temperature in delovanja v posameznih delih aplikacije.

Želja je, da se elektrotehniški del in mehanski del združita in nastane zaključen preplet fizikalnih vplivov. Prvi korak k temu cilju je predstavljen z odebeljeno puščico na sliki 1. V tem članku je narejena ena iteracija, kjer smo izračunali elektromagnetna stanja v motorju ter nato z mehansko analizo opisali vpliv teh stanj na sistem rotor-stator.



Slika 1: Sklopitev elektromagnetne in mehanske analize

3 Analiza

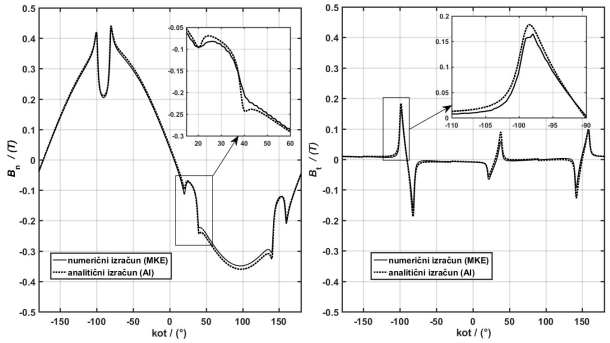
3.1 Trifazni dvopolni motor s trajnim magnetom

Obravnavan motor je trifazni dvopolni motor s trajnim magnetom in koncentriranimi navitji. Uporablja se ga v pomivalnih strojih, kjer kot ventilator služi za pomoč pri sušenju posode. Rotor motorja je iz sintranega feritnega materiala in je paralelno magneten.

3.2 Elektromagnetna analiza

Za izračun elektromagnetno-mehanske sklopitve je bil potreben izračun elektromagnetnega navora v odvisnosti od pozicije rotorja. Naredili smo nekaj poenostavitev za hitrejše numerične in analitične izračune ter zaradi lažje primerjave in vrednotenja modela. Prva poenostavitev je konstantna amplituda vhodnega toka. Druga poenostavitev je sinusna odvisnost elektromagnetnega navora v funkciji razlike električnega in mehanskega (rotorskega) kota. V prvem delu smo na podlagi verifikacije z meritvijo popravili remanentno gostoto magnetnega pretoka magneta iz 385 mT na 400 mT. Tako smo v naslednjem koraku analitično [1] in numerično [2] izračunali in preverili gostoto magnetnega pretoka v zračni reži motorja v odvisnosti od kota zasuka rotorja.

Rezultat analitičnega izračuna [3] sta permeančni funkciji. Ti upoštevata statorske zobe in ju dalje uporabimo za izračun gostote magnetnega pretoka v zračni reži pri poljubnem kotu θ. Ta kot predstavlja zasuk rotorja proti statorju. Primerjava rezultatov obeh izračunov za normalno (B_n) in tangencialno (B_t) komponento gostote magnetnega pretoka pri kotu pozicije θ=15° je prikazana na sliki 2. Na slikah je opaziti zelo dobro ujemanje obeh komponent pretoka.



Slika 2: Gostota magnetnega pretoka v zračni reži. Levo normalna, desno tangencialna.

Eksperimentalno, numerično in analitično smo izračunali inducirano napetost ene faze. Uporabili smo izračun magnetnega pretoka ϕ_i na eni fazi za vsako pozicijo rotorja, kjer smo uporabili enačbo (1),

$$\Phi_\mathrm{i} = l \int_{-90°}^{30°} B_{\mathrm{n_i}} r d\varphi \,;\; i = 1 \ldots N_\theta \qquad (1)$$

kjer je $B_{\mathrm{n_i}}$ gostota magnetnega pretoka v normalni smeri i-tega koraka premika rotorja, l efektivna dolžina motorja, r polmer rotorja in N_Θ število korakov premika rotorja. Dalje lahko izračunamo inducirano napetost E_ik z:

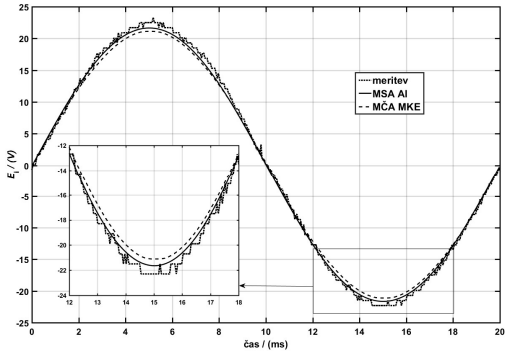
$$E_{i_k} = N \frac{\Phi_{k+1} - \Phi_{k-1}}{2\delta t};\; 2 \le k \le N_\theta - 1 \qquad (2)$$

V enačbi (2) N predstavlja število ovojev navitja, δt pa časovni inkrement, ki ga izračunamo po enačbi (3),

$$\delta t = \frac{1}{n_n N_\theta} \qquad (3)$$

kjer n_n predstavlja hitrost vrtenja v vrtljajih na minuto.

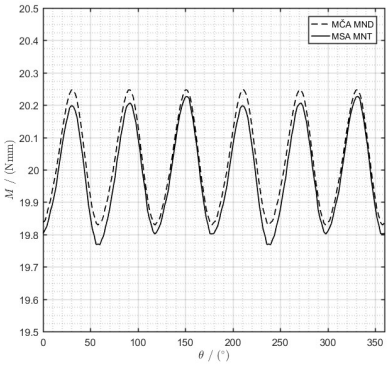
Primerjava merilnih rezultatov z analitičnim izračunom (MSA AI) in metodo končnih elementov (MČA MKE) je prikazana na sliki 3.



Slika 3: Inducirana napetost pri 3000 vrt/min

V zadnjem delu elektromagnetne analize izračunamo še valovitost navora motorja. Na sliki 4 primerjamo rezultate magnetostatične analize z analitičnim izračunom po metodi Maxwellovega napetostnega tenzorja (MSA MNT) s časovno odvisno elektromagnetno analizo z numeričnim izračunom na podlagi končnih elementov. Pri slednjem se uporablja metoda navideznega dela (MČA MND).

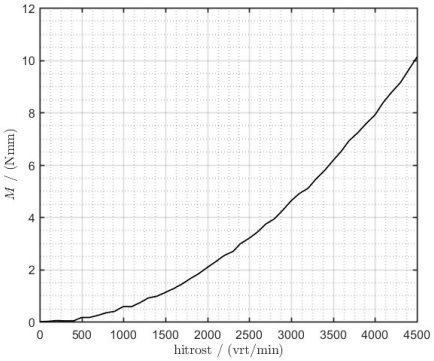
$$M = \frac{1}{\mu_0} r \delta S \sum_{n=1}^{N_\varphi} B_{t_n} B_{r_n} \qquad (4)$$



Slika 4: Valovitost navora motorja.

3.3 Mehanska analiza

V mehanskem delu analize bomo najprej analitično in numerično izračunali in primerjali časovni odziv hitrosti vrtenja rotorja na sunek bremena. V naslednjem koraku bomo motor zavrteli iz mirovanja na želenih 3000 vrt/min in znova opazovali, kako nenadna sprememba navora na gredi vpliva na kotno hitrost (slika 7) in kotni pospešek rotorja in statorja (slika 8). Na koncu bomo analizirali še glavne komponente proizvedenih vibracij, ki nam jih predstavlja akcija-reakcija sistema rotor-stator. Kot rezultat bomo opazovali amplitudni spekter ustaljenega odziva kotnega pospeška statorja kot posledico napajalnega toka in priključenega bremena v obliki ventilatorja. Ventilatorju smo že pred tem izmerili navorno karakteristiko. Dobljena karakteristika je prikazana na sliki 5.



Slika 5: Navorna karakteristika bremena v obliki ventilatorja.

Navorno karakteristiko smo interpolirali in dobili funkcijo navora bremena v odvisnosti od hitrosti. Zapisali smo še vsoto navorov za rotor.

$$M(\theta,\alpha_i) = M_e(\theta,\alpha_i)\sin(\alpha_i-\theta) - M_b(\dot{\theta}) \quad (5)$$

Odziv kota zasuka rotorja proti statorju dobimo po enačbi (6) in je označen z θ,

$$\ddot{\theta} = \frac{M(\theta,\alpha_i)}{J_0} = G(\theta,\alpha_i) \quad (6)$$

kjer je J_0 vztrajnostni moment rotorja in je $G(\theta,\alpha_i)$ nova funkcija, ki smo jo uporabili zaizračun diferencialne enačbe (6). Za numerično reševanje enačbe (6) potrebujemo sledeče nastavke:

$$\begin{aligned} \alpha_{i_n} &= 2\pi f_{el} t_n \\ \ddot{\theta}_n &= G(\theta_n,\alpha_{i_n}) \\ \dot{\theta}_{n+1} &= \delta t\ddot{\theta}_n + \dot{\theta}_n \\ \theta_{n+1} &= \delta t\dot{\theta}_{n+1} + \theta_n \end{aligned} \quad (7)$$

V nastavkih (7) je f_{el} frekvenca napajanja navitij in t_n časovni korak simulacije.

Zapišimo sedaj enačbo (8) za kotni pospešek statorja $\ddot{\beta}$. Predpostavili bomo, da se stator ne more vrteti.

$$\ddot{\beta} = \frac{M_e(\theta)\sin(\alpha_i-\theta)}{J_1} \quad (8)$$

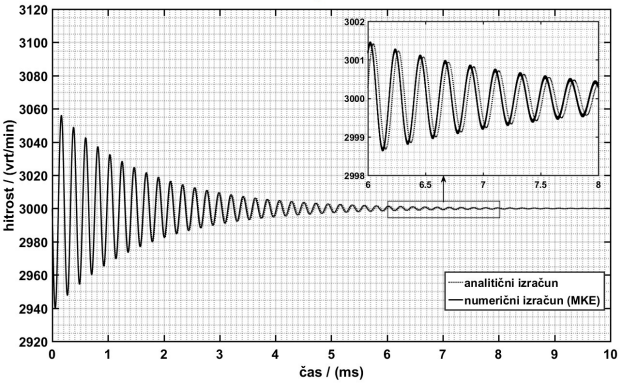
V zgornji enačbi (8) J_1 predstavlja vztrajnostni moment statorja.

Z numeričnim programom Ansys smo opazovali časovni odziv hitrosti rotorja pri istih začetnih pogojih. Ročno je bil izračunan vztrajnostni moment J_0 in vnešena bremenska karakteristika. Rezultati so prikazani na sliki 6. Vidimo, da se analitični in numerični izračun povsem ujemata.

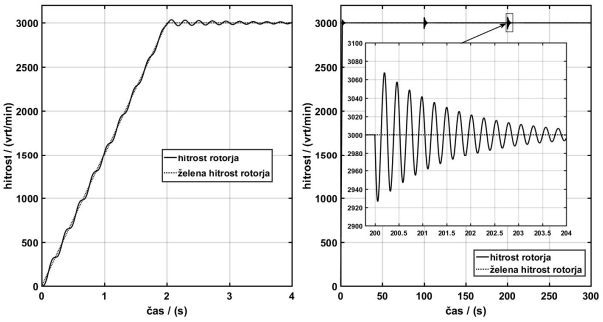
V naslednjem koraku smo opazovali odziv rotorja na želeno hitrost. Ob času 100 s in 200 s smo gred dodatno obremenili s konstantnim enosmernim navorom. Na navitje motorja smo priključili trifazni izmenični tok s konstantno amplitudo. Časovni odziv hitrosti je prikazan na sliki 7.

Na sliki 8 sta kotni pospešek na rotorju in statorju za namenjen interpretaciji vibracij. Opazimo, da se amplituda vibracij povečuje z obremenitvijo na rotorju.

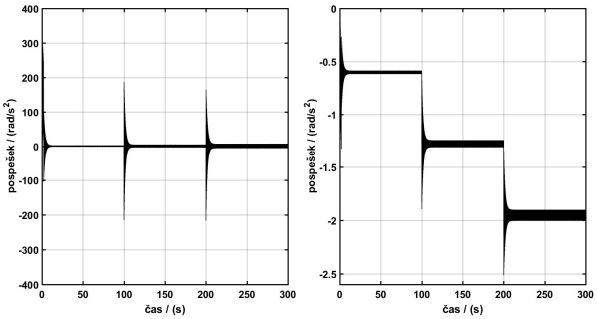
Analizirajmo sedaj še amplitudni spekter kotnega pospeška statorja pri ustaljenem stanju primera na sliki 6, ki opisuje časovni odziv hitrosti rotorja. Potrebno je bilo rešiti diferencialno enačbo (8), pri čemer smo uporabili rezultate kota premika θ. Izsek časovnega odziva kotnega posopeška statorja iz slike 8 (desno) je prikazan na sliki 9. Vidimo, da je na grafu kotnega pospeška statorja prisoten sinus s frekvenco, ki je 6-krat višja od frekvence osnovnega vrtilnega magnetnega polja (50 Hz). Ta višjeharmonska frekvenca je posledica valovitosti navora.



Slika 6: Časovni odziv rotorja



Slika 7: Odziv rotorja na želeno hitrost. Levo od 0-4 s, desno od 0-300 s.



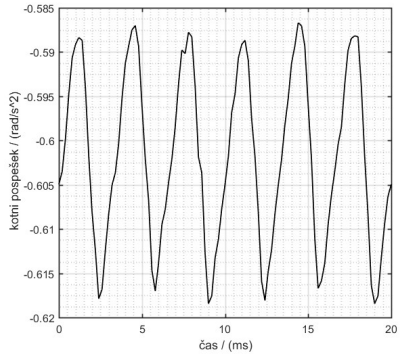
Slika 8: Kotni pospešek rotorja (levo) in statorja (desno).

Na dobljenih rezultatih za kotni pospešek statorja naredimo Fourierovo transformacijo in dobimo naslednji spekter, ki je prikazan na sliki 10. Na spektrogramu (slika 10) vidimo frekvence, različne od osnovne frekvence vrtilnega magnetnega polja. Najvišjo amplitudo imajo frekvence 300 Hz, 600 Hz in 900 Hz. Te frekvence že obstajajo v spektru navorne karakteristike M iz enačbe 4. Vse frekvence pa ne ohranijo svoje začetne vrednosti. Njihove spremembe amplitude so rezultat operacij med trigonometričnimi funkcijami. Za primer pokažimo frekvenčni spekter

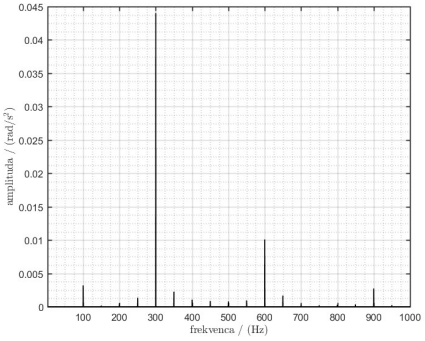
$$\sin(X + Y\cos(\omega_0 t))$$

ker imamo tak primer v odvisnosti $M(\theta)$ in $\sin(\alpha_i-\theta)$. V našem primeru imamo sicer manjše amplitude, ampak primerjava je smiselna za primer generiranja novih frekvenc. Na sliki 11 je prikazan primer, kjer je X=1, Y=1 in ω_0=2π300.

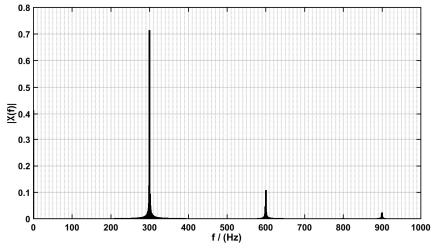
Vidimo, da nastanejo nove frekvence. S spreminjanjem koeficientov X in Y se generirajo različne amplitude različnih frekvenc.



Slika 9: Izsek iz ustaljenega stanja kotnega pospeška statorja

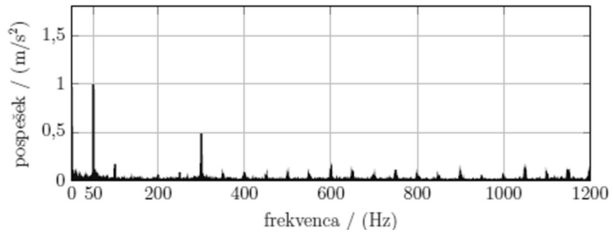


Slika 10: Amplitudni spekter kotnega pospeška statorja



Slika 11: Amplitudni spekter funkcije: sin(1+cos(2π300t))

Ovrednotimo sedaj naše izračune z meritvami. Izmerili smo tangencialne pospeške na statorskem paketu. Motor je bil bremenjen z ventilatorjem. Na sliki 12 je prikazan amplitudni spekter pospeška v tangencialni smeri pri hitrosti 3000 vrt/min.



Slika 12: Amplitudni spekter pospeška v tangencialni smeri na statorskem paketu

Na sliki 12 najprej opazimo frekvenco 50 Hz, ki v našem spektru na sliki 10 ni opažena. To je frekvenca vrtenja rotorja. Zaradi necentriranega rotorja imamo v navorni karakteristiki tudi komponento vrtilne hitrosti.

Druga amplitudno najvišja komponenta ima frekvenco 300 Hz, kar je posledica statorskih čevljev, števila faz in magnetnega polovega para na rotorju. V spektru opazimo tudi ujemanje frekvenc 600 Hz in 900 Hz, kateri sta naslednji največji, kot posledica njihovega obstoja že v navorni funkciji. Ostali večkratniki 50 Hz so posledica računskih operacij med harmoničnimi signali, kot je opisano zgoraj v članku.

4 Zaključek

V predstavljenem delu je prikazana povezava med elektromagnetno analizo in delom mehanske analize. Seveda je to samo mali korak v smeri sklopitve različnih fizikalnih pojavov.

Že na tem preprostem primeru pokažemo, da kljub zanemaritvi debalansa rotorja, trenja ležajev in ostalih parazitnih efektov dobimo širok spekter amplitud kotnih pospeškov statorja.

Dosegli pa vendarle nismo samo sklopitve, temveč smo s primerom tudi pokazali, da lahko z analitično matematičnimi pristopi za izračun gostote magnetnega pretoka v zračni reži izračunamo navor, ki je praktično enak izračunom v programu Ansys. Iz dobljenih rezultatov navora pa dalje izračunamo odziv rotorja in statorja. Za celoten postopek analitičnega izračuna časovnega odziva vrtenja rotorja ob še ne optimiziranem programu potrebujemo maksimalno 5 s, v primerjavi z numeričnim programom Ansys, kjer za isto dolžino časovne analize (300 s) s časovnim korakom (0,1 ms) in redko mrežo končnih elementov v zračni reži potrebujemo približno 1666 ur (70 dni).

S slike 8 smo ugotovili tudi, da se pri bolj obremenjenem motorju (povišan navor) posledično poveča amplituda kotnega pospešek, ki ga čuti stator.

Idej za nadaljevanje analiz in študij se tako odpre ogromno. Najprej bi bilo potrebno dodati debalans rotorja in realno krmiljenje motorja (MTPA). Tak model pa bi potem že bil pripravljen za optimizacijo izbranih parameterov (cena, amplituda vibracij pri izbranih frekvencah, navor, dimenzije motorja, časovni odziv hitrosti rotorja). Dalje pa bi bilo potrebno modelirati še tretjo dimenzijo (vzdolžno os) in aksialne vibracije, ter postaviti mehanski model ležajev in ostalega ohišja.

Zahvala

Projekt (Električni motorji premium energijskega razreda, L2-8187 (C)) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

[1] D. Žarko "A systematic approach to optimized design of permanent magnet motors with reduced torque pulsations" *PhD thesis, Uiversity of Winsconsin - Madison,* 2004.

[2] ANSYS, https://www.ansys.com (11.4.2018).

[3] M. Mavrič "Analiza magnetnih razmer v zračni reži električnega motoja s trajnimi magnet" *Magistrsko delo, Univerza v Ljubljani, 2017*

Metode za izračun ravnotežnih stanj sinhronskega stroja

Marcel Topler, Jožef Ritonja, Boštjan Polajžer

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru,
Koroška cesta 46, 2000 Maribor, Slovenija
E-pošta: marcel.topler@um.si

Methods for calculation of equilibrium states of the synchronous machine

Abstract. *The synchronous machines (SM) are among the most frequently used sources of electrical energy. Consequently, there is a great interest and need for static and dynamic mathematical SM models for the analysis of operating states. The mathematical model should precisely describe the static and dynamic properties of the SM while being as simple as possible. Furthermore, the quantities of the mathematical model should be easily obtained and their calculation should be possible from the basic construction data or from basic tests of SM.*

Most of the time the SM operates in the vicinity of equilibrium state, therefore there is a great interest in methods for calculation of equilibrium states of the SM. The problem in the article is limited to SM connected to an infinite bus of constant amplitude and frequency trough a transmission line i. e. single machine infinite bus (SMIB). The article presents different methods of equilibrium states calculation that are: mixed boundary conditions, infinite bus conditions and SM conditions. After determining the voltages, stator current, active and reactive power, the remaining SM quantities of equilibrium states are calculated using the classical SM phasor diagram.

1 Uvod

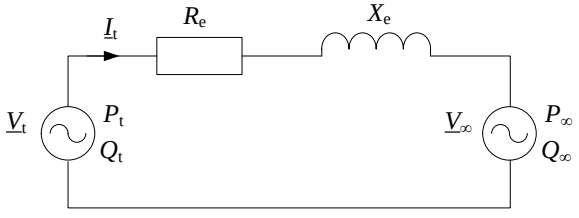
Sinhronski stroji (SS) spadajo med najpogostejše in najpomembnejše proizvajalce električne energije. Posledično obstaja veliko zanimanje in potreba po statičnih in dinamičnih matematičnih modelih SS za analizo obratovalnih stanj. Pri izbiri matematičnega modela si želimo, da le ta čim natančneje opiše statične in dinamične lastnosti SS, pri tem pa naj bo model čim enostavnejši. Veličine matematičnega modela naj bodo enostavno določljive, njihov izračun pa naj bo možen iz osnovnih konstrukcijskih podatkov ali nezahtevnih preizkusov SS [1].

V literaturi [2] in [3] obstaja veliko število modelov SS, ki se med seboj razlikujejo v natančnosti opisa in v zahtevnosti predstavitve modela SS. Med najenostavnješe modele spada t. i. klasični model, kjer je model SS predstavljen z omrežjem konstantne napetosti in serijske reaktance. Nasprotno je najpodrobnejši in hkrati najzahtevnješi model SS t. i. nelinearni d - q model sedme stopnje, kjer so magnetne sklopljenosti statorskih, vzbujalnih in dušilnih navitij odvisne od položaja in nasičenja rotorja [4], [5] in [6]. Med ta dva »skrajna« modela lahko umestimo več modelov SS, ki se razlikujejo v stopnji zahtevnosti. Posebej pomemben je t. i. linearen model tretje stopnje (Heffron - Phillipsov), ki temelji na nekaterih predpostavkah in poenostavitvah, kljub temu pa pravilno opiše dinamiko SS, vendar le v okolici ravnotežnega stanja [7].

Nelinearni modeli SS omogočajo natančno dinamično analizo SS, če poznamo natančne vrednosti veličin SS, ki pa v večini primerov niso poznane. Za njihov izračun so potrebne zahtevne in dolgotrajne meritve SS, zato je uporaba nelinearnih modelov SS zelo omejena. Odpravo omenjenih pomanjkljivosti omogoča poenostavitev in linearizacija nelinearnih modelov SS. Razlog za to predstavlja način delovanja SS v sinhronizmu, ki večino časa obratuje v okolici ravnotežnega stanja. Zaradi tega je brez bistvenega zmanjšanja uporabnosti smiselno linearizirati model SS za izbrano ravnotežno stanje [1] in [2]. Ravnotežna stanja lahko izračunamo z različnimi metodami; v članku so obravnavane tri. Najbolj zahtevna je metoda mešanih robnih pogojev, pri kateri so podane vrednosti fazorja napetosti togega omrežja, ter vrednosti delovne in jalove moči SS. Ostali dve metodi temeljita bodisi na podanih vrednostih fazorja napetosti ter delovne in jalove moči togega omrežja, bodisi na podanih vrednostih fazorja napetosti ter delovne in jalove moči SS.

2 Predpostavke in omejitve

V normalnem obratovanju elektroenergetskega sistema (EES) nas zanima delovanje SS v okolici ravnotežnega stanja. Za potrebe obravnave SS v EES predpostavimo, da je SS preko povezovalnega voda priključen na omrežje sinusne napetosti konstantne amplitude in frekvence. To je t. i. sistem z enim SS, ki je priključen na togo omrežje oz. neskončne zbiralke, kar v angleščini označimo kot SMIB (single machine infinite bus) in je prikazan na sliki 1 [8], kjer $\underline{V}_t$ označuje fazor napetosti na statorskih sponkah, P_t in Q_t delovno in jalovo moč, ki ju SS oddaja v povezovalni vod. $\underline{I}_t$ označuje fazor toka statorskega navitja, R_e in X_e upornost in reaktanco povezovalnega voda, $\underline{V}_\infty$ fazor napetosti togega omrežja, P_∞ in Q_∞ delovno in jalovo moč, ki ju togo omrežje prejema preko povezovalnega voda.



Slika 1: SS priključen na togo omrežje oz. neskončne zbiralke.

3 Predstavitev metod za izračun ravnotežnih stanj

Pri matematičnih modelih SS so potrebni začetni pogoji sistema, zato je potrebno izračunati ravnotežno stanje SS, za kar obstaja več metod. Izbira metode za izračun ravnotežnega stanja SS je odvisna od podanih veličin SS, ki je preko povezovalnega voda priključen na neskončne zbiralke. Obravnavali bomo tri različne metode.

3.1 Mešani robni pogoji: P_t, Q_t in $\underline{V}_\infty$

Določitev veličin ravnotežnega stanja iz mešanih robnih pogojev, kjer so podane vrednosti fazorja napetosti togega omrežja (1), delovne moči (2) in jalove moči (3), ki ju SS oddaja v povezovalni vod, spada med zahtevnejše primere.

$$\underline{V}_\infty = V_{\infty,\mathrm{re}} + jV_{\infty,\mathrm{im}} \tag{1}$$

$$P_\mathrm{t} = \Re\left\{\underline{V}_\mathrm{t}\,\underline{I}_\mathrm{t}^*\right\} \tag{2}$$

$$Q_\mathrm{t} = \Im\left\{\underline{V}_t\,\underline{I}_t^*\right\} \tag{3}$$

Iz $\underline{V}_\infty$, P_t in Q_t lahko ob upoštevanju upornosti R_e in reaktance X_e povezovalnega voda določimo potreben fazor napetosti na statorskih sponkah (4) in fazor toka statorskega navitja (5).

$$\underline{V}_\mathrm{t} = V_{\mathrm{t,re}} + jV_{\mathrm{t,im}} \tag{4}$$

$$\underline{I}_\mathrm{t} = I_{\mathrm{t,re}} + jI_{\mathrm{t,im}} \tag{5}$$

Izhodišče predstavlja kazalčni diagram, prikazan na sliki 2, na podlagi katerega zapišemo fazor napetosti na statorskih sponkah (6).

$$\underline{V}_\mathrm{t} = \underline{V}_\infty + \underline{I}_\mathrm{t}R_\mathrm{e} + j\underline{I}_\mathrm{t}X_\mathrm{e} \tag{6}$$

Za primer na sliki 2, ob upoštevanju (1) in (4) – (6), zapišemo realno (7) in imaginarno (8) komponento fazorja napetosti na statorskih sponkah.

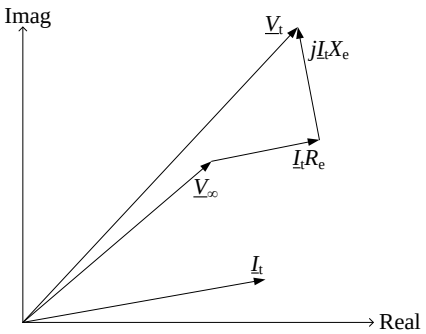
$$V_{\mathrm{t,re}} = V_{\infty,\mathrm{re}} + I_{\mathrm{t,re}}R_\mathrm{e} - I_{\mathrm{t,im}}X_\mathrm{e} \tag{7}$$

$$V_{\mathrm{t,im}} = V_{\infty,im} + I_{\mathrm{t,im}}R_\mathrm{e} + I_{\mathrm{t,re}}X_\mathrm{e} \tag{8}$$

Podobno ob upoštevanju (2) – (5) zapišemo delovno (9) in jalovo (10) moč, ki ju SS oddaja v povezovalni vod.

$$P_\mathrm{t} = V_{\mathrm{t,re}}I_{\mathrm{t,re}} + V_{\mathrm{t,im}}I_{\mathrm{t,im}} \tag{9}$$

$$Q_\mathrm{t} = V_{\mathrm{t,im}}I_{\mathrm{t,re}} - V_{\mathrm{t,re}}I_{\mathrm{t,im}} \tag{10}$$



Slika 2: Kazalčni diagram fazorja statorske napetosti in fazorja statorskega toka.

Ob upoštevanju (7) in (8) v (9) in (10) lahko enačbi preuredimo ju zapišemo kot (11) in (12).

$$R_\mathrm{e}I_{\mathrm{t,re}}^2 + R_\mathrm{e}I_{\mathrm{t,im}}^2 + V_{\infty,\mathrm{re}}I_{\mathrm{t,re}} + V_{\infty,\mathrm{im}}I_{\mathrm{t,im}} - P_\mathrm{t} = 0 \tag{11}$$

$$X_\mathrm{e}I_{\mathrm{t,re}}^2 + X_\mathrm{e}I_{\mathrm{t,im}}^2 + V_{\infty,\mathrm{im}}I_{\mathrm{t,re}} - V_{\infty,\mathrm{re}}I_{\mathrm{t,im}} - Q_\mathrm{t} = 0 \tag{12}$$

Geometrično predstavlja (11) krožnico s polmerom r_1 in središčem $S_1(x_1, y_1)$ (13) ter (12), krožnico s polmerom r_2 in središčem $S_2(x_2, y_2)$ (14), kot je prikazano na sliki 3.

$$\begin{aligned} r_1 &= \sqrt{\left(\frac{V_{\infty,\mathrm{re}}}{2R_\mathrm{e}}\right)^2 + \left(\frac{V_{\infty,\mathrm{im}}}{2R_\mathrm{e}}\right)^2 - \frac{P_\mathrm{t}}{R_\mathrm{e}}} \\ x_1 &= -\frac{V_{\infty,\mathrm{re}}}{2R_\mathrm{e}},\ \mathrm{y}_1 = -\frac{V_{\infty,\mathrm{im}}}{2R_\mathrm{e}} \end{aligned} \tag{13}$$

$$\begin{aligned} r_2 &= \sqrt{\left(\frac{V_{\infty,\mathrm{im}}}{2X_\mathrm{e}}\right)^2 + \left(-\frac{V_{\infty,\mathrm{re}}}{2X_\mathrm{e}}\right)^2 + \frac{Q_\mathrm{t}}{X_\mathrm{e}}} \\ x_2 &= -\frac{V_{\infty,\mathrm{im}}}{2X_\mathrm{e}},\ \mathrm{y}_2 = -\frac{V_{\infty,\mathrm{re}}}{2X_\mathrm{e}} \end{aligned} \tag{14}$$

Oddaljenost med središčema krožnic označimo z S_{12} (15), kot je prikazano na sliki 3.

$$S_{12} = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2} \tag{15}$$

Sistem enačb (11) in (12) lahko ima nič, eno ali dve rešitvi:

- Nič rešitev: $x_1 = x_2$ in $y_1 = y_2$.
- Ena rešitev: $r_1 + r_2 = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2}$.
- Dve rešitvi: $S_{12} = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2}$.

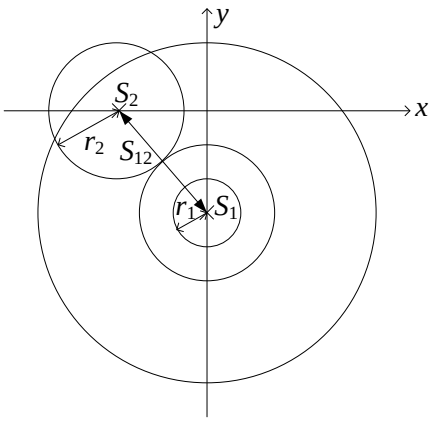
Na podlagi tega lahko sklenemo, da ima sistem eno ali dve rešitvi, ko so izpolnjeni pogoji (16).

$$r_1 + r_2 \geq S_{12} \wedge r_1 < r_2 + S_{12} \wedge r_2 < r_1 + S_{12} \tag{16}$$

V primeru, ko ima fazor napetosti togega omrežja $\underline{V}_\infty$ realno komponentno $V_{\infty,\mathrm{re}} = 1$ in imaginarno komponento $V_{\infty,\mathrm{im}} = 0$, lahko enačbi (11) in (12) poenostavimo v (17) in (18).

$$R_\mathrm{e}I_{\mathrm{t,re}}^2 + R_\mathrm{e}I_{\mathrm{t,im}}^2 + I_{\mathrm{t,re}} - P_\mathrm{t} = 0 \tag{17}$$

$$X_\mathrm{e}I_{\mathrm{t,re}}^2 + X_\mathrm{e}I_{\mathrm{t,im}}^2 - I_{\mathrm{t,im}} - Q_\mathrm{t} = 0 \tag{18}$$



Slika 3: Geometrična predstavitev s krožnicami.

Poenostavijo se enačbe (13) – (15) v (19) – (21).

$$r_1 = \frac{1}{R_e}\sqrt{\frac{1}{4} + P_t R_e},\ x_1 = -\frac{V_{\infty,\mathrm{re}}}{2R_e} \tag{19}$$

$$r_2 = \frac{1}{X_e}\sqrt{\frac{1}{4} + Q_t X_e},\ y_2 = -\frac{V_{\infty,\mathrm{re}}}{2X_e} \tag{20}$$

$$S_{12} = \sqrt{(x_1)^2 + (-y_2)^2} \tag{21}$$

Sistem enačb (17) in (18) ima običajno dve rešitvi, ki sta lahko enaki ali različni, če so izpolnjeni pogoji (22).

$$r_1, r_2, S_{12} > 0 \wedge r_1 + r_2 \geq S_{12} \wedge r_1 \geq r_2 + S_{12} \wedge r_2 \geq r_1 + S_{12} \tag{22}$$

Sistem enačb (17) in (18) preoblikujemo v (23), velja tudi (24).

$$\left[\left(\frac{R_e}{X_e}\right)^2 + 1\right] I_{t,\mathrm{im}}^2 + \left[2\frac{R_e}{X_e}\left(\frac{R_e}{X_e}Q_t - P_t\right) - \frac{1}{X_e}\right] I_{t,\mathrm{im}} + \left(\frac{R_e}{X_e}Q_t - P_t\right)^2 - \frac{Q_t}{X_e} = 0 \tag{23}$$

$$I_{t,\mathrm{re}} = -\left(\frac{R_e}{X_e}Q_t + \frac{R_e}{X_e}I_{t,\mathrm{im}} - P_t\right) \tag{24}$$

Iz kvadratne enačbe (23) izračunamo tok $I_{t,\mathrm{im}}$, katere rešitev sta dve vrednosti, običajno negativna in pozitivna. Izberemo ustreznejšo, to je pozitivno. S pomočjo toka $I_{t,\mathrm{im}}$ lahko iz (24) izračunamo $I_{t,\mathrm{re}}$. Po določitvi $\underline{I}_t$ in $\underline{V}_t$ izračunamo ostale veličine SS v ravnotežnem stanju s pomočjo klasičnega kazalčnega diagrama SS na način, ki je opisan v poglavju 3.4.

3.2 Pogoji togega omrežja: P_∞, Q_∞ in $\underline{V}_\infty$

Podane imamo vrednosti togega omrežja, ki so delovna (25) in jalova (26) moč, ki ju prejema togo omrežje ter fazor napetosti togega omrežja (1). V primerjavi z mešanimi robnimi pogoji je sedaj določitev veličin ravnotežnega stanja precej enostavna.

$$P_\infty = \Re\left\{\underline{V}_\infty \underline{I}_t^*\right\} \tag{25}$$

$$Q_\infty = \Im\left\{\underline{V}_\infty \underline{I}_t^*\right\} \tag{26}$$

Iz $\underline{V}_\infty$, P_∞ in Q_∞ lahko ob upoštevanju upornosti R_e in reaktance X_e povezovalnega voda določimo potreben fazor napetosti na statorskih sponkah (4) in fazor toka statorskega navitja (5). Izhodišče predstavlja kazalčni diagram, prikazan na sliki 2, na podlagi katerega zapišemo fazor napetosti na statorskih sponkah (7). Ob upoštevanju (1), (25) in (26) zapišemo delovno (27) in jalovo (28) moč, ki ju prejema togo omrežje.

$$P_\infty = V_{\infty,\mathrm{re}} I_{t,\mathrm{re}} + V_{\infty,\mathrm{im}} I_{t,\mathrm{im}} \tag{27}$$

$$Q_\infty = V_{\infty,\mathrm{im}} I_{t,\mathrm{re}} - V_{\infty,\mathrm{re}} I_{t,\mathrm{im}} \tag{28}$$

Enačbe (7), (8), (27) in (28) predstavljajo linearen sistem štirih enačb s štirimi neznankami. Ob upoštevanju (29) izrazimo iz enačb $I_{t,\mathrm{re}}$, $I_{t,\mathrm{im}}$, $V_{t,\mathrm{re}}$ in $V_{t,\mathrm{im}}$ in jih zapišemo v matrični obliki (30).

$$\left|\underline{V}_\infty\right|^2 = V_{\infty,re}^2 + V_{\infty,im}^2 \tag{29}$$

$$\begin{bmatrix} I_{t,\mathrm{re}} \\ I_{t,\mathrm{im}} \\ V_{t,\mathrm{re}} \\ V_{t,\mathrm{im}} \end{bmatrix} = \begin{bmatrix} \frac{V_{\infty,\mathrm{re}}}{\left|\underline{V}_\infty\right|^2} & \frac{V_{\infty,\mathrm{im}}}{\left|\underline{V}_\infty\right|^2} & 0 & 0 \\ \frac{\underline{V}_{\infty,\mathrm{im}}}{\left|\underline{V}_\infty\right|^2} & -\frac{\underline{V}_{\infty,\mathrm{re}}}{\left|\underline{V}_\infty\right|^2} & 0 & 0 \\ \frac{R_e V_{\infty,\mathrm{re}} - X_e V_{\infty,\mathrm{im}}}{\left|\underline{V}_\infty\right|^2} & \frac{R_e V_{\infty,\mathrm{im}} + X_e V_{\infty,\mathrm{re}}}{\left|\underline{V}_\infty\right|^2} & 1 & 0 \\ \frac{R_e V_{\infty,\mathrm{im}} + X_e V_{\infty,\mathrm{re}}}{\left|\underline{V}_\infty\right|^2} & \frac{-R_e V_{\infty,\mathrm{re}} + X_e V_{\infty,\mathrm{im}}}{\left|\underline{V}_\infty\right|^2} & 0 & 1 \end{bmatrix} \begin{bmatrix} P_\infty \\ Q_\infty \\ V_{\infty,\mathrm{re}} \\ V_{\infty,\mathrm{im}} \end{bmatrix} \tag{30}$$

Po določitvi $\underline{V}_t$ in $\underline{I}_t$ izračunamo ostale veličine SS v ravnotežnem stanju s pomočjo klasičnega kazalčnega diagrama SS na način, ki je opisan v poglavju 3.4.

3.3 Pogoji SS: P_t, Q_t in $\underline{V}_t$

Podane imamo vrednosti SS, ki so delovna (2) in jalova (3) moč, ki ju SS oddaja v povezovalni vod in fazor napetosti na statorskih sponkah (6). V primerjavi z mešanimi robnimi pogoji je sedaj določitev veličin ravnotežnega stanja precej enostavna. Iz $\underline{V}_t$, P_t in Q_t ob upoštevanju upornosti R_e in reaktance X_e povezovalnega voda določimo fazor napetosti togega omrežja (1) in fazor toka statorskega navitja (5). Ponovno predstavlja izhodišče kazalčni diagram, prikazan na sliki 2, na podlagi katerega zapišemo fazor napetosti togega omrežja (31).

$$\underline{V}_\infty = \underline{V}_t - \underline{I}_t R_e - j\underline{I}_t X_e \tag{31}$$

Ob upoštevanju (4), (5) in (31) zapišemo realno (32) in imaginarno (33) komponento napetosti togega omrežja.

$$V_{\infty,\mathrm{re}} = V_{t,\mathrm{re}} - I_{t,\mathrm{re}} R_e + I_{t,\mathrm{im}} X_e \tag{32}$$

$$V_{\infty,im} = V_{t,\mathrm{im}} - I_{t,\mathrm{im}} R_e - I_{t,\mathrm{re}} X_e \tag{33}$$

Ob upoštevanju (1), (2) in (3) zapišemo delovno (9) in jalovo (10) moč SS. Enačbe (9), (10), (32) in (33) predstavljajo linearen sistem štirih enačb s štirimi neznankami. Ob upoštevanju (34) izrazimo iz enačb $I_{t,\mathrm{re}}$, $I_{t,\mathrm{im}}$, $V_{\infty,\mathrm{re}}$ in $V_{\infty,\mathrm{im}}$ in jih zapišemo v matrični obliki (35).

$$\left|\underline{V}_t\right|^2 = V_{t,\mathrm{re}}^2 + V_{t,\mathrm{im}}^2 \tag{34}$$

$$\begin{bmatrix} I_{t,\mathrm{re}} \\ I_{t,\mathrm{im}} \\ V_{\infty,\mathrm{re}} \\ V_{\infty,\mathrm{im}} \end{bmatrix} = \begin{bmatrix} \frac{V_{t,\mathrm{re}}}{\left|\underline{V}_t\right|^2} & \frac{V_{t,\mathrm{im}}}{\left|\underline{V}_t\right|^2} & 0 & 0 \\ \frac{\underline{V}_{t,\mathrm{im}}}{\left|\underline{V}_t\right|^2} & -\frac{\underline{V}_{t,\mathrm{re}}}{\left|\underline{V}_t\right|^2} & 0 & 0 \\ \frac{-R_e V_{t,\mathrm{re}} + X_e V_{t,\mathrm{im}}}{\left|\underline{V}_t\right|^2} & \frac{-R_e V_{t,\mathrm{im}} - X_e V_{t,\mathrm{re}}}{\left|\underline{V}_t\right|^2} & 1 & 0 \\ \frac{-R_e V_{t,\mathrm{im}} - X_e V_{t,\mathrm{re}}}{\left|\underline{V}_t\right|^2} & \frac{R_e V_{t,\mathrm{re}} - X_e V_{t,\mathrm{im}}}{\left|\underline{V}_t\right|^2} & 0 & 1 \end{bmatrix} \begin{bmatrix} P_t \\ Q_t \\ V_{t,\mathrm{re}} \\ V_{t,\mathrm{im}} \end{bmatrix} \tag{35}$$

Po določitvi $\underline{V}_\infty$ in $\underline{I}_t$ izračunamo ostale veličine SS v ravnotežnem stanju s pomočjo klasičnega kazalčnega diagrama SS na način, ki je opisan v poglavju 3.4.

3.4 Določitev ostalih veličin SS

Ko imamo podane in izračunane vse potrebne veličine ravnotežnega stanja SS, ki so $\underline{V}_t$, P_t, Q_t, $\underline{I}_t$, R_e, X_e, $\underline{V}_\infty$, P_∞ in Q_∞, izračunamo še ostale veličine SS v ravnotežnem stanju s pomočjo klasičnega kazalčnega diagrama SS, ki je prikazan na sliki 4 [2].

Najprej izračunamo kolesni kot (36), kjer sta posamezni komponenti napetosti $E_{qa,\mathrm{im}}$ in $E_{qa,\mathrm{re}}$ določeni s (37) in (38).

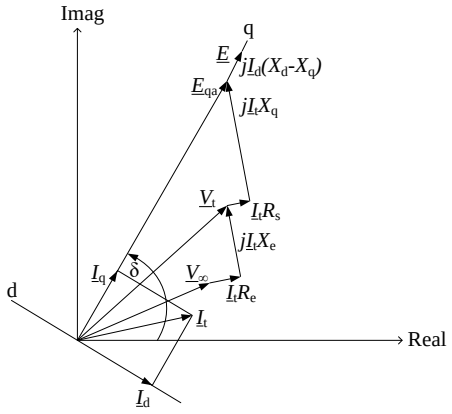
$$\delta = \frac{E_{qa,\mathrm{im}}}{E_{qa,\mathrm{re}}} \tag{36}$$

$$E_{qa,\mathrm{im}} = V_{t,\mathrm{im}} + R_s I_{t,\mathrm{im}} + X_q I_{t,\mathrm{re}} \tag{37}$$

$$E_{qa,\mathrm{re}} = V_{t,\mathrm{re}} + R_s I_{t,\mathrm{re}} - X_q I_{t,\mathrm{im}} \tag{38}$$

Skupna efektivna inducirana napetost statorja E_{qa} je določena z (39).

$$E_{qa} = \sqrt{E_{qa,\mathrm{re}}^2 + E_{qa,\mathrm{im}}^2} \tag{39}$$



Slika 4: Klasični kazalčni diagram SS.

Skupno efektivno inducirano napetost zračne reže E zapišemo s (40), kjer sta posamezni komponenti E_{re} in E_{im} določeni z (41) in (42). Inducirane napetosti $E_{qa,im}$, $E_{qa,re}$, E_{qa}, E, E_{re} in E_{im} so posledica magnetne sklopljenosti SS.

$$E = \sqrt{E_{re}^2 + E_{im}^2} \tag{40}$$

$$E_{re} = E_{qa,re} - I_d (X_d - X_q)\cos\delta \tag{41}$$

$$E_{im} = E_{qa,im} - I_d (X_d - X_q)\sin\delta \tag{42}$$

Toka I_q in I_d izrazimo v matrični obliki (43).

$$\begin{bmatrix} I_q \\ I_d \end{bmatrix} = \begin{bmatrix} \cos\delta & \sin\delta \\ -\sin\delta & \cos\delta \end{bmatrix} \begin{bmatrix} I_{t,re} \\ I_{t,im} \end{bmatrix} \tag{43}$$

Podobno izrazimo tudi napetosti V_q in V_d (44).

$$\begin{bmatrix} V_q \\ V_d \end{bmatrix} = \begin{bmatrix} \cos\delta & \sin\delta \\ -\sin\delta & \cos\delta \end{bmatrix} \begin{bmatrix} V_{t,re} \\ V_{t,im} \end{bmatrix} \tag{44}$$

4 Primer izračuna ravnotežnega stanja SS

Obravnavan je SS s podatki iz tabele 1, kjer so: P_t in Q_t delovna in jalova moč SS, V_∞ napetost togega omrežja, R_e in X_e upornost in reaktanca povezovalnega voda, X_d in X_q reaktanci d in q osi SS, X_d' in X_d'' prehodna (tranzientna) in začetna (subtranzientna) reaktanca d osi SS, l_d in l_q sklopljeni reaktanci d in q osi SS, R_s, R_F, R_D in R_Q upornosti statorskega, vzbujalnega in dušilnega navitja d in q osi SS, H vztrajnostna konstanta, D dušenje, T_{d0}' prehodna časovna konstanta v praznem teku, ω_s električna sinhronska hitrost in cosφ faktor moči.

Za obravnavani SS so z metodo mešanih robnih pogojev izračunane veličine ravnotežnega stanja za primer P_t = 1 pu in cosφ = 0,85. Rezultati so podani v tabeli 2.

Tabela 1: Podane veličine SS.

P_t = 1 pu	Q_t = 0,6197 pu	V_∞ = 1 pu
R_e = 0,2 pu	X_d = 1,7 pu	X_q = 1,64 pu
X_d' = 0,245 pu	X_d'' = 0,185 pu	l_d = 0,15 pu
R_s = 0,001096 pu	T_{d0}' = 5,9 s	l_q = 0,15 pu
R_F = 0,000742 pu	R_D = 0,0131 pu	R_Q = 0,054 pu
X_e = 0,4 pu	D = 2 pu	H = 2,37 s
$\omega_s = 2\pi 50$ s^{-1}	cosφ = 0,85	

Tabela 2: Izračunane veličine ravnotežnega stanja SS.

V_t = 1,1723 pu	I_t = 1,0036 pu	δ = 0,9379 pu
V_d = -0,6628 pu	V_q = 0,9670 pu	I_d = -0,9183 pu
I_q = 0,4047 pu	E_{re} = 1,4957 pu	E_{im} = 2,0388 pu
E = 2,5286 pu	$E_{qa,im}$ = 1,4631 pu	$E_{qa,im}$ = 1,9943 pu
E_{qa} = 2,4735 pu		

5 Zaključek

Pri pregledu literature na področju izračunov ravnotežnih stanj SS nismo zasledili prispevkov, kjer so podrobneje predstavljene in opisane metode za izračun ravnotežnih stanj SS. Tako je glavni prispevek članka pregled in podrobnejši opis metod za izračun ravnotežnih stanj SS, ki so potrebna za natančen opis matematičnih modelov SS za analizo obratovalnih stanj. Obravnava je omejena na SS, priključen na neskončne zbiralke. Glede na podane veličine SS se uporabljajo različne metode za izračun ravnotežnih stanj SS. Ko imamo podane in izračunane vse potrebne veličine ravnotežnega stanja SS, je izveden še izračun ostalih veličin SS s pomočjo klasičnega kazalčnega diagrama SS.

Sekundarni prispevek članka je izračun ravnotežnih stanj SS z metodo mešanih robnih pogojev. Reševanje sistema enačb za izračun realne in imaginarne komponente statorskega toka $\underline{I}_t$ SS je izvedeno s pomočjo dveh enačb krožnic. Rešitev takšnega sistema je odvisna od lege in polmera krožnic.

6 Literatura

[1] J. Ritonja, M. Petrun: Matematični modeli sinhronskega generatorja. Ventil, let. 21, št. 5, str. 42-48, januar 2015.

[2] P. M. Anderson, A. A. Fouad: Power System Control and Stability. The Iowa State University Press, Ames, Iowa, USA. 1977.

[3] P. Kundur: Power System Stability and Control. McGraw-Hill, Inc., New York. 1994.

[4] J. Ritonja, M. Petrun, J. Černelič, R. Brezovnik, B. Polajžer: Analysis and Applicability of Heffron–Phillips Model. Elektronika ir Elektrotechnika, let. 22, št. 4, str. 3-10, 2016.

[5] J. Ritonja, B. Grčar, B. Polajžer: Mathematical models for design and synthesis of power system stabilizers. Proceedings of the 36th Chinese control conference, str. 2299-2304, Dalian, China, July 26-28, 2017.

[6] J. Ritonja, B. Grčar: Applicability of linearized model of synchronous generator for power system stability analysis. International journal of energy and power engineering, let. 10, št. 11, str. 1361-1367, 2016.

[7] W. G. Heffron, R. A. Phillips: Effect of a modern amplidyne voltage regulator on underexcited operation of large turbine generators. AIEE Transactions, let. 71, str. 692–697, 1952.

[8] J. Ritonja: Izboljšanje dinamične stabilnosti sinhronskega generatorja s pomočjo adaptivnega modelno referenčnega vodenja. Doktorska disertacija, UM-FERI, 1996.

Zmanjšanje toplotnih izgub elektromagnetnega ventila

Simon Nosan[1], Urban Tomc[1], Katja Klinar[1], Danjel Vončina[2], Andrej Kitanovski[1]

[1]*Univerza v Ljubljani, Fakulteta za strojništvo*
[2]*Univerza v Ljubljani, Fakulteta za elektrotehniko*
E-pošta: simon.nosan@fs.uni-lj.si

Heat loss reduction of an electromagnetic valve

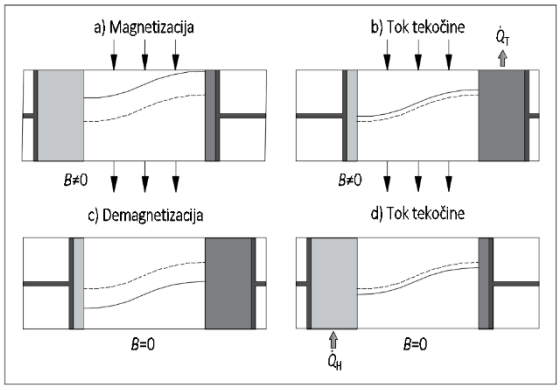
Abstract. *This paper explains a principle of reducing heat losses of solenoid valve in magnetocaloric cooling application. Magnetocaloric material is the core of operating structure. Other important parts are magnetic source (which can include permanent magnets or electromagnets) and hydraulic system. Heat gains are unwanted due to reduction of total efficiency and cooling power.*

1 Uvod

Magnetno hlajenje pri temperaturi okolice je okolju prijazna tehnologija v razvoju, ki veliko obeta. V prihodnosti bi lahko nadomestila razširjeno parno-kompresorsko tehnologijo. Tehnologija magnetnega hlajenja temelji na t. i. magnetokaloričnem efektu (MKE), prisotnem v magnetokaloričnih materialih (MKM) ob spremenljivem magnetnem polju. Ob magnetizaciji MKM temperatura v njem naraste, medtem ko ob demagnetizaciji upade [1]. Zaradi svetovne težnje po napravah s čim višjim izkoristkom je potrebno spreminjanje magnetnega polja izvajati čim bolj učinkovito, sicer se vloženo delo pretvori v toploto. Dandanes je večina raziskav s področja magnetenja MKM usmerjenih v trajne magnete [2-4]. Ti za obstoj magnetnega polja ne potrebujejo dovajanja energije, potrebujejo pa gibljiv sklop za spreminjanje magnetne poljske jakosti v MKM. MKM je oblikovan v porozno strukturo, ki jo imenujemo aktivni magnetni regenerator (AMR). Preko njega se pretaka tekočina za prenos toplote, za katero je potreben hidravlični sistem primerne vezave, ki omogoča pretok tekočine v eno ali drugo smer. S stališča AMR je nezaželeno, da bi hidravlični sistem vplival na hladilni sistem s toplotnimi dobitki, zato je v prispevku obširneje predstavljen način, kako zmanjšati del toplotnih dobitkov hidravličnega sistema.

2 Aktivni magnetni regenerator

AMR je porozna struktura iz MKM in predstavlja osrednji del magnetne hladilne naprave. Magnetna hladilna naprava poleg AMR vsebuje še magnetilni sistem, ki je lahko izveden s pomočjo trajnih magnetov ali z elektromagnetno strukturo. Hidravlični sistem sestavljata dva zunanja prenosnika toplote, tekočina za prenos toplote, štirje ventili za usmerjanje toka tekočine in črpalka. Tekočina prenaša toploto iz AMR na okolico preko zunanjih prenosnikov toplote. Tekom delovanja naprave se izvajajo štiri faze, ki jih lahko vidimo na sliki 1:



Slika 1: Štiri glavne faze delovanja AMR.

A: magnetenje, ko se MKM zaradi MKE segreje.
B: Tok tekočine iz hladnega prenosnika toplote skozi segret MKM v topel prenosnik toplote, pri čemer ohranjamo konstantno (visoko) magnetno polje. Tekočina, namenjena prenosu toplote, se ob obtekanju MKM segreje, MKM pa se ohladi. V topel prenosnik toplote vstopa tekočina s temperaturo, ki je višja od okoliške, zato odda toploto v okolico.
C: Razmagnetenje, ko se MKM zaradi MKE ohladi.
D: Tok tekočine iz toplega prenosnika toplote skozi ohlajen MKM v hladen prenosnik toplote, pri čemer ohranjamo konstantno (nizko) magnetno polje. Tekočina, namenjena prenosu toplote, se ob obtekanju MKM ohladi, MKM pa se segreje. V hladen prenosnik toplote vstopa tekočina s temperaturo, ki je nižja od temperature hlajenega prostora, zato iz njega absorbira toploto.

3 Krmiljenje ventilov

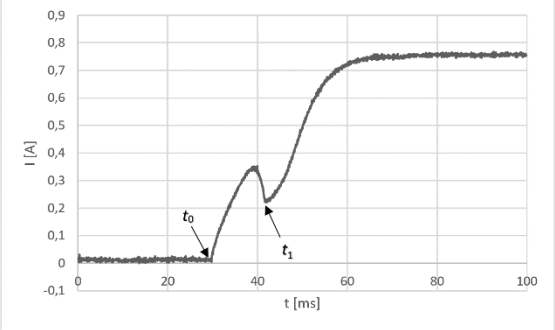
Smer pretoka tekočine skozi MKM spreminjamo s štirimi električno krmiljenimi direktnimi 2/2 (NC) solenoidnimi ventili. Tekočina enkrat ogreva, drugič ohlaja AMR, zato je pomembno, da nanjo ne vplivamo nekontrolirano. Ventili se ob svojem delovanju segrevajo, s stališča delovanja naprave pa bi bilo bolje, če se ne bi. Zato je v nadaljevanju opisan način, kako zmanjšati izgube na ventilu in s tem prenos toplotnih dobitkov na tekočino.

Znotraj ventila je navitje, ki ob električnem toku generira magnetno polje. Takrat magnetna sila pritegne feromagnetno jedro – ventil se iz zaprtega stanja premakne v odprto. Čas preklopa iz enega stanja v drugo podaja proizvajalec za vsak ventil posebej ob predpisanih obratovalnih pogojih. Nanj lahko vplivamo z velikostjo magnetne sile, ki je odvisna od toka po navitju, vrednost toka pa je različna pri različnih priključnih napetostih. Ker je navitje ohmsko-induktivnega značaja, bo tok večji ob višji napetosti. Po odklopu napajalne napetosti vzmet vrne jedro ventila v prvotno (zaprto) stanje.

Največja vrednost toka skozi navitje ventila je odvisna od priključne napetosti, katere nazivno vrednost poda proizvajalec, in upornosti navitja, ki je določena s snovno-geometrijskimi lastnostmi navitja. Ker se energija ne more hipoma spremeniti, lahko po enačbi 1 vidimo, da se tudi tok skozi navitje ne bo hipoma spremenil – je zvezen.

$$W_{mag} = \frac{LI^2}{2} \quad (1)$$

Za primer vzemimo zaprt ventil, ki ga ob času t_0 (glej sliko 2) priklopimo na napajanje.



Slika 2: Potek toka skozi ventil ob nazivni napetosti.

Tok najprej narašča po eksponentni funkciji, dokler ne doseže prvega maksimuma. Tedaj je magnetna sila dovolj velika, da začne premikati feromagnetno jedro proti zračni reži, zaradi česar se povečuje induktivnost tuljave. V enačbi 2 za inducirano napetost vidimo, da se mora ob nespremenjeni napajalni napetosti (ventil je priključen na napajalnik, ki vzdržuje konstantno napetost) in ob povečevanju induktivnosti, tok zmanjševati.

$$U_i = \frac{d(LI)}{dt} + RI \quad (2)$$

Ko se jedro ne more več premikati, se induktivnost ne spreminja več. Feromagnetno jedro bo v končni poziciji tako, da bo magnetni pretok skozi njega največji, kar vidimo na sliki 2 ob času t_1. Tok potem še nekaj časa narašča, nato pa se ustali – prehodni pojav izzveni. Takrat je priključna napetost enaka ohmskemu padcu na navitju.

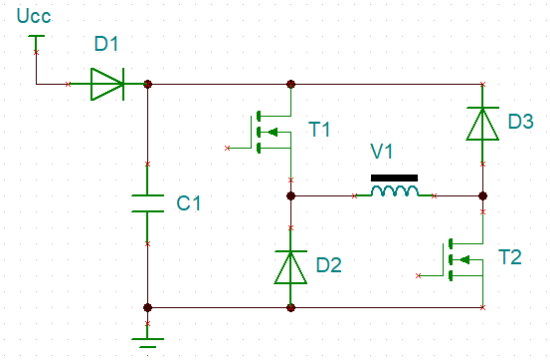
V odprtem stanju ventila se bo na njem trošila moč v obliki toplote, ki jo lahko popišemo z enačbo 3.

$$P_{Joule} = RI^2 \quad (3)$$

Sila, ki jo potrebujemo za držanje feromagnetnega jedra v poziciji, ko je ventil odprt, je zaradi mirujočega jedra v končni legi sedaj manjša, kot na začetku, ko je morala hitro premakniti jedro. Magnetna upornost je po premiku v odprto stanje najmanjša zaradi najmanjše zračne reže. Držalno magnetno silo lahko zmanjšamo do te mere, da bo še vedno večja od sile vzmeti, vendar manjša od maksimalne vrednosti. S tem se zmanjša tudi potreben tok skozi navitje, ki nam magnetno silo ustvarja, to pa pomeni manjše toplotne izgube.

3.1 Uporaba pulzno širinske modulacije

Tok skozi navitje elektromagneta lahko zmanjšamo s pomočjo pulzno širinske modulacije (PŠM) [5]. Vezje, ki dopušča napajanje ventila s PŠM, je prikazano na sliki 3.



Slika 3: Krmilno vezje.

Ventil V_1 je preko MOSFET tranzistorjev T_1 in T_2 vezan na kondenzator C_1, ki se napaja preko diode D_1 iz napajalnika U_{cc}. Ob vklopu krmilnega signala na tranzistorjih, ki je isti za T_1 in T_2, bosta oba tranzistorja v prevodnem stanju. Tok bo iz napajalnika stekel skozi diodo D_1, tranzistor T_1, ventil V_1, tranzistor T_2 in se zaključil v napajalnik. Ob električnem toku skozi ventil je v njem shranjene nekaj magnetne energije. Ko krmilni signal na tranzistorjih izklopimo, se le-ti zaprejo. Tok bo prešel iz ventila na diodo D_3, stekel v kondenzator C_1 in se zaključil preko diode D_2 nazaj do ventila V_1. V tem primeru imamo ob vklopu in izklopu ventila enak prehodni pojav, saj je višina inducirane napetosti ob vklopu in izklopu enaka (ob kondenzatorju z dovolj veliko kapacitivnostjo, da napetost na njem ne zraste bistveno). Torej potrebuje ventil enako časa za preklop iz zaprtega stanja v odprto in obratno.

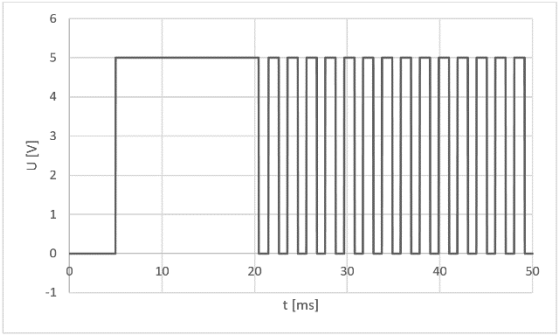
Za vzdrževanje ventila v odprtem stanju je potrebno dovajati vsaj tolikšno moč, da vzmet jedra ne potisne nazaj v zaprto stanje. Vse, kar je več od tega minimuma, se pretvori v nepotrebno dodatno toploto. Vendar pa morajo biti za zagotavljanje preklopnih časov ventila s strani proizvajalca prisotni nazivni obratovalni pogoji, kar lahko dosežemo s kombinacijo nazivne vrednosti napetosti in PŠM.

3.2 Kombinacija nazivne vrednosti napetosti in PŠM

V nadaljevanju je predstavljen princip krmiljenja, kjer ventilu omogočimo hiter preklop skupaj z nizkim

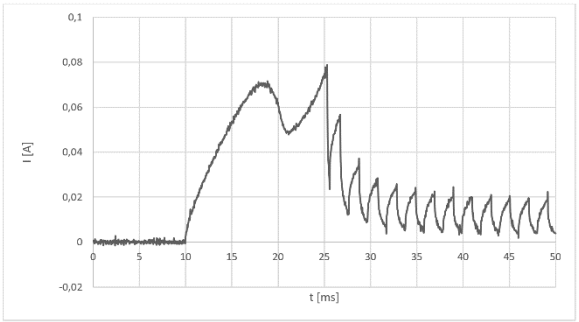
segrevanjem v odprtem stanju. Proizvajalec ventilov običajno poda preklopni čas ventila iz zaprtega v odprto stanje. Med tem časom se na ventilu zahteva nazivna priključna napetost, po tem času pa je lahko napetost nižja. Krmilni signal tranzistorja mora biti na visokem nivoju vsaj za čas preklopa, potem pa se lahko spreminja glede na frekvenco in delovno razmerje PŠM.

Na sliki 4 vidimo krmilni signal za tranzistorja.



Slika 4: Krmilni signal za tranzistorja.

Na začetku je tranzistor izklopljen in s tem tudi ventil zaprt. Sledi visok nivo, ki mora biti prisoten vsaj toliko časa, da ventil zagotovo preklopi v odprto stanje. Potem sledi vklapljanje in izklapljanje krmilnega signala. Obliko toka, posneto na osciloskopu s tokovnimi kleščami, vidimo na sliki 5. Ob visokem nivoju krmilnega signala tok narašča, ob nizkem nivoju pa pada. Srednja vrednost toka, ko je ventil v prevodnem stanju, in s tem posledično magnetna sila, mora biti večja od sile vzmeti, ki potiska jedro ventila proti zaprtemu stanju. Opazimo, da je vrednost toka skozi ventil ob uporabi PŠM po 20 ms od vklopa krmilnega signala znatno nižja.



Slika 5: Tok skozi ventil ob PŠM.

V naslednjem podpoglavju si bomo ogledali, kako lahko ovrednotimo zmanjšanje potrebnega toka skozi ventil v prevodnem stanju.

3.3 Določanje toplotnih izgub ventila

Toplota, s katero se segreva ventil, se pretvori iz električne energije. Moč, ki se troši na ventilu, je v vsakem trenutku enaka

$$P_V = U_V i_V \quad (4)$$

ali

$$P_V = R_V i_V^2 \quad (5)$$

kjer je U_V enosmerna nazivna napajalna napetost ventila, R_V upornost navitja v ventilu, i_V pa trenutna efektivna vrednost toka v navitju. Energija enega cikla delovanja ventila je odvisna od frekvence, s katero želi uporabnik preklapljati ventil. V režimu delovanja, kjer ventil najprej priklopimo na enosmerno nazivno napetost, potem pa vklopimo PŠM, lahko energijo enega cikla, ki se na ventilu pretvori v toploto, razdelimo na dva prispevka: energijo za preklop in energijo, ko je magnetna sila zaradi PŠM manjša.

$$W_V = W_{preklop} + W_{PŠM} \quad (6)$$

Energija, ki jo dovedemo ventilu v času preklopa, se ne spreminja glede na obratovalno frekvenco. Preklop ventila je v vseh primerih enak, zato je enaka tudi energija za preklop. Energija, ki je potrebna za vzdrževanje prevodnega stanja, je odvisna od frekvence delovanja. Višja, kot je frekvenca, krajši bo čas, ko bo ventil v prevodnem stanju, torej bo posledično energija manjša.

Energijo, potrebno za preklop, bomo izračunali s pomočjo enačbe 4, ki jo je potrebno še pomnožiti s časom.

$$W_{V,preklop} = \int_{t_0}^{t_{preklop}} U_V i_V(t) dt \quad (7)$$

Moč integriramo v času od začetka vklopa krmilnega signala t_0 do časa $t_{preklop}$, ko ventil prenehamo napajati z napajalno napetostjo in začnemo s PŠM načinom.

Nazivno napetost poda proizvajalec, časovni potek toka pa smo določili s pomočjo osciloskopa. Energijo v odprtem stanju ventila lahko določimo s pomočjo enačbe 5.

$$W_{V,ON} = \int_{t_{preklop}}^{t_{OFF}} R_V i_V^2(t) dt \quad (8)$$

Moč integriramo v času od preklopa ventila v prevodno stanje $t_{preklop}$ do izklopa krmilnega signala ob času t_{OFF}, ko ventil preide v zaprto stanje.

4 Rezultati

V tem poglavju so podani rezultati izboljšav glede zmanjšanja toplotnih izgub na elektromagnetnem ventilu proizvajalca Jakša. V tabeli 1 so navedeni uporabljena oprema in elementi.

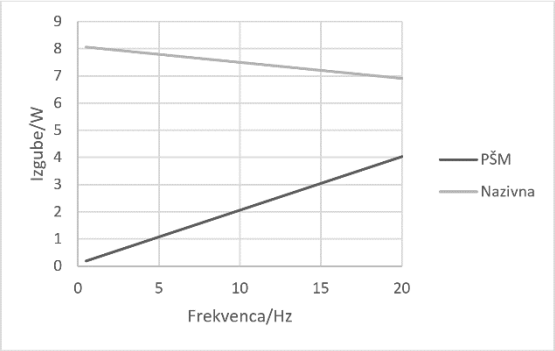
Tabela 1: Uporabljena oprema in elementi.

Oprema	Proizvajalec	Tip
Ventil	Jakša	D221
Osciloskop	Keysight	EDUX1002A
Tokovne klešče	Chauvin Arnoux	E3N 10A/V
Napajalnik	Tenma	DC 0-30V5AX2
Krmilnik	Arduino	Mega 2560
Diode	ON Semiconductor	MBR1045G
Tranzistorji	infineon	IRF540NPBF
Kondenzator	Panasonic	ECA-1VHG472

Nazivne vrednosti uporabljenega ventila, ki jih podaja proizvajalec, so: $U_V = 24$ V, $P_V = 16$ W, nazivni preklopni čas iz zaprtega v odprto stanje pa 10 ms. Iz tega lahko določimo upornost navitja $R = 36\ \Omega$ in nazivni tok $I = 0{,}67$ A.

Krmilni signal smo generirali s pomočjo krmilnika Arduino Mega, ki lahko na digitalnih izhodih generira PŠM signal prednastavljene frekvence 490 Hz. Delovno razmerje, ki je znašalo 50%, smo določili eksperimentalno tako, da smo izenačili tlačni padec in pretok skozi ventil ob napajanju s PŠM s tistim, ki ustreza obratovanju z nazivno napetostjo. Tako je ventil v obeh načinih krmiljenja popolnoma odprt. Čas, ko ventil prehaja iz zaprtega v odprto stanje, je podan s strani proizvajalca. Zaradi varnosti smo ta čas, ko je ventil priklopljen na nazivno napetost (pred izvajanjem PŠM), podaljšali z 10 ms na 16 ms.

Ker je preklopni čas ventila neodvisen od frekvence, s katero želimo odpirati in zapirati ventil, sta na sliki 6 podani moči na ventilu ob nazivnem delovanju in ob delovanju s PŠM, obe v odvisnosti od frekvence preklopov ventila iz zaprtega v odprto stanje in nazaj v zaprto. Izgube krmilne elektronike smo zanemarili, saj ne vplivajo na segrevanje ventila.



Slika 6: Primerjava izgub na ventilu.

Opazimo lahko, da je ob krmiljenju ventila s PŠM moč na njem ves čas nižja, kakor bi bila ob napajanju z nazivno napetostjo. Izgube se ob nazivnem delovanju s povečevanjem števila vklopov na časovno enoto znižujejo, kar je logično, saj tok takoj ob vklopu ne naraste skočno na nazivno vrednost. Torej se ob višjem številu vklopov na časovno enoto daljši delež časa nahajamo v prehodnem območju, ko je produkt napetosti in toka nižji, kot je ob povsem odprtem ventilu.

Po drugi strani potrebna moč ob krmiljenju ventila s PŠM narašča s številom preklopov v časovni enoti, kar lahko utemeljimo s tem, da je efektivna vrednost toka v času preklapljanja večja od toka ob času PŠM. Energija, potrebna za preklop, je neodvisna od frekvence, prišteje se ji pa še energija med PŠM, ki pa se s povečevanjem preklopov zmanjšuje.

Frekvenca preklopov ventila je izbrana glede na ostali hidravlični sistem, ki najbolje deluje pri nižjih frekvencah, zato nadaljnje povečevanje števila preklopov ni bilo izvedeno.

5 Zaključek

V prispevku je opisano zmanjšanje toplotnih izgub elektromagnetnega ventila, ki spreminja smer pretoka tekočine za prenos toplote. Toplotne izgube predstavljajo toplotne dobitke v hladilnem sistemu in ga nezaželeno segrevajo.

Zmanjšanje izgub na ventilu je izvedeno s pomočjo pulzno širinske modulacije. Predstavljena je električna shema, ustrezna oblika krmilnega signala, izračun potrebne energije za preklop ventila in rezultati. Potrebna moč za delovanje ventila s pomočjo pulzno širinske modulacije se izkaže za 7-krat manjšo od moči ob nazivnem delovanju pri frekvenci preklopov ventila 5 Hz. Z višanjem frekvence preklopov ventila pa metoda s pulzno širinsko modulacijo izgublja na prednosti.

Literatura

[1] Kitanovski A, Tušek J, Tomc U et all (2015) Magnetocaloric energy conversion: From theory to applications. Series: Green energy and technology. Springer Publications

[2] Lu DW, Xu XN, Wu HB et al (2005) A permanent magnet magneto-refrigerator study using Gd/Gd-Si-Ge/Gd-Si-Ge-Ga alloys, in: First IIF-IIR International Conference on Magnetic Refrigeration at Room Temperature, Thermag I, Montreux, Switzerland, 2005, p 291-296

[3] Okamura T, Yamada K, Hirano N et al (2005) Performance of a room-temperature rotary magnetic refrigerator. In: First IIF-IIR International Conference on Magnetic Refrigeration at Room Temperature, Thermag I, Montreaux, Switzerland, 2005, p 319-324

[4] Huang J, Liu J, Jin P et al (2006) Development of permanent magnetic refrigerator at room temperature. Rare Metals 25, Spec Issue, p 641

[5] Taghizadeh M, Ghaffari A, Najaf F (2009) Modeling and identification of a solenoid valve for PWM control applications, C. R. Mecanique 337: 131-140

Detekcija napak v feromagnetnih materialih z metodo vrtinčnih tokov

Matic Markl, Henrik Lavrič, Klemen Drobnič, Danilo Makuc, Rastko Fišer

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

Fault detection in ferromagnetic materials by eddy current method

The paper presents development of a probe for eddy current testing of ferromagnetic materials. Multiple types of probes were tested with FEM analysis. All relevant parameters of the probe (geometrical and electrical) were evaluated and discussed. The same procedure was used also for evaluating probability of detecting the irregularity in material. In addition to simulations real probe was constructed, as well as auxiliary electric circuits and an user interface for controlling the examination procedure. Comparison between simulation and measurement results show good agreement and some further improvements of measuring systems were proposed.

1 Uvod

Metoda preiskovanja materialov z vrtinčnimi toki je ena izmed mnogih metod, ki spadajo med neporušne preiskave NDT (Non-Destructive Testing). Neporušne se imenujejo zato, ker se vzorcu med preiskavo ne spremenijo lastnosti, kar pomeni, da je vzorec enako uporaben kot pred preiskavo. Materiale je možno preiskovati z magnetnimi delci, ultrazvokom, radiografijo, penetranti idr. Preiskave se opravljajo tako med samo proizvodnjo, kot tudi kasneje med obratovanjem. Vzorce, ki ne dosegajo zahtev po različnih standardih, se zavrže že med proizvodnjo, s čimer se doseže višja stopnja zanesljivosti končnega izdelka. Med obratovanjem je s pomočjo *on-line* preiskave vzorca možno spremljati njegovo obrabo, nastanek različnih razpok itd. ter tako povečati zanesljivost obratovanja in obenem napovedati morebitni remont. Če *online* spremljanje ni možno, se vzorec podvrže preiskavi na periodičnih pregledih. V strokovni literaturi se pogosto pojavi tudi kratica NDE (Non-Destructive Evaluation), ki je v bistvu nadpomenka za NDT. NDE zahteva pridobitev dodatnega geometrijskega parametra napake iz podatkov, pridobljenih v preiskavi, medtem ko NDT zahteva le potrditev, da gre za nepravilnost v vzorcu [1].

2 Princip delovanja preiskave

Za preiskavo z vrtinčnimi toki je potrebno ustvariti časovno spreminjajoče magnetno polje, ki v prevodnem materialu inducira vrtinčne toke. Ti toki povzročijo magnetno polje, ki po smeri nasprotuje vzbujalnemu. Na mestu, kjer je v materialu napaka, se inducira manj vrtinčnih tokov in posledično je tudi manjše nasprotujoče magnetno polje. Z merjenjem inducirane napetosti v navitju sonde je možno odkriti rahlo deviacijo v skupnem magnetnem polju, ko je sonda nad napako [2]. Izmenično magnetno polje se vzbuja s konstantnim sinusnim tokom nizke frekvence skozi vzbujalno navitje. Vdorna globina δ, ki jo izračunamo z $\delta = \sqrt{(\pi f \mu \sigma)^{-1}}$, je poleg frekvence f odvisna tudi od permeabilnosti μ in prevodnosti σ materiala. Pri preiskavi feromagnetnih materialov je zaradi visoke relativne permeabilnosti zelo pomembno, da se uporablja nizka frekvenca, reda velikosti nekaj Hz. Ker se preiskuje feromagnetni material, je potrebno paziti tudi na amplitudo vzbujalnega toka, da magnetenje ne doseže nasičenja na BH-karakteristiki materiala. Po drugi strani pa mora biti ta tok dovolj velik, da v vzorcu inducira ustrezno količino vrtinčnih tokov.

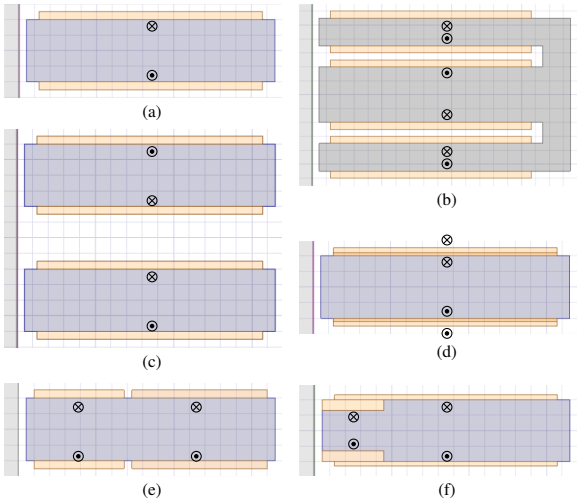
V članku je predstavljena metoda, po kateri se vzorec preiskuje po korakih, kar pomeni, da se meritev izvaja medtem ko sonda miruje. V vsaki točki sta zabeležena dva parametra detekcijske napetosti (amplituda in fazni kot glede na vsiljeni tok). Odkrivanje napake temelji na odklonu parametrov od povprečne vrednosti, ki nastopi, ko je sonda nad napako. S tako metodo ni potrebno določiti absolutne vrednosti parametrov napetosti ob napaki, ampak le mejno vrednost odklona parametrov od povprečne vrednosti. S tem se izniči vpliv spremembe permeabilnosti materiala.

Poleg beleženja parametrov detekcijske napetosti se sproti izračunavata tudi gradienta spreminjanja amplitude in faznega kota detekcijske napetosti. Majhen gradient predstavlja počasno spremembo teh dveh veličin skozi meritve. To je značilno za nehomogene materiale, kjer prevodnost in permeabilnost materiala nista enaki po vsem volumnu. Višji gradient pa predstavlja hitro spremembo razmer v materialu. Sem sodita približevanje sonde robu materiala ali pa približevanje napaki v materialu. Zaradi tega sta poleg samih potekov vrednosti parametrov detekcijske napetosti med preiskavami vedno prikazana tudi njuna gradienta.

3 Konstrukcija in evalvacija sonde v simulacijskem modelu

Za preiskovanje materialov z vrtinčnimi toki se uporablja več različnih sond. Deliti jih je možno po izvedbi

(površinske, za izvrtine, za pregledovanje notranjih sten cevi, za pregledovanje palic) in po načinu delovanja (absolutne, diferencialne, refleksijske in hibridne) [3]. Zato je prvi korak konstrukcije sonde izbira pravilnega načina delovanja sonde. V ta namen je bil razvit 2D simulacijski model v programskem paketu Ansys Simplorer, ki temelji na metodi končnih elementov. Opravljenih je bilo več izračunov z različnimi sondami, ki so označene na sliki 1.



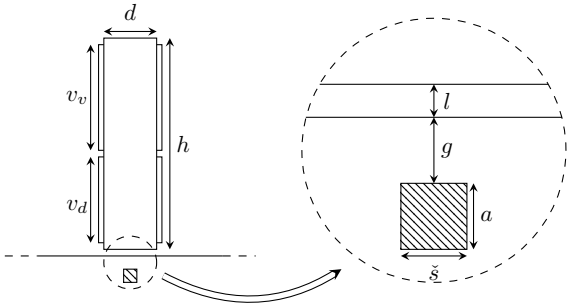
Slika 1: Sonde za izvajanje preiskav: (a) absolutna, (b) hibridna, (c) diferencialna, (d) refleksijska v2, (e) refleksijska v1, (f) refleksijska v3.

Vse različice sonde so bile napajane s konstantnim sinusnim tokovnim virom in na vseh je bila izmerjena inducirana napetost na navitju sonde. Da se lahko izvrši primerjava med različnimi tipi sond, je vse potrebno oceniti po enakem kriteriju. Izbran je bil kriterij SNR (Signal-to-noise ratio) [4]. Algoritem poišče največji odklon od povprečne vrednosti parametra detekcijske napetosti med meritvami, ki so zajete v okolici napake, in ga primerja z najvišjim odklonom od povprečne vrednosti med meritvami, ki so bile zajete daleč stran od napake. Razmerje SNR je bilo izračunano posebej za vsak parameter napetosti. Kot optimalna izbira glede na pridobljene rezultate (visok SNR) v tabeli 1 in enostavnost izdelave se je izkazala refleksijska sonda v1 (slika 1(e)). Značilnost refleksijske sonde je, da ima ločeno vzbujalno in detekcijsko navitje. To pomeni, da je možno vsako navitje posebej optimirati. Za vzbujalno navitje se lahko uporabi nekoliko debelejša žica, ki prevaja višji tok, detekcijsko navitje pa je navito z zelo tanko žico. S tem lahko na enaki površini naredimo več ovojev, kar zviša inducirano napetost v detekcijskem navitju. Čeprav je hibridna sonda v simulacijah pokazala veliko višje vrednosti SNR, je bila izbrana refleksijska sonda v1, ker le-ta povzroči veliko ožji odklon (indikacijo) ob napaki. Prednost ožje indikacije je predvsem v primerih, ko je več manjših napak na majhni medsebojni razdalji.

Ko je bila izbrana osnovna postavitev navitij v sondi, je bilo potrebno preveriti še vplive vseh parametrov sonde na uspešnost detekcije napake. Poleg geometrijskih parametrov, označenih na sliki 2, so bili preverjeni še vplivi amplitude in frekvence vzbujalnega toka, števila ovojev posameznih navitij, permeabilnosti vzorca in jedra sonde ter oblike jedra sonde.

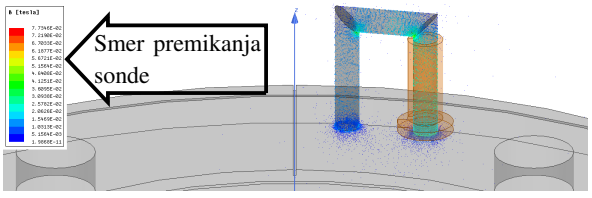
Tabela 1: Primerjava SNR gradienta osnovnega signala.

	Abs.	Dif.	Ref v1.	Ref v2.	Ref v3.	Hib.
SNR_{amp}	83	36	196	56	159	87
SNR_{θ}	10	38	21	1	22	110

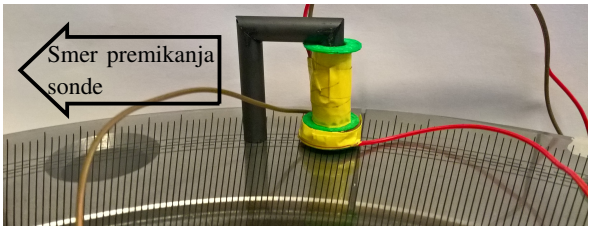


Slika 2: Geometrijski parametri sonde in merjenca.

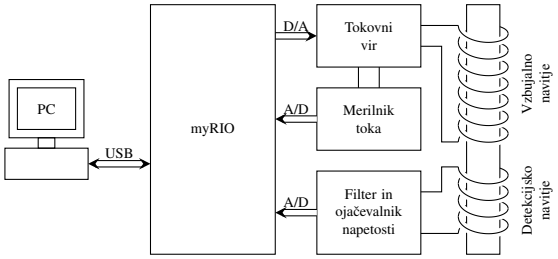
Za izračun vplivov vsakega posameznega parametra sonde je bilo izvedenih mnogo simulacij. V vsaki izmed njih se je naenkrat spremenil le en parameter, preostali pa so imeli nespremenjene privzete vrednosti. Ker je bilo v tem koraku potrebno le preveriti trend SNR, ki ga povzroči posamezen parameter, je bila večina simulacij izvedenih v 2D modelu (npr. če se globina napake prepolovi, se SNR_{amp} ne spremeni, SNR_{θ} pa pade za približno 30%). Vplivi vsakega posameznega parametra so predstavljeni v [4]. Absolutne vrednosti napetosti v 2D simulacijah niso enake tistim na realnem modelu, zato jih je potrebno pomnožiti s prilagoditvenim faktorjem. Največja prednost 2D simulacij je, da so zelo hitre pri izračunih, kar omogoči večje število preverjenih parametrov v enakem časovnem obdobju. Za vse parametre ni bilo možno preveriti trendov v 2D kartezičnem koordinatnem sistemu, zato je bilo nekaj izračunov izvedenih tudi v valjnem koordinatnem sistemu. V zaključni fazi simulacijskega procesa je bil narejen tudi 3D model, ki omogoča pregled trendov, ki jih ni možno preveriti v 2D simulacijah, primerjavo simulacijskih izračunov z meritvami in nenazadnje sledljivost rezultatov od meritev vse do 2D modela. Slika 3 predstavlja končni simulacijski model. Končna oblika sonde, ki je bila predlagana v simulacijah, je bila tudi fizično izdelana in je prikazana na sliki 4.



Slika 3: Simulacijski model sonde za preiskovanje feromagnetnega diska.



Slika 4: Izdelana sonda za preiskovanje feromagnetnega diska.



Slika 5: Shema postavitve merilnih instrumentov.

4 Meritve na realnem vzorcu

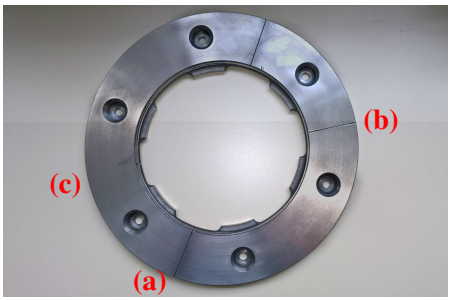
4.1 Merilni sistem

Za praktično opravljanje meritev je bilo potrebno konstruirati poleg same sonde tudi vsa pomožna električna vezja, ki so skrbela za napajanje in merjenje električnih veličin na sondi. Izdelan je bil konstantni sinusni tokovni vir v vezju izboljšane Howlandove vezave, ojačevalnik detekcijske napetosti skupaj z nizkoprepustnim filtrom ter merilnik toka. Pri konstrukciji napajalnika je zelo pomembno, da ta v vzbujalno navitje pošilja tok s čim manj višjimi harmoniki. Ti povzročajo težave pri izračunu faznega kota sinusnega signala, ki je pomemben za detekcijo napake v vzorcu. Tudi vezji za zajemanje električnega toka in napetosti morata signal dobro filtrirati, da je izračun faznega kota čim bolj natančen. Sprememba faznega kota ob napaki v vzorcu se meri v desetinkah kotnih stopinj, sprememba amplitude detekcijske napetosti pa v milivoltih.

Krmiljenje tokovnega vira in zajem električnih veličin je naloga algoritma v programskem paketu LabVIEW. Računalniški program zajema časovne poteke vzbujalnega toka in detekcijske napetosti ter iz njiju izračunava amplitudi ter fazni kot med njima. Za izračun amplitudne vrednosti program izračuna efektivno vrednost, katero pomnoži s $\sqrt{2}$, za izračun faznega kota pa uporablja Fourierovo vrsto. Poleg osnovnih parametrov algoritem izračunava še njuna gradienta. Izdelan je bil tudi celoten uporabniški vmesnik, preko katerega se lahko spreminja parametre vzbujalnega toka. Program v realnem času prikazuje rezultate meritev, iz katerih je možno razbrati, ali se pod sondo nahaja napaka. Komunikacija s pomožnimi vezji poteka skozi razvojno ploščo myRIO. Na razvojni plošči so uporabljeni en digitalno-analogni pretvornik kot izhod ter dva analogno-digitalna pretvornika kot vhoda. Pretvorniki delujejo na območju ± 10 V z 12-bitno ločljivostjo [5]. To pomeni, da je napetost na vhodu možno odčitati/nastaviti z ločljivostjo 4,9 mV. Zaradi tega so bili v pomožnih vezjih uporabljeni ojačevalniki z visokim faktorjem ojačanja, da je bilo možno izmeriti tako majhne razlike v parametrih detekcijske napetosti. Na sliki 5 je prikazana celotna shema povezave merilnih instrumentov.

4.2 Testni vzorec

Pri meritvah je bil kot vzorec uporabljen zavorni disk iz sive litine, na katerem so bile namerno povzročene tri napake. Disk je bil zarezan z žago na površini do globine 2 in 4 mm ter z zadnje strani do 2 mm pod površino. Tako sta bili umetno narejeni dve površinski in ena napaka pod površino. Sonda se je po površini vzorca pomikala tako, da je bila razdalja med merilnima mestoma 2 mm. Smeri premikanja sonde sta označeni že na slikah 3 in 4. Kot je razvidno s teh dveh slik, napako naprej prečka steber sonde, ki ne nosi navitja in nato še steber, ki nosi navitje.



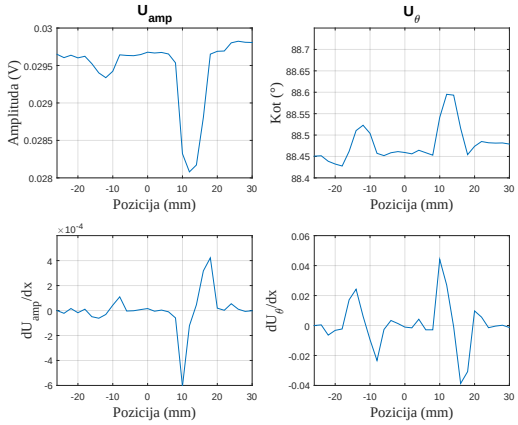
Slika 6: Zavorni disk z napakami kot vzorec za preskušanje: (a) površinska 2 mm, (b) površinska 4 mm, (c) 2 mm pod površino.

5 Primerjava meritev in simulacijskih izračunov

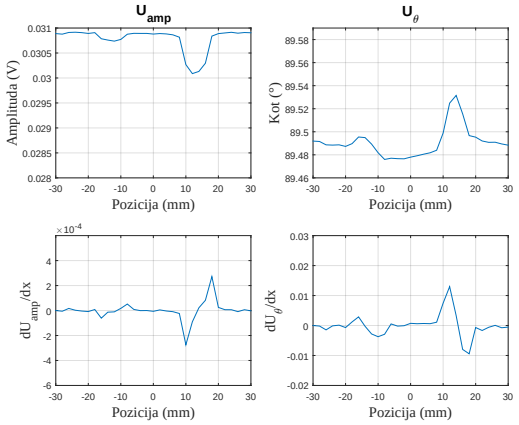
5.1 Rezultati meritev

Na sliki 7 sta prikazana poteka parametrov ob pregledu vzorca s sondo. V prvi vrstici grafa prikazujeta izračunana parametra detekcijske napetosti, druga vrstica pa pripadajoče gradiente. V levem stolpcu sta prikazana poteka amplitude detekcijske napetosti, desna pa potek faznega kota. Odklon od povprečne vrednosti (indikacijo) se lahko opazi kot znižanje amplitude in povišanje faznega kota detekcijske napetosti. Levo indikacijo na vseh štirih grafih povzroči prehod stebra sonde brez navitja čez napako. Desno, večjo indikacijo, pa povzroči prehod stebra sonde, ki nosi detekcijsko navitje. Oba odklona sta na grafih za 12 mm oddaljena od središča, saj je koordinatno izhodišče točno na sredini napake. Indikaciji se pojavita izmaknjeni za točno toliko, kolikor znaša vsota polovice razdalje med stebroma in radija stebra.

Enaki meritvi sta bili opravljeni tudi na ostalih dveh poškodbah na disku. Absolutne vrednosti maksimalnega odklona posameznega parametra od povprečne vrednosti so podane v tabeli 2. Opazno je, da je sprememba amplitude pri obeh napakah na površini enaka, ne glede na globino napake. Razlika je le v odklonu faznega kota. Iz



Slika 7: Spreminjanje parametrov napetosti ob napaki na površini globine 2 mm - meritev



Slika 8: Spreminjanje parametrov napetosti ob napaki na površini globine 2 mm - simulacija

tega lahko zaključimo, da je globino napake možno oceniti na podlagi spremembe v faznem kotu. Pri dvakratni globini napake je odklon faznega kota približno dvakrat večji. Pri meritvi spremembe amplitude nad napako, ki je pod površino, pa se odklon v amplitudi prepolovi. V tem primeru se je bolje zanesti na spremembo v faznem kotu, saj je ta večja. Ta je skoraj enaka tisti pri 4 mm globoki napaki na površini.

Tabela 2: Sprememba amplitude in faznega kota ob napaki - meritev

	ΔU_{amp} [mV]	ΔU_θ [°]
napaka na površini (2 mm)	-1.5	+0.12
napaka na površini (4 mm)	-1.5	+0.20
napaka 2 mm pod površino	-0.8	+0.17

5.2 Simulacijski izračuni

Z namenom primerjave simulacijskih rezultatov s tistimi, pridobljenimi na praktičnem modelu, je bil narejen simulacijski model, ki kar najbolj posnema realne razmere. Model je izveden v 3D prostoru. Na tem modelu so bile preskušane enake napake kot pri meritvah. Da sta poteka parametrov podobna, kaže slika 8. Glede na meritve je opazno, da je povprečna vrednost amplitude nekoliko višja. Do razlike pride zaradi nepoznavanja natančne vrednosti permeabilnosti materiala. Višja permeabilnost v simulacijskem modelu rezultira v navzgor premaknjeni karakteristiki. Tudi med potekoma faznega kota obstaja razlika med povprečnima vrednostima parametra. Ta razlika nastane zaradi filtra, ki je na vhodu merilnika detekcijske napetosti.

Absolutne vrednosti odklonov od povprečne vrednosti za oba parametra detekcijske napetosti so bile v simulacijah nekoliko manjše za vse tri napake. To je posledica žaganja materiala in tako lokalno spremenjene (znižane) permeabilnosti v okolici napake na zavornem disku. Ob pregledu razmerij med rezultati, pridobljenimi z meritvami in izračuni, pa lahko zaključimo, da se rezultati skladajo. Tako je tudi v simulacijskem modelu ob dvakrat globlji površinski napaki odklon faznega kota približno dvakrat večji.

6 Zaključek

Metoda z vrtinčnimi toki ne more povsem nadomestiti drugih metod neporušnega testiranja, lahko pa jih odlično dopolnjuje. Ena izmed največjih prednosti te metode je predvsem velika možnost avtomatizacije. Metoda predvideva zaznavo indikacije napake z električnim signalom. Zaradi tega je možno izvajanje preiskav prepustiti mikrokrmilniku.

Izbira parametra napetosti, po katerem je zaznava napak zanesljivejša, je odvisna od vrste napak, ki jih poskušamo odkriti. Pri površinskih napakah se je smiselno v večji meri zanesti prav na parameter amplitude, pri napakah pod površino pa na parameter faznega kota. V splošnem je seveda potrebno opazovati in ustrezno reagirati na oba parametra.

Dosedanje delo dokazuje, da je detekcija napak z vrtinčnimi toki možna. Sistem zaenkrat deluje v laboratorijskih razmerah, kjer je relativno malo motenj. Za praktično uporabo sistema je sondo in vezja najprej potrebno nadgraditi glede robustnosti. K aplikaciji te metode bi zelo pripomogla tudi avtomatizacija premikanja sonde po površini vzorca in programsko prepoznavanje napak na vzorcu. Sem sodi razširitev algoritma, ki bi izračunaval geometrijske razsežnosti napak.

Literatura

[1] J. Prasad and C. Gouri Krishnadas Nair: Non-Destructive Test And Evaluation Of Materials, Tata McGraw-Hill, 2009.

[2] J. García-Martín, J. Gómez-Gil in E. Vázquez-Sánchez: Non-destructive techniques based on eddy current testing, Sensors, vol. 11, no. 3, 2011, pp. 2525-2565.

[3] Iowa State University, Center for NDE: ECT NDT Course Material, Dosegljivo: https://www.nde-ed.org/EducationResources/CommunityCollege/EddyCurrents/cc_ec_index.htm, Dostopano: 9.10.2017.

[4] M. Markl: Detekcija napak v feromagnetnih materialih z metodo vrtinčnih tokov, Magistrsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani, 2018.

[5] National Instruments: User guide and specifications, NI myRIO-1900, Datasheet, 2016.

Vodenje dveh zaporedno vezanih sinhronskih motorjev s trajnimi magneti z enim frekvenčnim pretvornikom

Matej Čadež, Rastko Fišer, Henrik Lavrič, Klemen Drobnič

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: mc2881@student.uni-lj.si, rastko.fiser@fe.uni-lj.si, henrik.lavric@fe.uni-lj.si, klemen.drobnic@fe.uni-lj.si

Control of dual serial-connected PMSM by single frequency converter

The main aim of this paper is to analyse dual serial-connected permanent magnet synchronous motors (PMSM) controlled by single frequency converter (FC) and to develop a control method which will ensure stable operation of the system when different load torque is applied to each PMSM. First, basic aspects of multi-connected motors controlled by single FC are presented, followed by an overview of control solutions for parallel-connected PMSM. Mathematical model of dual serial-connected PMSM is given in section 3. The stability analysis of the system described in section 4 showed that change of magnetization is needed in order to stabilise the operation of dual serial-connected PMSM. In the final part, different solutions of magnetizing current controller are proposed and substantiated with simulation results.

1 Uvod

Pri načrtovanju sodobnih elektromotorskih pogonov pogosto naletimo na zahtevo po njihovi čim nižji ceni. Eno izmed mogočih rešitev predstavlja vodenje večjega števila izmeničnih motorjev z enim frekvenčnim pretvornikom (FP), ki jo je mogoče aplicirati v pogonih, kjer se motorji v stacionarnem stanju vrtijo z enako vrtilno hitrostjo, opravljajo enako ali podobno funkcijo in so po zgradbi ter lastnostih enaki. Takšne aplikacije so npr. pogoni ventilatorjev, črpalk, transportnih trakov in električna vleka tirnih vozil. Kljub cenovno ugodni rešitvi zaradi manjšega števila potrebnih pretvorniških stikal, omrežnih in izhodnih filtrov FP, se je potrebno zavedati pomanjkljivosti predlagane konfiguracije, kot npr. nezmožnost neodvisne regulacije posameznih motorjev, s čimer se zmanjša fleksibilnost sistema, in potreba po večji moči FP.

Ena izmed glavnih zahtev vezave enega FP in večjega števila izmeničnih motorjev je zagotoviti njihovo stabilno obratovanje v primeru različne obremenitve posameznih motorjev. Sprememba bremenskega navora pri enem motorju povzroči električno motnjo, ki jo občutijo vsi motorji v sistemu. Pri uporabi asinhronskih motorjev se zaradi slipa ne srečamo s problemom stabilnosti sistema, ki pa je močno prisoten pri sinhronskih motorjih (SM), katerih rotorji se vrtijo v sinhronizmu z vrtilnim magnetnim poljem, katerega frekvenca je določena s frekvenco napetosti razsmernika. V primeru povečanja bremenskega navora se SM poveča kolesni kot δ. To je kot med vektorjema inducirane in statorske napetosti. Le-ta ne sme preseči vrednosti $\pi/2$, saj takrat SM preide v nestabilno delovanje in pade iz sinhronizma.

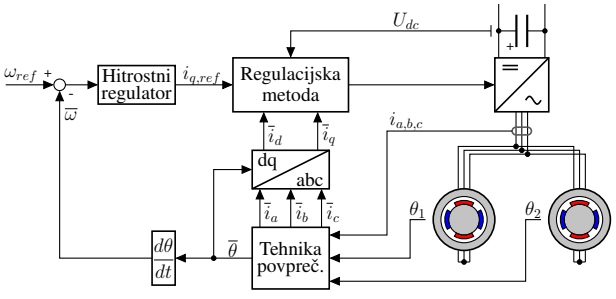
V pričujočem članku se posvetimo analizi vodenja dveh zaporedno vezanih sinhronskih motorjev s površinsko nameščenimi trajnimi magneti (SMTMP) z enim FP s ciljem razvoja ustrezne strategije vodenja, ki bo zagotavljala stabilno obratovanje sistema pri različno obremenjenih SMTMP. V poglavju 2 sta predstavljena koncepta dveh osnovnih strategij vodenja dveh vzporedno vezanih SMTMP. Temu v poglavju 3 sledi izpeljava matematičnega modela zaporedne vezave dveh SMTMP. Kratek opis stabilnostne analize takšne konfiguracije podaja poglavje 4. Simulacijski rezultati preizkusa različnih načinov nastavljanja magnetilnega toka (i_d), s katerimi lahko zagotovimo stabilnost sistema, so opisani v poglavju 5.

2 Rešitve pri vzporedni vezavi

Zaradi pomanjkanja tuje literature o vodenju dveh enakih zaporedno vezanih SMTMP, smo preverili obstoječe rešitve pri vzporedni vezavi. V nadaljevanju opisani strategiji lahko z določenimi popravki apliciramo tudi pri zaporedni vezavi.

2.1 Tehnika povprečenja (angl. *Average Technique*)

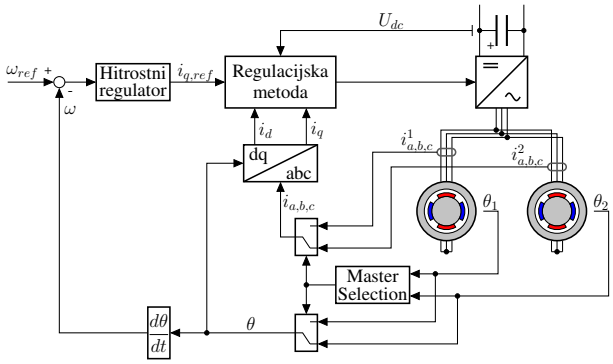
V [1] in [2] najdemo najenostavnejši pristop, katerega osnovni princip podaja shema na sliki 1. SMTMP 1 in SMTMP 2 sta z regulacijskega vidika na podlagi njunih povprečnih vrednosti tokov, kota zasuka in kotne hitrosti rotorja združena v ekvivalentni SMTMP. Pravzaprav gre za klasično hitrostno regulacijo enega SMTMP, kateri dodamo izračun povprečnih vrednosti veličin.



Slika 1: Shema tehnike povprečenja pri vzporedni vezavi dveh SMTMP.

2.2 Izbor nadrejenega motorja (angl. *Master/Slave*)

Osnovna ideja te strategije je v vsakem vzorčnem intervalu izbrati bolj obremenjen motor, saj ta, zaradi potrebnega večjega električnega navora, zahteva večji tok oz. moč. SMTMP, ki je določen kot nadrejeni motor (angl. *master*), je zaprtozančno reguliran. Preostali SMTMP, ki prevzame vlogo podrejenega motorja (angl. *slave*), je voden odprtozančno. Tako tudi pri tej strategiji preidemo z regulacije dveh SMTMP na klasično regulacijo enega. Obstajajo različni načini, kako izbrati nadrejeni motor. V [3] in [4] je podrobneje opisan izbor s primerjanjem kota zasuka rotorjev obeh SMTMP.



Slika 2: Shema izbora nadrejenega motorja pri vzporedni vezavi dveh SMTMP.

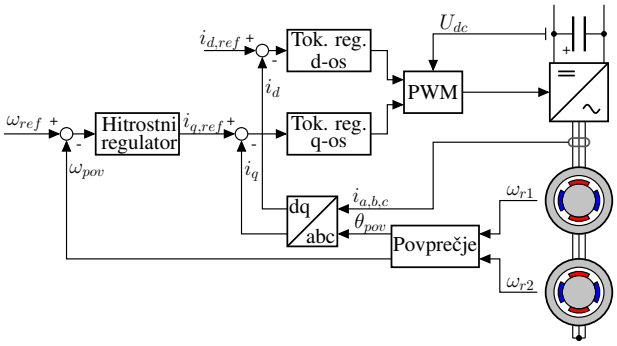
3 Model zaporedne vezave dveh SMTMP

Matematični model zaporedne vezave dveh SMTMP razvijemo na podlagi tehnike povprečenja, saj ima le-ta v primerjavi z izborom nadrejenega motorja določene prednosti:

- zaradi enostavnega koncepta ni teoretičnih zadržkov za njeno apliciranje pri zaporedni vezavi SMTMP,
- omogoča enostavnejšo razširitev na vodenje večjega števila zaporedno vezanih SMTMP,
- izvedba brezsenzorske regulacije zahteva le oceno povprečne kotne hitrosti, kar pri praktični realizaciji pomeni uporabo manjšega števila merilnikov toka oz. napetosti kot v primeru izbrane strategije izbore nadrejenega motorja.

Blokovno shemo modela s tehniko povprečenja prikazuje slika 3. Regulacija v koordinatnem sistemu polja (KSP) temelji na povprečnih vrednostih tokov, kotne hitrosti in kota zasuka rotorjev. Ob tem predpostavimo, da kotno hitrost oz. pozicijo rotorja posameznega SMTMP merimo neposredno z dajalnikom kota zasuka. Čeprav imamo v tem primeru opravka z mehanskimi kotnimi hitrostmi in koti zasuka rotorja (ω_{m1}, ω_{m2}, θ_{m1}, θ_{m2}), pri teoretični izpeljavi enačb modela uporabimo električne veličine. Dejansko kotno hitrost SMTMP 1 in SMTMP 2 označimo z ω_{r1} oz. ω_{r2}. Povprečno kotno hitrost ω_{pov} izračunamo z enačbo

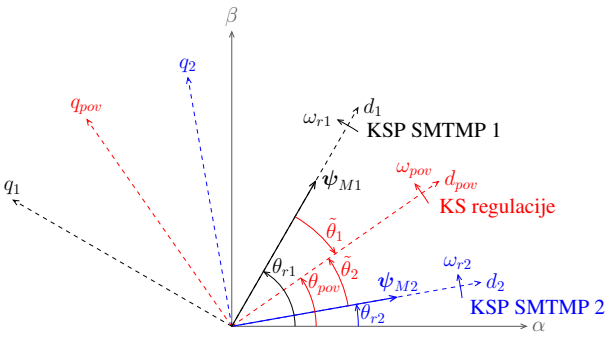
$$\omega_{pov} = \frac{\omega_{r1} + \omega_{r2}}{2}. \tag{1}$$



Slika 3: Shema tehnike povprečenja pri zaporedni vezavi dveh SMTMP.

Pri regulaciji se kot povratna informacija o rotorskem kotu uporabi povprečna vrednost θ_{pov}, ki definira pozicijo koordinatnega sistema (KS) regulacije. KSP SMTMP 1 in KSP SMTMP 2 ne sovpadata nujno s KS regulacije. Ker je pravilna informacija o položaju rotorja nujna za pravilno delovanje modela, je potrebno definirati kota $\tilde{\theta}_1$ in $\tilde{\theta}_2$ (enačba (2)), ki podajata razliko kotov med KS regulacije in KSP posameznega SMTMP. Lego vseh treh KS za opisan primer podaja kazalčni diagram na sliki 4.

$$\begin{aligned} \frac{d\tilde{\theta}_1}{dt} &= \frac{d(\theta_{r1} - \theta_{pov})}{dt} = \omega_{r1} - \omega_{pov} \\ \frac{d\tilde{\theta}_2}{dt} &= \frac{d(\theta_{r2} - \theta_{pov})}{dt} = \omega_{r2} - \omega_{pov} \end{aligned} \tag{2}$$



Slika 4: Kazalčni diagram lege KS pri tehniki povprečenja.

Napetostni enačbi za SMTMP 1 v KS regulacije sta

$$\begin{aligned} u_{d1} &= R_S i_d + L_S \frac{di_d}{dt} - \omega_{pov} L_S i_q - \omega_{r1} \psi_M \sin\tilde{\theta}_1 \\ u_{q1} &= R_S i_q + L_S \frac{di_q}{dt} + \omega_{pov} L_S i_d + \omega_{r1} \psi_M \cos\tilde{\theta}_1. \end{aligned} \tag{3}$$

Podobno sta v KS regulacije določeni tudi napetostni enačbi (4) za SMTMP 2.

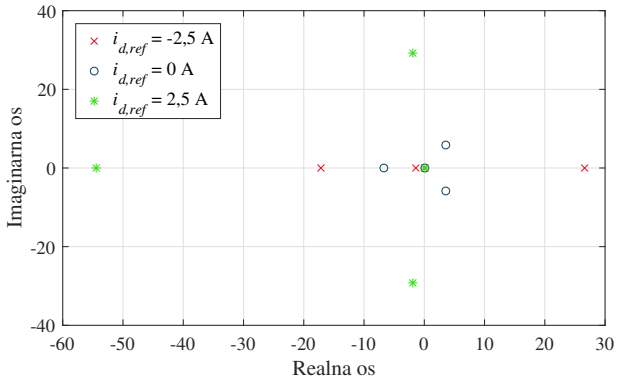
$$\begin{aligned} u_{d2} &= R_S i_d + L_S \frac{di_d}{dt} - \omega_{pov} L_S i_q - \omega_{r2} \psi_M \sin\tilde{\theta}_2 \\ u_{q2} &= R_S i_q + L_S \frac{di_q}{dt} + \omega_{pov} L_S i_d + \omega_{r2} \psi_M \cos\tilde{\theta}_2 \end{aligned} \tag{4}$$

S primerjanjem enačb (3) in (4) lahko ugotovimo, da so tokovi enaki v obeh sklopih enačb. Če enačbi seštejemo, dobimo skupno napetostno enačbo, ki jo nato preuredimo

za implementacijo v programu Simulink. Opisan postopek je podrobneje predstavljen v [5], kjer sta podani tudi navorni in mehanski enačbi posameznega SMTMP.

4 Stabilnostna analiza

Preizkus simulacijskega modela je potrdil nestabilno obratovanje pogona pri različni obremenitvi SMTMP. Iz tega razloga je potrebno opraviti analizo stabilnosti sistema, katere rezultat bo lega polov v kompleksni ravnini. Slednjo naredimo s pomočjo orodja *Linear Analysis Tool*, ki je že integrirano v programu Simulink. Z njim linearizirajmo sistem v izbrani delovni točki. Le-to izberemo pri referenčni vrtilni hitrosti $n_{ref} = 2000\,\mathrm{min}^{-1}$ in razliki v obremenitvi SMTMP $10\,\%\ M_n$. V [6] je ugotovljeno, da pozitivna d komponenta toka zagotovi stabilno obratovanje dveh vzporedno vezanih SMTMP, ki sta vodena z enim FP. Za ovrednotenje njegovega vpliva pri zaporedni vezavi izberemo tri vrednosti $i_{d,ref}$, in sicer $-2{,}5\,\mathrm{A}$, $0\,\mathrm{A}$ ter $2{,}5\,\mathrm{A}$, s katerimi najbolj reprezentativno prikažemo premikanje polov. Na sliki 5 so prikazani poli v bližini koordinatnega izhodišča, ki določajo stabilnost sistema. Razvidno je, da se pri pozitivnem toku i_d vsi karakteristični poli nahajajo na levi polovici kompleksne ravnine, kar pomeni, da je sistem v tem primeru stabilen. Dodatno smo preverili tudi obnašanje sistema pri različnih referenčnih hitrostih ($500\,\mathrm{min}^{-1}$, $1000\,\mathrm{min}^{-1}$ in $1500\,\mathrm{min}^{-1}$). Ugotovljeno je bilo, da n_{ref} nima vpliva na stabilnost sistema.



Slika 5: Lega polov v področju blizu imaginarne osi pri $n_{ref} = 2000\,\mathrm{min}^{-1}$ v odvisnosti od toka $i_{d,ref}$.

5 Regulator $i_{d,ref}$

Na podlagi rezultata stabilnostne analize je smiselno razviti ustrezen regulator, s katerim nastavimo minimalno potrebno pozitivno referenčno vrednost d komponente toka, ki bo zagotovila stabilno obratovanje zaporedno vezanih SMTMP. S tem uspemo znižati dodatne toplotne izgube v navitjih motorjev zaradi dodatnega toka i_d, ki je pri regulaciji enega SMTMP običajno enak nič, razen v primeru slabljenja polja. Za preizkus različnih rešitev nastavljanja $i_{d,ref}$, ki so podrobneje opisane v nadaljevanju, določimo profil obremenitve posameznega SMTMP, ki ga podaja tabela 1. Zaradi želje po univerzalni rešitvi so spremembe bremenskega navora skočne. Poteki toka $i_{d,ref}$ in vrtilne hitrosti, ki pripadajo posamezni izvedbi regulatorja, so prikazani skupaj na slikah 6 in 7, saj s tem slikoviteje prikažemo ustreznost posamezne izvedbe.

Tabela 1: Profil obremenitve posameznega SMTMP.

t (s)	0,1 - 2	2 - 6	6 - 10	10 - 14	14 - 18
SMTMP 1 (% M_n)	90	90	100	80	100
SMTMP 2 (% M_n)	80	110	70	120	90

5.1 Konstantna vrednost $i_{d,ref}$

Nastavitev zadosti velike pozitivne vrednosti $i_{d,ref}$ omogoča stabilno delovanje sistema za širok razpon različnih obremenitev obeh SMTMP. Z dodatnimi testiranji je bilo ugotovljeno, da sistem prenese razliko v obremenitvi tudi do $90\,\%\ M_n$. S slike 7 (a) je razvidno, da so oscilacije vrtilne hitrosti manjše kot pri drugih izvedbah z izjemo zadnje rešitve. Bistvena pomanjkljivost tega načina je prevelik tok i_d, kot bi bil večino časa potreben.

5.2 Skaliran tok i_q

V [7] je tok i_d po obliki podoben i_q, zato smo prišli na idejo, da referenčni tok $i_{d,ref}$ izračunamo kot absolutno vrednost razlike med referenčnim tokom $i_{q,ref}$, ki se spreminja glede na trenutno obremenitev SMTMP, in konstantnim nazivnim tokom $i_{q,n}$, kar podaja enačba (5). S tem ocenimo razliko v trenutni obremenitvi obeh motorjev glede na nazivno stanje, pri katerem ni potrebe po pozitivnem toku i_d. Nazivni tok $i_{q,n}$ je določen z (6).

$$i_{d,ref} = K_1 |i_{q,ref} - i_{q,n}| \tag{5}$$

$$i_{q,n} = \frac{2}{3}\frac{M_n}{p_p \psi_M} \tag{6}$$

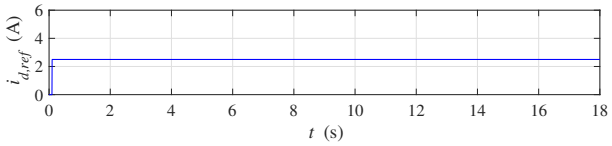
Na slikah 6 (b) in 7 (b) lahko opazimo, da se ob času $t = 6\,\mathrm{s}$ pojavi nedušeno nihanje toka $i_{d,ref}$ oz. vrtilne hitrosti n. Samo informacija o toku i_q torej ni dovolj, zato je potrebno dodati še nek podatek, ki nosi informacijo, da je v sistemu prišlo do določene spremembe bremenskega navora, in bo povzročil takojšnje povečanje toka $i_{d,ref}$.

5.3 Skaliran tok $i_{q,ref}$ z dodanim odvodom $u_{q,ref}$

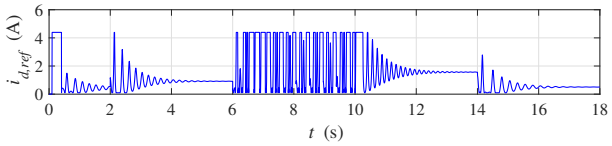
Ena izmed možnih rešitev je, da skaliranemu toku $i_{q,ref}$ dodamo odvod referenčne napetosti $u_{q,ref}$, ki ga aproksimiramo z razliko vrednosti $u_{q,ref}$, ki sta časovno zamaknjeni za 5 vzorčnih intervalov ($t_{vz} = 0{,}1\,\mathrm{ms}$). Trajanje časovnega zamika je določeno eksperimentalno. V k-tem vzorčnem intervalu izračunamo $i_{d,ref}$ po enačbi (7).

$$\begin{aligned} i_{d,ref}[k+1] = & K_1 |i_{q,ref}[k] - i_{q,n}| + \\ & K_2 |u_{q,ref}[k] - u_{q,ref}[k-5]| \end{aligned} \tag{7}$$

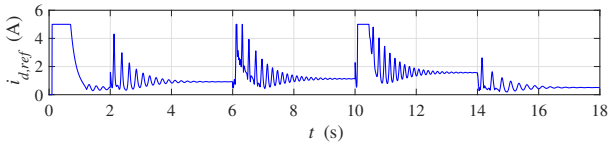
Takšna izvedba regulatorja $i_{d,ref}$ omogoča stabilno delovanje obeh SMTMP v omejenem področju obratovanja, kjer razlika v njuni obremenitvi SMTMP ne sme preseči $40\,\%\ M_n$. Iz oscilogramov na slikah 6 (c) in 7 (c) je razvidno, da je dušenje prehodnega pojava še vedno relativno dolgotrajno, zato smo poskusili najti boljšo rešitev.



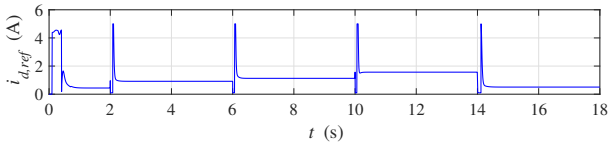
(a) Konstantna vrednost $i_{d,ref}$.



(b) Skaliran tok $i_{q,ref}$.

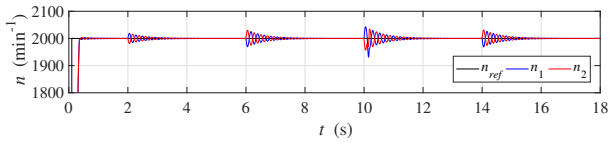


(c) Skaliran tok $i_{q,ref}$ z dodanim odvodom $u_{q,ref}$.

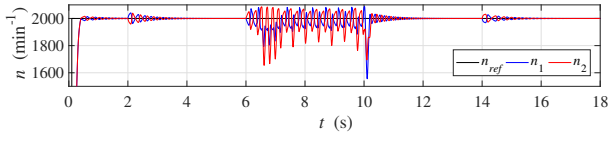


(d) Skaliran tok $i_{q,ref}$ z dodano razliko hitrosti.

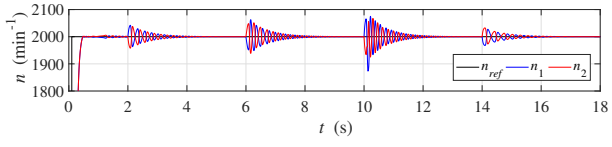
Slika 6: Poteki $i_{d,ref}$ pri posamezni izvedbi regulatorja referenčne vrednosti d komponente toka.



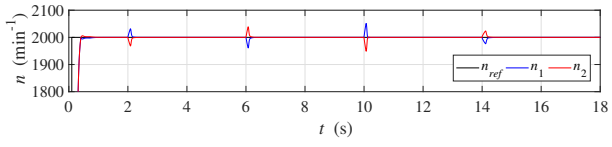
(a) Konstantna vrednost $i_{d,ref}$.



(b) Skaliran tok $i_{q,ref}$.



(c) Skaliran tok $i_{q,ref}$ z dodanim odvodom $u_{q,ref}$.



(d) Skaliran tok $i_{q,ref}$ z dodano razliko hitrosti.

Slika 7: Poteki vrtilnih hitrosti obeh SMTMP pri posamezni izvedbi regulatorja $i_{d,ref}$.

5.4 Skaliran tok $i_{q,ref}$ z dodano razliko hitrosti

To dobimo, če namesto odvoda $u_{q,ref}$ skaliranemu toku $i_{q,ref}$ dodamo razliko hitrosti med podrejenim in nadrejenim motorjem (oz. manj in bolj obremenjenim motorjem), kar podaja enačba (8).

$$i_{d,ref} = K_1|i_{q,ref} - i_{q,n}| + K_2(\omega_{slave} - \omega_{master}) \quad (8)$$

V tem primeru je dušenje oscilacij toka $i_{d,ref}$ (slika 6 (d)) in vrtilne hitrosti (slika 7 (d)) skorajda hipno z manjšim prenihajem. Takšna rešitev omogoča stabilno obratovanje sistema do 90 % razlike v obremenitvi obeh SMTMP glede na M_n. Pomanjkljivost te izvedbe je potreba po informaciji o kotu zasuka rotorja posameznega SMTMP in je posledično ni možno uporabiti pri konfiguraciji, kjer ocenimo le povprečno kotno hitrost obeh SMTMP.

Vrednosti parametrov K_1 in K_2, ki nastopata v podanih enačbah pri posameznih načinih nastavljanja $i_{d,ref}$ in sta določeni eksperimentalno, lahko bralec najde v [5].

6 Zaključek

V delu je predstavljen koncept izgradnje simulacijskega modela zaporedno vezanih SMTMP, ki ju vodimo z enim FP. Rezultat opravljene stabilnostne analize sistema je razkril, da vsiljevanje dovolj velike pozitivne vrednosti toka i_d zagotovi stabilno obratovanje različno obremenjenih SMTMP, pri čemer njuno kotno hitrost neposredno merimo. Simulacijsko so bile preverjene različne izvedbe regulatorja $i_{d,ref}$. S stališča izgub v navitjih motorjev in dinamike najustreznejšo rešitev predstavlja skaliran tok $i_{q,ref}$ z dodano razliko hitrosti. V prihodnosti bi bilo potrebno preveriti možnosti za izvedbo pogona brez uporabe merilnikov rotorske pozicije in praktično verificirati predstavljene načine nastavljanja toka $i_{d,ref}$.

Literatura

[1] T. Liu in M. Fadel, "Comparative study of different predictive torque control strategies for mono-inverter dual-PMSM system," v *2016 18th Mediterranean Electrotechnical Conference (MELECON)*, str. 1–6, April 2016.

[2] A. D. Pizzo, D. Iannuzzi in I. Spina, "High performance control technique for unbalanced operations of single-VSI dual-PM brushles motor drives," v *2010 IEEE International Symposium on Industrial Electronics*, str. 1302–1307, July 2010.

[3] M. N. L. Nguyen, *Commande Predictive de deux Machines Synchrones alimentees en parallele par un Onduleir de Tension Triphase*. PhD dissertation, INP Toulouse, Genie Electrique, Electronique et Telecommunications (GEET), Laboratorire Laplace - UMR5213, 2013.

[4] D. Bidart, M. Pietrzak-David, P. Maussion in M. Fadel, "Mono inverter multi-parallel permanent magnet synchronous motor: structure and control strategy," *IET Electric Power Applications*, vol. 5, str. 288–294, March 2011.

[5] M. Čadež, "Optimizacija vodenja pogonskega sklopa ventilatorskega sistema," magistrsko delo, Fakulteta za elektrotehniko, 2018.

[6] Y. Lee in J. I. Ha, "Analysis and control of mono inverter dual parallel SPMSM drive system," v *2014 IEEE Energy Conversion Congress and Exposition (ECCE)*, str. 4843–4849, Sept 2014.

[7] D. Seto, F. Sun, J.F. Gieras and N. Hootsmans, "Single electronic drive controlling two synchronous motors via modified vector control," *Journal of Power Electronics*, 2003.

Ugotavljanje energijske učinkovitosti frekvenčno vodenega pogona vodne črpalke v industrijskem hladilnem sistemu

Henrik Lavrič, Klemen Drobnič, Rastko Fišer

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: henrik.lavric@fe.uni-lj.si, klemen.drobnic@fe.uni-lj.si, rastko.fiser@fe.uni-lj.si

Energy efficiency assesment of variable speed pump in industrial cooling system

Abstract. *Energy policy of the European Union strives for lower greenhouse gases emissions by lowering the overall energy consumption, by introducing more renewables and by improving the energy efficiency. In coherence with the latter item, energy audits must be conducted every four years in larger companies in order to comply with the directive 2012/27/ES. In this paper a comprehensive approach for determination of energy efficiency indicators for drives without an archive of electrical parameters is presented. Functional relationship between electrical and hydraulic parameters is established by means of measurements. Then an analytical procedure is developed to calculate electrical power and drive efficiency out of archived hydraulic parameters. It was found out that the analysed drives are heavily oversized and the overall efficiency is low, less than 60 %. Improvement in efficiency can be made in range of 15 % by using new properly sized pump and electric motor.*

1 Uvod

Odgovorna raba energije s ciljem ohranjanja energetskih zalog in zmanjšanja škodljivih emisij je poleg zmanjšanja stroškov za energijo zelo velik izziv novodobnega sveta. Izboljšanje energijske učinkovitosti je eden izmed najpomembnejših ciljev energetske politike in strategije v razvitih državah. Zmanjšanje energijske odvisnosti (delež uvožene energije v oskrbi z energijo) in podnebne spremembe predstavljata ključna izziva za vse članice Evropske unije (EU). Energijska odvisnost Slovenije je bila najvišja v letu 2008, ko je znašala 54,8 %, v letu 2017 pa je bila še vedno na visokih 47,5 % [1]. Slovenija naftne derivate in zemeljski plin uvaža v celoti, premog in električno energijo pa delno.

V splošnem obstajata dva dejavnika, ki lahko spodbudita k izboljšanju energijske učinkovitosti, in sicer tehnološke inovacije in prilagoditev deleža porabe energentov. V Sloveniji predstavlja industrija skoraj 25 % celotne porabe končne energije [2]. Po podatkih statističnega urada Republike Slovenije se je v obdobju od leta 2010 do leta 2013 končna poraba energije v industrijskem sektorju zmanjšala za 6 %, v obdobju med letoma 2000 in 2013 pa za več kot 18 %, trend porabe zadnjih let pa je zopet pozitiven [1]. Na žalost je bil upad v omenjenem obdobju precej bolj posledica recesije kot pa učinek smernic evropske direktive na področju energijske učinkovitosti, ki diktira zmanjšanje izpustov toplogrednih plinov za 20 %, povečanje deleža obnovljivih virov na 20 % ter zmanjšanje porabe energije za 20 % do leta 2020 glede na leto 2005. Električna energija v industriji je v letu 2016 predstavljala največji delež (43 %) porabe končne energije, zemeljski plin 34 %, ostala fosilna goriva 10 % (trdna goriva 3 %, naftni proizvodi 7 %), toplota 4 % ter obnovljivi viri in odpadki 9 % [1]. Zaradi omenjenih direktiv na področju energijske učinkovitosti je potrebno v večjih industrijskih obratih izvajati redne energetske preglede najmanj vsaka 4 leta od datuma zadnjega energetskega pregleda, kar določa direktiva 2012/27/ES.

Pri iskanju izboljšav za doseganje višjega izkoristka so bistvenega pomena električni pogonski sistemi (EPS), ki tvorijo povezavo med oskrbo z električno energijo in večino mehanskih in hidravličnih procesov. V industrijskih obratih EPS porabljajo kar dve tretjini električne energije [3], zato je njihovo učinkovito delovanje bistveno z vidika dvigovanja energijske učinkovitosti. Ocenjuje se, da EPS predstavljajo med 43 % in 46 % celotne svetovne porabe električne energije, kar povzroča približno 6.040 Mt emisij CO_2 [4]. Do leta 2030 bi se brez celovitih in učinkovitejših ukrepov politike energijske učinkovitosti vrednost emisij CO_2 lahko povečala za 40 % [4].

Waide in Brunner [4] ugotavljata, da bi ob uporabi motorjev z najboljšimi izkoristki prihranili od 4 % do 5 % celotne porabljene električne energije. Če bi k temu dodali primerne elektromehanske rešitve na bremenski strani, bi prihranili dodatnih 15 % do 25 %. Skupno bi izboljšali energijsko učinkovitost za 20 % do 30 %, kar bi zmanjšalo skupno svetovno povpraševanje po električni energiji za približno 10 %. Izkoristke lahko v industrijskih procesih povečamo z:

- uporabo ustrezno velikih in energijsko učinkovitih motorjev,
- uporabo frekvenčnih pretvornikov v pogonih, kjer je to smiselno,
- optimizacijo celotnega sistema, vključno z ustrezno velikimi cevovodi in vodniki,
- energijsko učinkovitimi stroji za pretvorbo mehanske energije (črpalke, ventilatorji, kompresorji, ...), ki zagotavljajo zahtevane pogoje z minimalnimi izgubami.

Brez poglobljene presoje omenjenih predlogov lahko številne ovire in motnje otežijo ali onemogočijo uresničitev prihrankov v trenutnem tržnem okolju.

V prispevku je predstavljen postopek ugotavljanja energijske učinkovitosti EPS na način, kot ga izvede podjetje, ki se zaveda potencialnih prihrankov in zato k izvajanju energetskih pregledov resno pristopi. Analiza je izvedena na podlagi arhiviranih podatkov procesnih

parametrov, med katerimi pa ni električnih moči EPS, zato tudi ni možno neposredno določiti porabe električne energije. V začetni fazi so bile izvedene meritve, s katerimi se je določila povezava med procesnimi parametri in električnimi parametri pogona. Nato je bil razvit matematični algoritem, s katerim se je obdelalo in določilo energijske kazalnike iz enoletnega podatkovnega arhiva. Na podlagi teh kazalnikov so izoblikovani predlogi za nadaljnje ukrepanje.

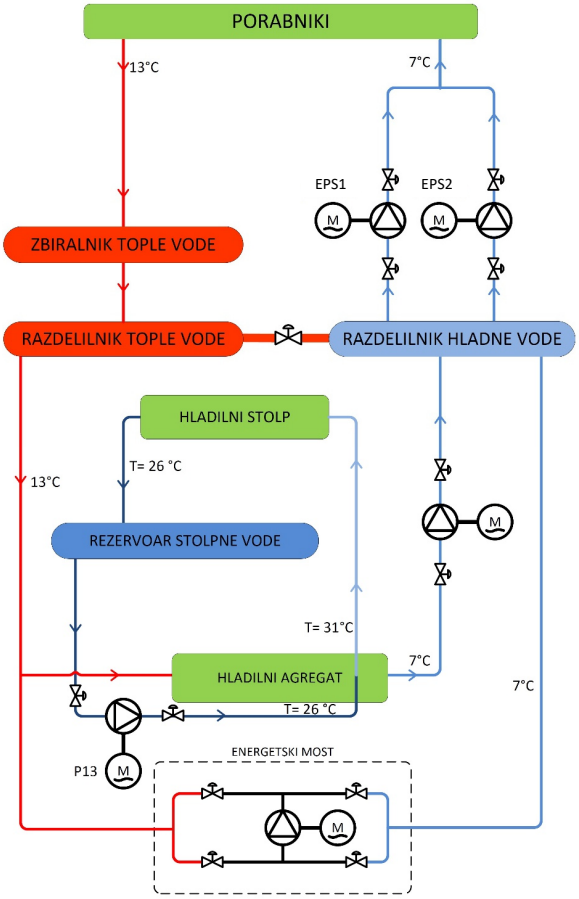
2 Tehnološki proces hlajenja vode

Izbrana pogona EPS1 in EPS2 sta v industrijskem obratu vgrajena v sistem hlajenja tehnološke vode in služita za transport hladilne vode do porabnikov [5]. Principialna shema centralnega hladilnega sistema tehnološke vode je na sliki 1. Pogona služita za črpanje hladne vode iz razdelilnika, v katerem je temperatura vode 7 °C, do porabnikov, kjer se uporablja za hlajenje različnih sistemov, naprav in prostorov. Takšni primeri so denimo klimati, ventilatorski konvektorji, hlajenje naprav, … Pri porabnikih se tehnološka voda segreje na približno 13 °C in v zaprtem ciklu jo je potrebno znova ohladiti na 7 °C. Povratna voda priteče do zbiralnika tople vode, nato imamo v sistemu dve možnosti ohlajevanja vode. Pri prvi s pomočjo hladilnega agregata ohladimo vodo na potrebnih 7 °C ter jo s pomočjo dodatne črpalke črpamo do razdelilnika hladne vode. To počnemo predvsem v času visokih zunanjih temperatur. Druga možnost pa je s pomočjo energetskega mostu, kjer se voda sama ohlaja in črpa preko dodatne črpalke. Energetski most predstavlja način ohlajevanja medija, ki se uporablja predvsem v zimskem času, kadar je potreba po hladni vodi precej manjša kot v poletnem času. Z uporabo energetskega mostu tako ne potrebujemo dodatnega ohlajevanja z agregatom, kar pomeni, da je poraba električne energije v zimskem času zaradi tega še dodatno zmanjšana. V poletnem času je za hlajenje medija uporaba hladilnega agregata nujna, kajti le z njim je možno ohladiti vodo na zahtevanih 7 °C.

Iz slike 1 je razvidno, da sta pogona EPS1 in EPS2 v sistem vezana vzporedno, s čimer je zagotovljena redundanca s pogoni oz. črpalkami. Ta je potrebna z vidika zanesljivosti. Vedno obratuje le ena črpalka, ker je zahtevane obratovalne parametre sistema sposobna zagotavljati samostojno. Če pride do okvare na katerem od členov v pogonu, je tako na voljo še en, popolnoma enak sestav, ki je sposoben zagotavljati primerno količino vode, ne glede na letni čas oz. tehnološke potrebe.

3 Merilna metoda

K ugotavljanju učinkovitosti EPS smo pristopili z uporabo direktne metode določanja izkoristka pogona. V tem primeru moramo izmeriti električne parametre na vhodu pogona in hidravlične parametre na izhodu pogona oz. v cevovodu, v katerem je vgrajena črpalka. Bistveno pri sami izvedbi meritve je bilo, da je ta potekala brez prekinjanja industrijskega proizvodnega procesa. To je bilo možno zaradi redundančne zasnove pogonov EPS1 in EPS2. V času nameščanja merilne opreme na enega izmed EPS je drugi normalno deloval in tako zagotavljal

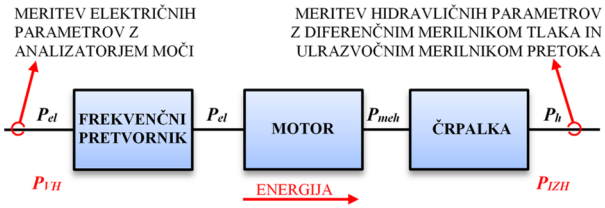


Slika 1. Principialna shema hladilnega sistema tehnološke vode

potrebe porabniških sistemov. Po zaključenem nameščanju opreme smo v dogovoru z nadzornikom preklopili med pogonoma, tako da je šel v obratovanje pogon z nameščeno merilno opremo. Meritve smo izvajali sredi junija, ko je potreba po hladilni vodi še zmerna. Želeli pa smo izmeriti parametre za obratovalna stanja z minimalno obremenitvijo, ki je prisotna pozimi in tudi pri maksimalni obremenitvi, ki je prisotna predvsem julija in eventualno še v avgustu, ko ni kolektivnega dopusta. Za izvajanje meritev tako nismo imeli na voljo pogojev, kot bi si jih želeli, oz. lahko ustvarili v laboratoriju, temveč smo morali kombinirati z realnimi pogoji, kot jih je narekoval proizvodni proces v okviru normalnega obratovalnega časa. Za izvajanje meritev pri parametrih, ki so bili različni kot trenutni, smo imeli na voljo do deset minutne časovne intervale, v katerih smo sistem podhlajevali ali pregrevali. Odvisno od tega, ali smo želeli meriti pri veliki obremenitvi ali nizki obremenitvi. Nato je bil pred ponovnim 10 minutnim intervalom za izvajanje meritev potreben daljši časovni interval, v katerem se je stanje v sistemu normaliziralo. Tako smo motenje proizvodnega procesa zaradi izvajanja naših meritev zelo omejili. Postopek smo ponavljali, dokler nismo posneli rezultatov v vseh predvidenih točkah.

Na sliki 2 je principialna blokovna shema pogona s prikazanimi merilnimi mesti električnih količin in hidravličnih količin za namen grafične ponazoritve merilnega principa. Električno moč, električni tok in ostale električne parametre smo na vhodni strani

frekvenčnega pretvornika merili z analizatorjem moči (PPA5530, Newtons4th), hidravlične parametre na izhodni strani črpalke pa s pomočjo diferenčnega merilnika tlaka (DP–4000, Klay Instruments) in ultrazvočnega merilnika pretoka (Fluxus ADM 6725, Flexim). Merjene vrednosti iz analizatorja so se shranjevale na prenosni računalnik preko namenskega programa, hidravlični parametri pa v podatkovni shranjevalnik (Almemo 2890-9). Ker moramo za pravilno določevanje izkoristka imeti na voljo istočasne odčitke vrednosti vhodnih in izhodnih količin, smo poskrbeli za sinhronizirano beleženje podatkov. V obeh sistemih smo nastavili čas vzorčenja na eno sekundo ter nato pri posamezni točki istočasno sprožali zajemanje podatkov v obeh sistemih in merili po eno minuto.



Slika 2. Principialna blokovna shema pogona z označenimi merilnimi mesti električnih in hidravličnih količin.

Meritve obratovalnih parametrov frekvenčno vodenega pogona smo izvajali tako, da smo pri določenem prednastavljenem dušenju v sistemu stopenjsko spreminjali hitrost vrtenja elektromotorja z nastavljanjem ustreznih referenčnih vrednosti frekvenčnemu pretvorniku preko nadzornega (SCADA) sistema. Merili smo od 10 Hz vse do 50 Hz s korakom po 10 Hz. Pri vsakokratnem prehodu z ene na drugo delovno točko smo počakali toliko časa, da so se vrednosti na merilniku pretoka umirile, kar je nakazovalo, da je prehodni pojav v sistemu izzvenel. Šele nato smo sprožili minutni interval merjenja. Različno dušenje in s tem različne potrebe po hladilni vodi v sistemu smo simulirali s pripiranjem ventila na tlačni strani črpalke. Vsak cevovod ima namreč pred in za črpalko vgrajene dušilne ventile (slika 1), ki jih lahko ročno zapiramo in odpiramo. V normalnem obratovanju so ventili vedno popolnoma odprti, za zagotavljanje različnih obratovalnih stanj v sistemu pa smo poleg meritev pri popolnoma odprtem ventilu merili še pri dveh delno odprtih položajih. Ko smo dosegli maksimalno vrtilno hitrost (50 Hz) pri popolnoma, to je 100 % odprtem ventilu, smo na tlačni strani ventil priprli na približno 60 % ter izvajali meritve v obratni smeri nastavljanja frekvence, od 50 Hz do 10 Hz. Potem smo ventil priprli na 30 % odprtosti ter še enkrat izvedli meritve od 10 Hz do 50 Hz.

4 Rezultati meritev

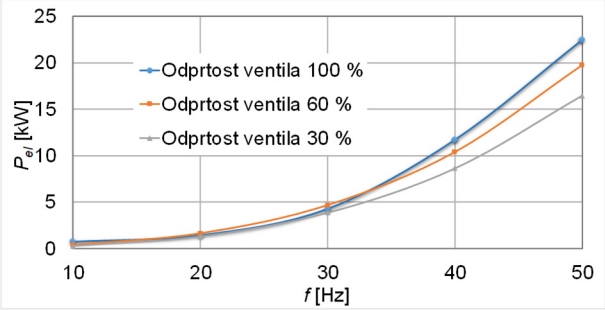
Ker sta pogona EPS1 in EPS2 identična in se izmenoma vključujeta v sistem na tedenski bazi, smo meritve izvedli samo na pogonu EPS2 in izmerjene karakteristike upoštevali kasneje tudi pri analizi obratovalnih stanj pogona EPS1. Oba pogona torej vsebujeta povsem enake gradnike z naslednjimi nazivnimi podatki:

Frekvenčni pretvornik:
U_n = 380 – 460 V
I_n = 61 A

Črpalka:
H = 20 m
Q = 270 m³/h
P_n = 30 kW
n = 1460 min^{-1}

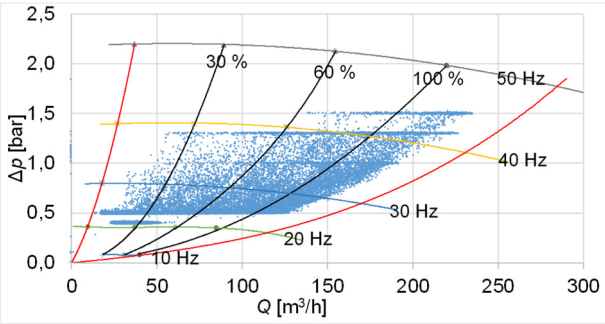
Motor:
P_n = 30 kW
U_n = 400 / 690 V
I_n = 56 / 31A
cosφ = 0,85
n = 1465 min^{-1}
Št. polov: 4
η = 91 % (IE1)

Na sliki 3 so poteki izmerjene električne moči na vhodu EPS2 v odvisnosti od frekvence napajalne napetosti motorja pri vseh treh položajih odprtosti ventila. Opazimo lahko križanje krivulj pri 60 % in 100 %, za kar nimamo eksaktne obrazložitve. Predvidevamo, da je pri popolnoma odprtem ventilu v času meritve na 20 Hz in 30 Hz točki prišlo do zmanjšanega odjema hladne vode in s tem do povečanega upora v sistemu, česar zaradi dolgotrajnega postopka izvajanj meritev ni možno izključiti. Deviacija je tako majhna, da je v rezultatih meritev na hidravlični strani niti ni opaziti.



Slika 3. Izmerjena delovna moč na vhodu EPS2 v odvisnosti od frekvence napajalne napetosti motorja pri treh odprtostih ventila.

Rezultati meritev tlaka in pretoka na hidravlični strani so prikazani na sliki 4. Krivulje, ki povezujejo točke meritev pri določeni nastavljeni frekvenci, potekajo dokaj vodoravno. Črne krivulje, ki potekajo v vertikalno-diagonalni smeri, pa povezujejo točke meritev, ki so bile izvedene pri posamezni odprtosti ventila. V graf so za potrebe analize dodani še rdeči krivulji in modre točke arhiviranih dejanskih obratovalnih stanj za obdobje enega leta.



Slika 4. Izmerjeni vrednosti količin na hidravlični strani z dodanimi modrimi točkami obratovalnih stanj v enem letu.

5 Ugotavljanje energijskih kazalnikov

Izkoristek celotnega EPS je odvisen od hidravlične moči na izhodu in vhodne električne moči P_{el}. Slednjo imamo

izmerjeno neposredno. Izhodno hidravlično moč P_h izračunamo iz merilnih rezultatov diference tlaka Δp in pretoka Q

$$P_h = \Delta p \cdot Q \,. \tag{1}$$

Izkoristek sistema η_{EPS} je potem določen z enačbo (2)

$$\eta_{EPS} = \frac{P_h}{P_{el}} \,. \tag{2}$$

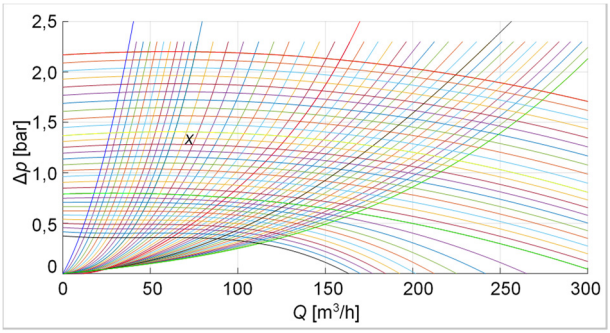
Kumulativno porabo električne energije dobimo z vsoto produktov trenutne moči in časovnega intervala Δt_i

$$W_{el} = \sum_{i=1}^{m} P_{el} \cdot \Delta t_i \,. \tag{3}$$

Težava je, da se električne moči posameznih pogonov ne merijo in ni arhiviranih podatkov, torej jo moramo določiti posredno.

Med obratovanjem industrijskega postroja se v desetminutnih intervalih shranjujejo podatki za tlačni padec v hladilnem sistemu in pretok vode, ter kateri pogon deluje. Hidravlični parametri obratovalnih stanj enega leta so vneseni v graf na sliki 4 (oblak modrih točk). Izkazalo se je, da vseh točk nimamo znotraj območja merjenih parametrov, zato sta z ekstrapolacijo dodani skrajni rdeči krivulji. Sedaj pa je treba matematično opisati lokacijo posamezne obratovalne točke in zanjo določiti potrebno električno moč. V ta namen smo vsak razdelek med krivuljami interpolirali s sto vmesnimi krivuljami, tako vertikalno kot horizontalno. Interpolacijska mreža na sliki 5 je zaradi preglednosti izrisana le z vsako deseto krivuljo. Na podoben način smo z ekstrapolacijo razširili območje moči v odvisnosti od frekvence in razdelke med krivuljami prav tako interpolirali s po sto krivuljami. Sedaj za določeno obratovalno točko *x* (slika 5) poiščemo najbližje presečišče krivulj in potem najprej določimo, pri kateri frekvenci mora pogon delovati ter nato še s kolikšno močjo. Ta postopek ponavljamo toliko časa, da za vsako izmed obratovalnih točk dobimo potrebno električno moč pogona. Sedaj imamo vse podatke, da poleg hidravlične moči (1) izračunamo tudi, s kakšnim izkoristkom deluje pogon v trenutni delovni točki (2), ter nadalje električno energijo (3) in povprečje moči ter izkoristka za obratovanje skozi celo leto.

Kazalniki učinkovitosti za obravnavana pogona so zbrani v tabeli 1. Poleg že omenjenih povprečnih vrednosti za celotno obratovanje v obdobju enega leta smo v celotnem naboru podatkov poiskali tudi maksimalne vrednosti moči in izkoristka. Pogon pri maksimalni moči ne obratuje z maksimalnim izkoristkom.



Slika 5. Interpolacijska mreža merjenih krivulj za aproksimacijo dejanskih obratovalnih točk.

Tabela 1. Kazalniki učinkovitosti za pogona EPS1 in EPS2

Kazalnik	EPS1	EPS2
η_{EPS} povprečje (%)	34,57	34,62
η_{EPS} maksimum (%)	59,42	60,30
P_{el} povprečje (kW)	3,60	3,38
P_{el} maksimum (kW)	18,97	18,83
P_h povprečje (kW)	1,53	1,42
P_h maksimum (kW)	9,81	9,72
W_{el} (MWh)	92,14	85,16

6 Sklep

V rezultatih analize najbolj izstopata dve dejstvi:

- maksimalni izkoristek pogona okoli 60 % je zelo slab,
- maksimalna električna moč malo pod 19 kW je veliko nižja od nazivne moči motorja 30 kW.

K tako slabemu izkoristku prispeva 2/3 črpalka in 1/3 motor. Z namestitvijo nove črpalke z visokim izkoristkom bi izboljšali izkoristek pogona od 10 % do 15 %. Z zamenjavo obstoječega motorja razreda učinkovitosti IE1 z učinkovitejšim, razreda IE3, bi skupni izkoristek dvignili le za 1,5 %. Lahko trdimo, da bi že sedaj v pogon lahko bil vgrajen motor z nazivno močjo 18,5 kW. Pri novemu pogonu z višjimi izkoristki pa bi imeli še rezervo moči.

Teoretični poskusi, da bi ovrednotili možne prihranke električne energije z upoštevanjem izkoristkov novih motorjev in črpalk, se izjalovijo. Problem je v tem, da tako za motorje, kakor tudi za črpalke ni na voljo funkcijske odvisnosti izkoristka od frekvence. Proizvajalci podajajo le poteke izkoristkov v odvisnosti od obremenitve za frekvenco omrežne napetosti. Da bi lahko izvedli eksaktni izračun, bi morali ponoviti celoten postopek analize, vključno z meritvami na sistemu, z uporabo novega pogona.

Literatura

[1] Statistični urad Republike Slovenije, Energetska bilanca in energetski kazalniki, http://pxweb.stat.si. (30.8.2018)

[2] M. Pusnik, F. Al-Mansour, B. Sucic, M. Cesen, "Trends and prospects of energy efficiency development in Slovenian industry", *Energy*, vol. 136, pp. 52 – 62, October 2017.

[3] ZVEI Automation, *Electric Motors and Variable Speed Drives, Standards and legal requirements for the energy efficiency of low-voltage tree-phase motors*, Frankfurt, December 2010.

[4] P. Waide, C. U. Brunner, "Energy-Efficiency Policy Opportunities for Electric Motor-Driven Systems", *International Eenergy Agency*, 2011.

[5] L. Koprivc, *Magistrsko delo: Analiza energijske učinkovitosti električnih pogonov v procesu hlajenja tehnološke vode*, FE, Ljubljana 2018.

Načrtovanje programske opreme za vodenje večfaznega pogona

Mitja Nemec, Andraž Rihar

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

Software design for the control of a multi-phase drive

Abstract. *Recent increase of current capability and output power of electric drives is a result of important technology-related improvements in the field of semiconductor power switches and utilization of multi-phase configurations. Multi-phase multi-system electric machines can be magnetically coupled at different levels, resulting in different mutual inductances. Control of magnetically coupled multi-system machines is not straightforward. The paper describes the effects of different control strategies on phase currents in multi-phase systems, presented with simulation results. Results show that phase-opposition control of magnetically-coupled machines is pointless. Furthermore, advanced control strategies, focused on using a combination of the main control algorithm, as well as additional difference-removal algorithms are presented. A graphical user interface was prepared for easier testing and interpretation of test results.*

1 Uvod

Razvoj električnih pogonov je neposredno povezan z razvojem polprevodniških močnostnih stikal. Tako z današnjo tehnologijo namreč hitro dosežemo tokove v območju 500 A – 1000 A, ki predstavljajo tako problem za močnostno elektroniko kot tudi za sam stroj. Tipična rešitev, ki se danes vedno bolj pogosto uveljavlja, je prehod s trifaznega pogona na večfazni pogon [1], [2]. Poleg povečanja moči ob uporabi enake tehnologije pa večfazni pogoni omogočajo tudi doseganje nižjih izgub [2], [3] in redundanco v primeru napak [4].

Vendar pa v večini primerov večfaznih pogonov ni dovolj samo razširitev na strojnem nivoju (pretvornik in stroj), bistvena sprememba je tudi v načinu regulacije.

Za potrebe razvoja in testiranje večfaznega pogona nujno potrebujemo programsko opremo s pripadajočim grafičnim uporabniškim vmesnikom, ki bi lahko bila splošno uporabna za testiranje različnih večfaznih strojev in za preverjanje vplivov različnih dejavnikov (principov vodenja, konstrukcijskih parametrov, karakteristik pretvornikov, itn.).

Članek predstavlja načrtovanje programske opreme za vodenje večfaznega pogona s poudarkom na uporabi kombinacije glavnih regulatorjev in regulatorjev razlike. Pristop je prikazan za primer 6–faznega dvosistemskega motorja.

2 Topologije

Če kot parameter delitve večfaznih strojev v obzir vzamemo tip magnetne sklopitve, torej magnetne povezave več navitij stroja, lahko električne stroje v grobem delimo na dve področji. Lahko so izdelani kot popolnoma neodvisni sistemi, torej med posameznimi sistemi oziroma istosmiselnimi fazami posameznih sistemov ni magnetne sklopitve (faktor sklopa $K = 0$) in jih označujemo kot n x 3 – fazni stroji (na primer 2 x 3–fazni stroj). Lahko pa med sistemi obstaja magnetna sklopitev, ki je načeloma med istosmiselnimi fazami različnih sistemov pozitivna ($0 < K < 1$). Faktor sklopa K je sicer poenostavljeno določen kot razmerje med medsebojno induktivnostjo M dveh navitij in korenom zmnožka lastnih induktivnosti L navitij. V našem primeru predvidevamo, da so lastne induktivnosti vseh navitij enake. V nadaljevanju so za lažjo interpretacijo dogajanja v večsistemskem stroju prikazani rezultati simulacije klasičnega H-mostiča, ki napaja dve bremenski dušilki z enakima lastnima induktivnostma (L_1 in L_2) in medsebojno induktivnostjo M, ki je povezana s faktorjem sklopa K (slika 1). Simulacije so bile izvedene v programskem okolju LTSpice pri stikalni frekvenci 20 kHz.

2.1 Brez sklopitve med sistemi

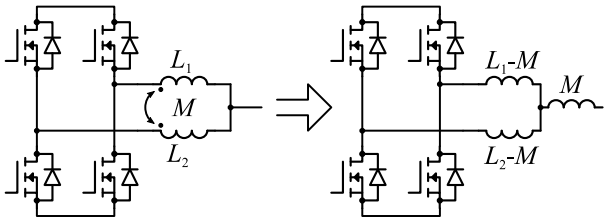
V primeru brez sklopitve sistemov ($K = 0$) lahko vsak sistem vodimo kot ločen stroj, saj veličine med seboj, razen prek položaja rotorja, niso povezane. Možna je tudi uporaba faznega zamika PWM signalov do 180°, ki je zelo pogosta na področju večvejnih DC/DC pretvornikov [5] in lahko v nekaterih konfiguracijah večfaznih strojev pomembno vpliva na zmanjšanje valovitosti toka v enosmerni vmesni tokokrog [6].

Rezultati simulacij kažejo, da je v primeru faktorja sklopa $K = 0$ tako ob uporabi sofaznega (fazni kot PWM je 0°, slika 2), kot tudi protifaznega proženja (fazni kot je 180°, slika 5), valovitost toka čez posamezno dušilko enaka, le v drugem primeru fazno zamaknjena. Z vidika celotnega bremenskega toka, torej skupnega toka obeh dušilk, je v drugem primeru valovitost enaka 0 A.

2.2 Sklopitev med sistemi

V primeru sklopitve sist%mov analiza zaradi medsebojne i.duktivnosti M, ki povezuje oba sistema oziroma v primeru predstavljenih simulacij obe bremenski dušilki, postane kompleksnejša.

Za lažjo interpretacijo dogajanja lahko lastnima induktivnostma posameznih bremenskih dušilk odštejemo medsebojno induktivnost in torej dobimo L_1-M oziroma L_2-M, na izhodno sponko pa dodamo dodatno dušilko z induktivnostjo M (slika 1) [7].

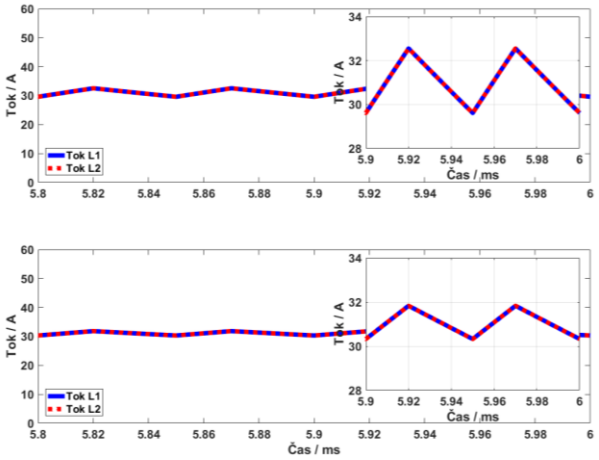


Slika 1. Shematski prikaz (levo) H-mostiča, ki napaja dve dušilki z lastnima induktivnostma L_1 in L_2 in medsebojno induktivnostjo M ter nadomestni shematski prikaz (desno) s tremi nadomestnimi dušilkami z vrednostmi L_1-M, L_2-M in M.

Stopnja sklopitve je seveda lahko različna in se spreminja od 0 do 1, pri čemer faktor sklopa K = 1 označuje uporabo bifilarnega navitja, pri čemer je vsak odcep napajan s svojo vejo. Običajno je seveda tudi v tem primeru še prisotne nekaj stresane induktivnosti posameznih navitij in zato raje predvidevamo faktor sklopa K = 0,95. Negativni faktorji sklopa so na področju električnih pogonov izredno redki, saj s posameznimi navitji želimo istosmiselno povečevati skupni navor motorjev. Negativni faktorji sklopa se pogosteje pojavljajo v DC/DC pretvornikih, kjer imajo določene prednosti [7].

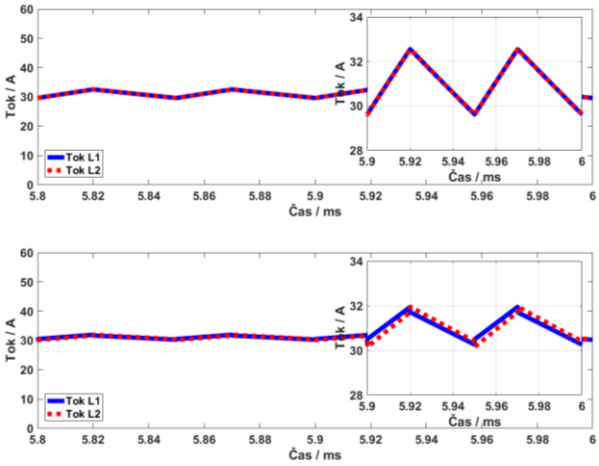
V primeru pozitivnega faktorja sklopa uporaba faznih zamikov signalov PWM pri proženju posameznih vej ni več povsem enostavna in ima različne vplive.

V primeru, da je sistem sklopljen pozitivno in je prožen popolnoma sofazno, se valovitost skozi posamezen sistem zmanjša na polovico, saj zaradi medsebojnega vpliva v tem primeru vsako navitje prevzame polovico valovitosti toka (slika 2).



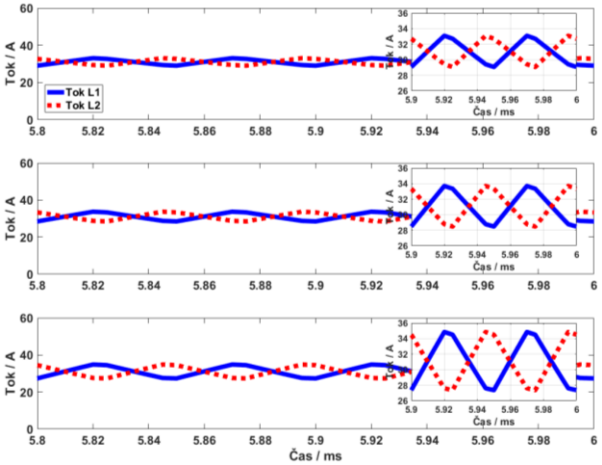
Slika 2. Potek tokov skozi dušilki L_1 (modra) in L_2 (rdeča) za sofazno proženje pri faktorju sklopitve K = 0 (zgornji del) in K = 0,95 (spodnji del)

Zaradi parazitnih efektov v sistemu (zakasnitve prožilnih vezij, zakasnitve MOSFET-ov, itn.) lahko pride do relativno majhne zakasnitve posameznih prožilnih PWM signalov. Majhna zakasnitev v primeru K = 0 praktično ne vpliva na potek tokov čez posamezni dušilki, ob K = 0,95 se posamična valovitost sicer ponovno razpolovi, pojavijo pa se manjše tokovne konice, ko sistema nista enako prožena (slika 3).



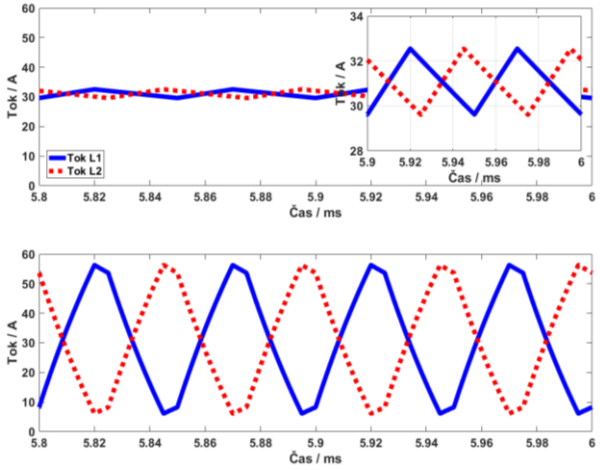
Slika 3. Potek tokov skozi dušilki L_1 (modra) in L_2 (rdeča) za skoraj sofazno proženje (zakasnitev 100 ns) pri faktorju sklopitve K = 0 (zgornji del) in K = 0,95 (spodnji del)

V primeru pozitivnega K in protifaznega proženja je valovitost na izhodu sicer enaka nič, saj sta toka ravno fazno premaknjena za 180°, povečevanje faktorja sklopa pa neposredno vpliva na povečevanje valovitosti toka skozi posamezni dušilki in je v določenih primerih (odvisno od aplikacije in različno sklopljenih strojev [8]) lahko smiselno (K = 0,33, 0,5 in 0,66 – slika 4).



Slika 4. Potek tokov skozi dušilki L_1 (modra) in L_2 (rdeča) za protifazno proženje (fazni zamik 180°) pri faktorjih sklopitve K = 0,33 (zgornji del), K = 0,5 (srednji del) in K = 0,66 (spodnji del)

V primeru bifilarnega navitja (K = 0,95) protifazno proženje ni smiselno (slika 5), saj se valovitost tokov poveča čez vse razumne meje. Tako je protifazno proženje smiselno samo pri strojih, pri katerih so posamezni sistemi razklopljeni oziroma zelo šibko sklopljeni. V vseh ostalih primerih je sofazno proženje edina smiselna izbira.



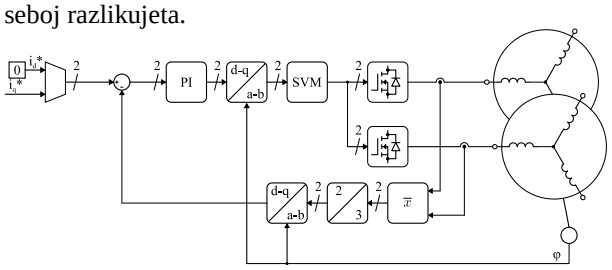
Slika 5. Potek tokov skozi dušilki L_1 (modra) in L_2 (rdeča) za protifazno proženje (fazni zamik 180°) pri faktorju sklopitve $K = 0$ (zgornji del) in $K = 0{,}95$ (spodnji del)

3 Tokovna regulacija pri večfaznih strojih

Tokovna regulacija električnih strojev temelji na dobro znani teoriji polja [9], kjer merjeni tok preslikamo v dvoosni koordinatni sistem in nato reguliramo prečno in vzdolžno komponento z dvema PI regulatorjema. Tak princip regulacije je uporaben tudi pri večfaznih strojih.

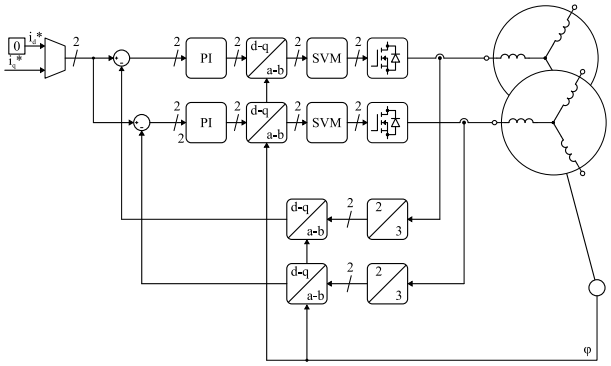
3.1 Klasična paralelna regulacija

V kolikor imamo opravka z 2 x 3–faznim strojem (šestfazni stroj, pri katerem sta dva sistema medsebojno popolnoma razklopljena), lahko tok reguliramo na dva načina. Najbolj preprost način je, da reguliramo povprečno vrednost toka obeh sistemov (slika 6) Prednost tega načina je predvsem v nezahtevnosti, saj ne potrebujemo dodanih regulatorjev, vendar pa lahko prihaja do asimetrije toka, v kolikor se sistema med seboj razlikujeta.



Slika 6. Regulacija povprečnega toka pri razklopljenem dvakrat trifaznem stroju.

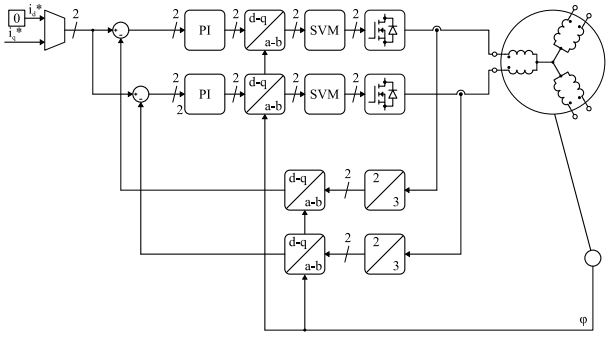
Nekoliko bolj zahtevna rešitev je regulacija toka vsakega posameznega sistema (slika 7). Z regulacijo toka v posameznemu sistemu dosežemo, da ne prihaja do asimetrije toka, vendar pa ta rešitev rabi nekoliko več računske moči.



Slika 7. Regulacija tokov v vsakem posameznem sistemu pri razklopljenem dvakrat trifaznem stroju.

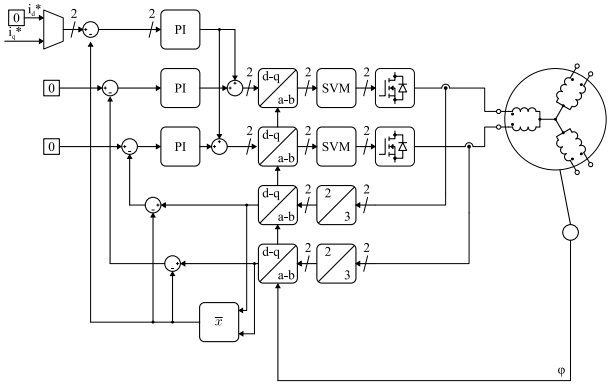
3.2 Glavni regulator ter regulatorji razlike

Pri strojih, kjer so sistemi medsebojno magnetno sklopljeni, pa ločena regulacija toka v posamezni veji ni dovolj. Zaradi medsebojne sklopitve se vsaka sprememba v enem sistemu obnaša kot motnja v sklopljenem sistemu. V kolikor se torej sistem odzove na motnjo, bo tako ponovno zmotil sklopljeni sistem. Tako klasična regulacija posameznih komponent deluje le v primeru, ko so spremembe toka dovolj majhne, le-to pa posledično pomeni slabo dinamiko (slika 8).



Slika 8. Regulacija tokov v vsakem posameznem sistemu pri sklopljenem šestfaznem stroju

Ta problem se lahko reši na več načinov, vendar se najlažja rešitev ponuja preko regulacije povprečnega toka (slika 9). V tem primeru glavno delo regulacije opravlja regulator povprečne vrednosti, s katerim dosežemo dovolj veliko dinamiko. K temu so dodani regulatorji, s katerimi odpravljamo asimetrijo v vsakem sistemu, ki pa imajo počasnejšo dinamiko.



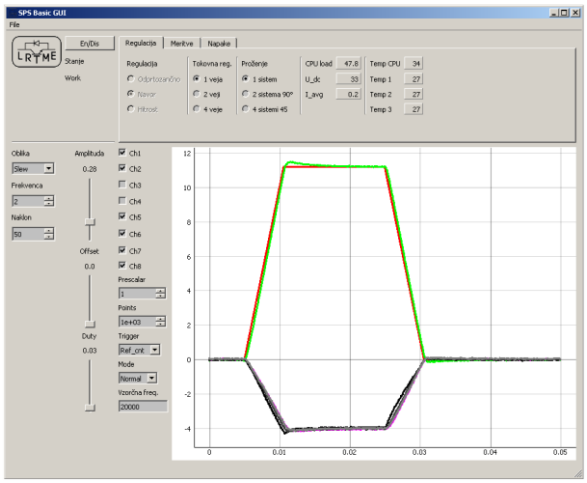
Slika 9. Regulacija povprečnega toka in posameznih tokov v vsakem sistemu pri sklopljenem šestfaznem stroju

4 Preostala programska oprema

V fazi preizkušanja večfaznega pogona je zaželeno imeti možnost spremljanja ne le vseh merjenih veličin (predvsem tokov), temveč tudi možnost izbire načina proženja (sofazno, protifazno) ter principa regulacije (regulacija povprečne vrednosti, regulacija posameznega sistema ali regulacija povprečne vrednosti ter asimetrije posameznega sistema). Tako je v samem mikrokrmilniku treba realizirati vse možne načine regulacije ter proženja.

4.1.1 Programska oprema in uporabniški vmesnik

Za nadzor izvajanja programa v mikrokrmilniku in izbiro načina delovanja je bil v programskem okolju Python pripravljen namenski grafični uporabniški vmesnik (slika 10). Ta omogoča izbiro številnih parametrov in načinov delovanja (zgoraj levo), spremljanje numeričnih vrednosti izmerjenih veličin (zgoraj desno), nastavljanje obratovalnih točk (levo) in grafični prikaz izmerjenih in izračunanih vrednosti za lažjo interpretacijo in sledenje postopku testiranja.



Slika 10. Namenski grafični uporabniški vmesnik z implementirano regulacijsko logiko za spremljanje in nastavljanje parametrov obratovalnih točk razsmernika.

5 Zaključek

Pri prehodu s trifaznega pogona na večfazni pogon je poleg same konstrukcije stroja in močnostne elektronike treba dovolj pozornosti nameniti tudi proženju in regulaciji toka. Pri tem velja upoštevati, da je najprimernejši način tako proženja kot tudi regulacije odvisen od same konstrukcije stroja. Pripravljena programska oprema, vključno z namensko izdelanim grafičnim uporabniškim vmesnikom, bo v sklopu nadaljnjih raziskav lahko veliko pripomogla k enostavnejši, kvalitetnejši in celovitejši analizi delovanja večfaznih večsistemskih pogonov z različnimi stopnjami magnetne sklopitve.

Literatura

[1] E. Levi, "Multiphase Electric Machines for Variable-Speed Applications," *IEEE Trans. Ind. Electron.*, vol. 55, no. 5, pp. 1893–1909, May 2008.

[2] E. Levi, R. Bojoi, F. Profumo, H. A. Toliyat, in S. Williamson, "Multiphase induction motor drives – a technology status review," *IET Electr. Power Appl.*, vol. 1, no. 4, pp. 489, 2007.

[3] J. Huang, M. Kang, J. q Yang, H. b Jiang, in D. Liu, "Multiphase machine theory and its applications," in *2008 International Conference on Electrical Machines and Systems*, 2008, pp. 1–7.

[4] M. J. Duran in F. Barrero, "Recent Advances in the Design, Modeling, and Control of Multiphase Machines - Part II," *IEEE Trans. Ind. Electron.*, vol. 63, no. 1, pp. 459–468, Jan. 2016.

[5] O. Garcia, P. Zumel, A. de Castro, in A. Cobos, "Automotive DC-DC bidirectional converter made with many interleaved buck stages," *IEEE Trans. Power Electron.*, vol. 21, no. 3, pp. 578–586, May 2006.

[6] S. Bhattacharya, D. Mascarella, G. Joos, J.-M. Cyr, in J. Xu, "A Dual Three-Level T-NPC Inverter for High-Power Traction Applications," *IEEE J. Emerg. Sel. Top. Power Electron.*, vol. 4, no. 2, pp. 668–678, Jun. 2016.

[7] A. Leban in D. Vončina, "Vpliv sklopljenih dušilk na lastnosti vzporedno vezanih pretvorniških modulov," presented at the Zbornik petnajste mednarodne Elektrotehniške in računalniške konference ERK 2006, 25. - 27. september 2006, Portorož, Slovenija, 2006, pp. 459-462.

[8] Y. Hu, Z. Q. Zhu, in M. Odavic, "Comparison of Two-Individual Current Control and Vector Space Decomposition Control for Dual Three-Phase PMSM," *IEEE Trans. Ind. Appl.*, vol. 53, no. 5, pp. 4483–4492, Sep. 2017.

[9] V. Ambrožič in P. Zajec, *Električni servo pogoni*. V Ljubljani: Slovensko združenje elektroenergetikov CIGRÉ-CIRED, 2016.

Vpliv mehanskih napetosti na magnetne lastnosti mehkomagnetnih materialov

Marko Petkovšek, Peter Zajec

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

Influence of a Mechanical Stress on Magnetic Properties of Soft-Magnetic Materials

Abstract. *This paper gives a deeper insight into magnetic properties of a soft-magnetic material that is exposed to a mechanical stress. Although soft-magnetic materials that are used as magnetic components in chokes, transformers or in rotating machines seem robust at first sight, this is far from the truth. Namely, as it is shown on a wound core example, either a short time or a continuous mechanical stress can significantly change the magnetic properties of the core. Therefore, in order to achieve the required properties of an assembled electromagnetic device, a proper handling is required throughout the whole process.*

1 Uvod

Končna kontrola je eno izmed ključnih področij v procesu proizvodnje mehkomagnetnih jeder iz transformatorske pločevine ali feritnega prahu. Poleg mehanske kontrole, kjer je treba preveriti dimenzijsko ustreznost in mehansko robustnost, je seveda za nadaljnjega uporabnika zanimiv predvsem podatek o magnetnih lastnostih. Te je v odvisnosti od vrste izdelka (npr. tračna/feritna jedra ali lamele statorskih in rotorskih paketov električnih rotacijskih strojev) mogoče dobiti z neposrednim ali posrednim merilnim postopkom, ki ju zaradi zagotavljanja sledljivosti in korektnega odnosa med poslovnima partnerjema izvajamo tako pri proizvajalcu kot pri kupcu.

Četudi so na prvi pogled izdelki iz tračnih mehkomagnetnih materialov za razliko od feritnih robustni in odporni na mehanske deformacije, temu v resnici ni tako. Prvotno odličen izdelek bo zaradi delovanja mehanske sile izkazoval bistveno spremenjene magnetne lastnosti [1, 2], kar se v prvi vrsti pokaže kot napačno ali neustrezno delovanje naprave, v katero je tak izdelek vgrajen, lahko pa je seveda tudi razlog za prekinitev pogodbe med poslovnima partnerjema. Mehanska sila, ki negativno vpliva na magnetne lastnosti materiala (gre za t.i. inverzno magnetostrikcijo oz. megnetoelastičnost – Villarijev pojav), je lahko posledica več dejavnikov. Najpogosteje je vzrok za prisotnost mehanskih napetosti v materialu neustrezen termični postopek po končanju vseh faz procesa, ki vključujejo mehansko preoblikovanje – npr. krivljenje pločevine, žaganje, brušenje, preoblikovanje v želeno obliko. Zato je poznavanje »metalurških« lastnosti materialov ključnega pomena za izbiro ustreznega temperaturnega profila (in ustrezne nadzorovane atmosfere), s čimer je mogoče mehanske napetosti izničiti in s tem doseči optimalne magnetne lastnosti materiala.

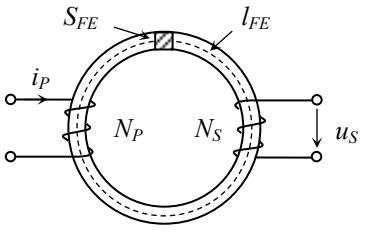
Poseben problem pri mehkomagnetnih materialih pa predstavlja neustrezno rokovanje po končani termični obdelavi in opravljenih kontrolnih meritvah magnetnih lastnosti, s katerimi smo sicer potrdili ustreznost izdelka. Magnetne lastnosti se namreč začasno ali celo trajno spremenijo, če je izdelek izpostavljen tako kratkotrajni (enkratni) mehanski obremenitvi – npr. zaradi padca z delovne mize ali neustreznega transporta, kot tudi trajni mehanski obremenitvi – npr. zaradi neustreznega pritrjevanja, neenakomernega ohlajanja zalivnega materiala transformatorskega jedra ali kot posledica namestitve navitij na jedro.

Namen tega prispevka je na primeru tračnih toroidnih jeder prikazati, kako in v kolikšni meri vplivajo mehanske obremenitve na nekatere ključne magnetne lastnosti materiala.

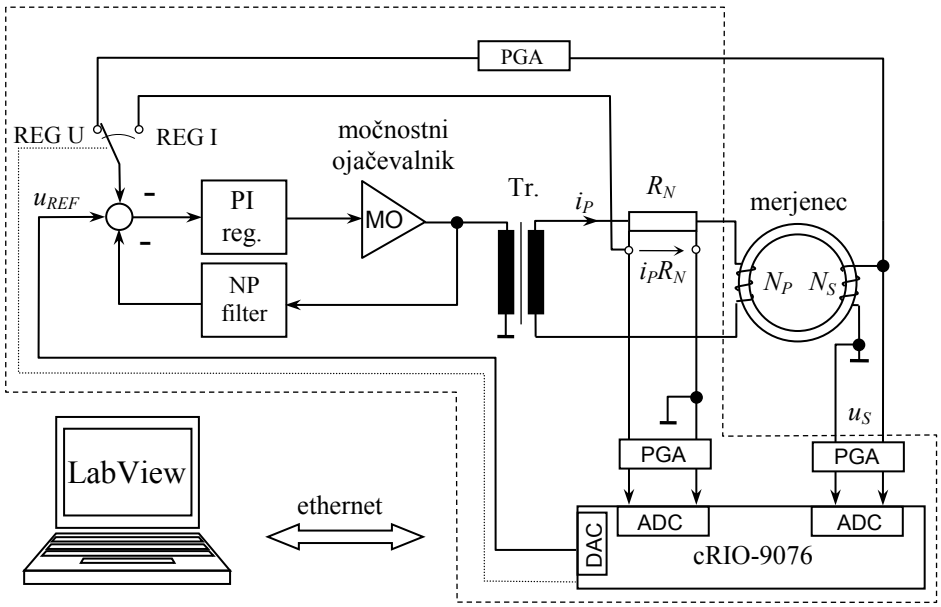
2 Osnovne značilnosti merilnega postopka

Magnetne lastnosti mehkomagnetnih tračnih jeder ugotavljamo s posrednim merilnim postopkom, pri katerem merjenec opremimo z dvema navitjema. S tokom i_P skozi primarno navitje z N_P ovoji (slika 1) poskrbimo za ustrezno magnetno poljsko jakost H, na sekundarnem navitju z N_S ovoji pa merimo inducirano napetost u_S kot posledico spreminjajoče se gostote magnetnega pretoka B v jedru.

Z željo po zagotavljanju primerljivosti meritev pri praktični izvedbi merilnega sistema ločimo dva pristopa [3], in sicer: a) vsiljeni primarni tok je sinusne oblike, posledično je inducirana napetost v splošnem nesinusna in b) primarni tok magneti jedro tako, da je inducirana napetost sinusne oblike. Na sliki 2 je podana blokovna shema merilnega sistema, ki je podrobneje opisan v [4, 5], in ki omogoča izvajanje meritev na oba predstavljena načina ter nam bo služil tudi za izvedbo meritev v nadaljevanju.



Slika 1. Toroidno jedro s primarnim in sekundarnim navitjem.



Slika 2: Shema merilnega sistema z možnostjo izbire načina vzbujanja.

3 Rezultati

V nadaljevanju so podani rezultati meritev magnetnih lastnosti treh mehkomagnetnih tračnih jeder s tipsko oznako 53/38/27 (zunanji in notranji premer ter višina jedra v mm) pri želeni gostoti magnetnega pretoka B_{ref} (1,6 T, 1,0 T in 0,2 T; sinusen potek) in pri $N_P = N_S = 2$. Poleg želene (referenčne) vrednosti gostote magnetnega pretoka si v stolpcih (tabela 1) sledijo oznaka vzorca, izmerjena vrednost gostote magnetnega pretoka B_m, efektivna vrednost magnetne poljske jakosti H, relativna permeabilnost, delovna (p_s) in navidezna (s_s) specifična izgubna moč.

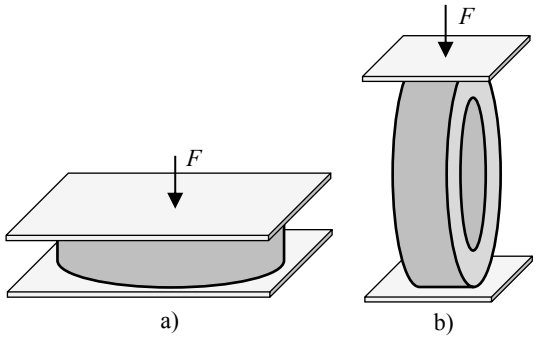
Pri poskusu so bila izbrana tri jedra s približno enakimi parametri, ki so bili pomerjeni/izračunani pred mehansko obremenitvijo (neosenčene vrstice). Do bistvenih sprememb v rezultatih pa pride, če jedro izpostavimo kratkotrajni mehanski obremenitvi, kar je razvidno iz osenčenih vrstic v tabeli 1 (krepko je napisan odstotek spremembe). Po mehanski obremenitvi, ki jo je predstavljal »naključni« padec z višine 1 m na vinilno talno oblogo, vrednosti pomerjenih parametrov pri istih vrednostih želene gostote B nakazujejo, da je zaradi padca pri obeh jedrih v njuni kristalni strukturi prišlo do deformacij, zaradi katerih bi predhodno sicer ustrezno jedro lahko za naročnika postalo neustrezno. Še posebej je to opazno pri vzorcu št. 2, medtem ko vzorec št. 1 ne izkazuje tako velikih odstopanj glede na meritev pred obremenitvijo. Glede na podane rezultate je vpliv kratkotrajnih mehanskih obremenitev na magnetne lastnosti jedra popolnoma nepredvidljiv. Dejstvo je, da je neustrezno rokovanje z jedrom po končani termični obdelavi zelo kritično, saj se lahko magnetne lastnosti jedra drastično spremenijo.

Druga možnost je trajno prisotna mehanska napetost v jedru kot posledica tehnoloških postopkov (zalivanje jedra v izolativno maso, namestitev navitij, pritrjevanje na podlago).

Tabela 1. Merilni rezultati za dva vzorca po kratkotrajni mehanski obremenitvi

B_{ref} [T]	vzorec	B_m [T]	H [A/m]	μ_r /	p_s [W/kg]	s_s [VA/kg]
1,6	1	1,601	21,53	41198	0,791	0,939
		1,600	23,34	34387	0,820	1,019
			8,4%	**-16,5%**	**3,7%**	**8,5%**
	2	1,600	22,29	37156	0,797	0,975
		1,599	27,35	26144	0,864	1,194
			22,7%	**-29,6%**	**8,4%**	**22,5%**
	3	1,600	21,83	39818	0,836	0,944
1,0	1	1,002	12,67	46690	0,307	0,346
		1,001	12,99	46266	0,315	0,355
			2,5%	**-0,9%**	**2,6%**	**2,6%**
	2	1,001	12,69	47208	0,305	0,347
		1,000	13,49	44552	0,326	0,368
			6,3%	**-5,6%**	**6,9%**	**6,1%**
	3	0,999	13,65	43813	0,340	0,372
0,2	1	0,198	3,61	31284	0,014	0,020
		0,200	3,72	30621	0,014	0,020
			3,0%	**-2,1%**	**0,0%**	**0,0%**
	2	0,200	3,62	31322	0,014	0,020
		0,200	3,88	29363	0,015	0,021
			7,2%	**-6,3%**	**7,1%**	**5,0%**
	3	0,200	3,68	30871	0,015	0,020

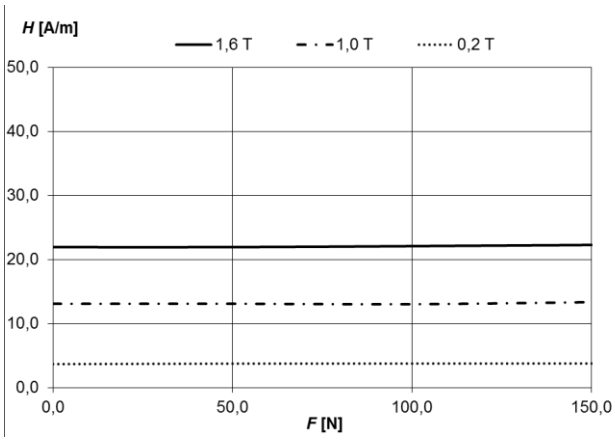
Vpliv trajne obremenitve je bil testiran s tretjim vzorcem, ki je bil vstavljen v namensko zgrajeni jarm, s katerim je bilo mogoče nastaviti obremenitev v razponu od 0 do 150 N, in sicer za primer a) v vzdolžni (osni) smeri in v primeru b) v prečni (radialni smeri), kot je razvidno s slike 3. V obeh primerih je na jedro delovala sila stiskanja zgolj v eni smeri, za pričakovati pa je bilo, da bo vpliv delovanja sile v vzdolžni smeri precej manjši kot v prečni.



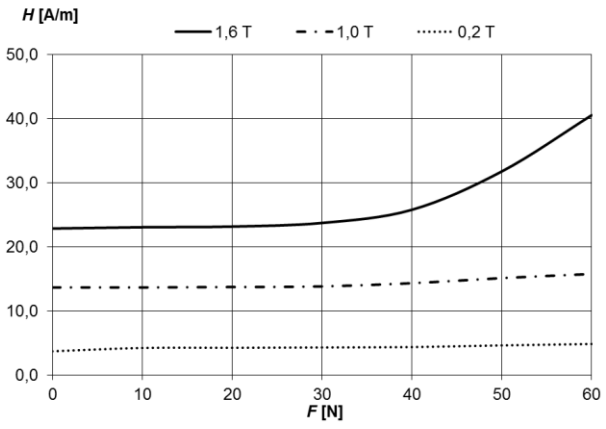
Slika 3. Princip obremenitve merjenca v vzdolžni (a) in prečni smeri (b).

Merilni rezultati magnetne poljske jakosti pri različnih gostotah magnetnega pretoka za vzorec št. 3 in vzdolžno obremenitev so podani na sliki 4, za prečno obremenitev pa na sliki 5.

Kot je bilo pričakovati, je vpliv mehanske obremenitve v vzdolžni smeri minimalen (pri sili 150 N je za enak B povečanje $H \approx 2\%$), opazen postane šele pri silah nad 150 N, ki pa jih z obstoječim merilnikom (LIBELA TENSO) ni bilo mogoče izmeriti. Povsem drugače je pri obremenitvi v prečni smeri. S slike 5 je razvidno, da je za doseganje želene gostote B v jedru v primeru delovanja prečne sile potrebna precej večja magnetna poljska jakost H (sploh pri 1,6 T).

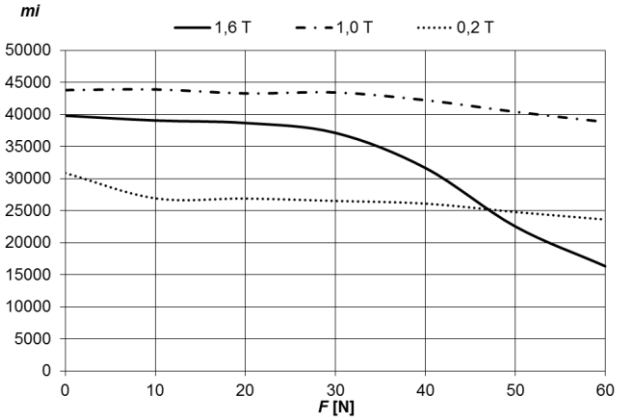


Slika 4. Magnetna poljska jakost pri različnih gostotah v odvisnosti od sile (vzdolžna smer).

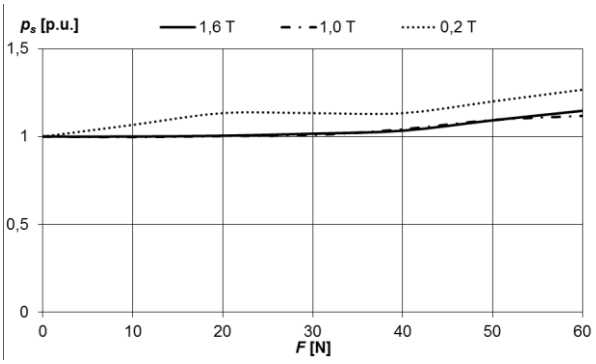


Slika 5. Magnetna poljska jakost pri različnih gostotah v odvisnosti od sile (prečna smer).

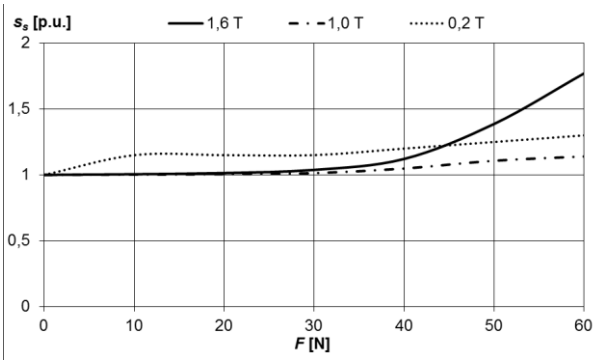
Podobne ugotovitve kot za magnetno poljsko jakost dobimo tudi za ostale izračunane parametre. Na sliki 6 je podan potek relativne permeabilnosti, na slikah 7 in 8 pa delovne in navidezne specifične izgubne moči (p.u.) pri različnih gostotah B v odvisnosti od sile v prečni smeri.



Slika 6. Relativna permeabilnost pri različnih gostotah v odvisnosti od sile (prečna smer).

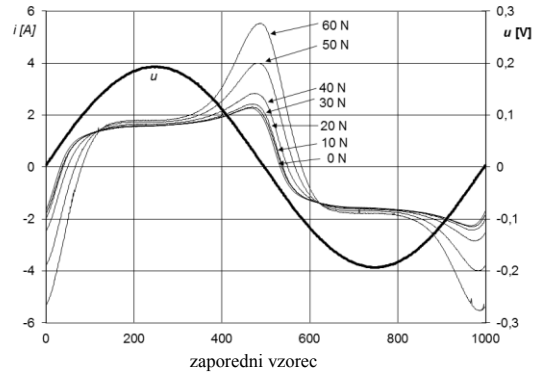


Slika 7. Delovna specifična izgubna moč pri različnih gostotah v odvisnosti od sile (prečna smer).

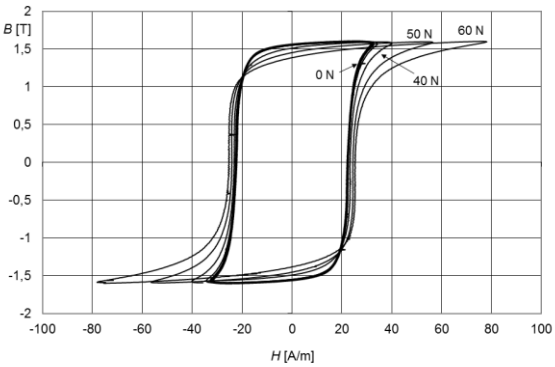


Slika 8. Navidezna specifična izgubna moč pri različnih gostotah v odvisnosti od sile (prečna smer).

Še boljši vpogled v razmere pri izpostavitvi jedra mehanski obremenitvi nam daje časovni potek ključnih dveh veličin med meritvijo – sekundarne inducirane napetosti in primarnega toka. Izhodišče za grafično predstavitev v obliki »oscilograma« napetosti in toka ter BH histerezne zanke sta tabeli s po 1000 vzorci toka in napetosti na periodo. Na sliki 9 je podan časovni potek sekundarne napetosti pri $B = 1{,}6$ T in primarnega toka za različne obremenitve v prečni smeri, na sliki 10 pa pripadajoča družina BH histereznih zank.

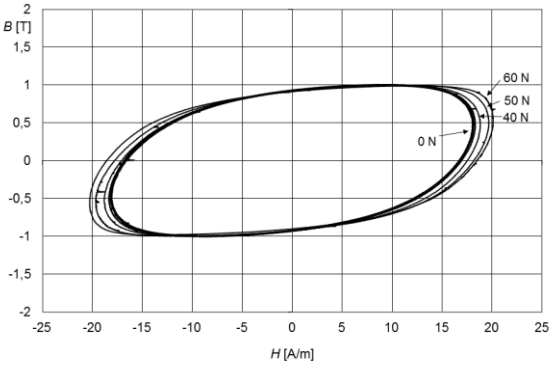


Slika 9: Časovni potek inducirane napetosti pri $B_{žel}$ = 1,6 T in časovni poteki toka pri različnih silah.

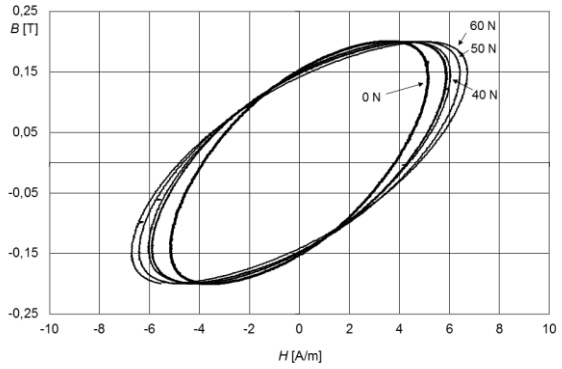


Slika 10: Družina histereznih zank pri $B_{žel}$ = 1,6 T in različnih silah.

Opazno je, da postaja z večanjem mehanske obremenitve histerezna zanka vedno bolj izrazita. Četudi je na prvi pogled (npr. glede na sliko 5) vpliv mehanske obremenitve na magnetne lastnosti jedra pri nižjih gostotah B minimalen, je že s slik 6 – 8, še bolj pa iz družine histereznih zank na slikah 11 in 12 razvidno, da mehanska sila tudi pri nižjih gostotah B nezanemarljivo vpliva na magnetne lastnosti jedra.



Slika 11: Družina histereznih zank pri $B_{žel}$ = 1,0 T in različnih silah.



Slika 12: Družina histereznih zank pri $B_{žel}$ = 0,2 T in različnih silah.

4 Zaključek

V prispevku je raziskan vpliv mehanske obremenitve na magnetne lastnosti tračnih jeder. Izkazalo se je, da se zaradi kratkotrajnega (»naključni« padec na tla) ali trajnega delovanja sile magnetne lastnosti jedra poslabšajo, saj je za doseganje enake vrednosti gostote magnetnega pretoka potreben večji magnetilni tok oz. večja magnetna poljska jakost. Velikost sile in njen vpliv na magnetne lastnosti jedra je sicer odvisen od togosti jedra, pri čemer je npr. pri jedru s tanjšo steno ($d_z \approx d_n$) vpliv pri enaki sili večji kot pri debelostenskem jedru. Na podlagi analize merilnih rezultatov za več tipov jeder (v prispevku so sicer podani rezultati samo za en tip jedra) je mogoče zaključiti, da je za doseganje želenih karakteristik elektromagnetne naprave treba posebno pozornost nameniti vsem fazam proizvodnega postopka, vključno s transportom in rokovanjem z že izdelano napravo.

Literatura

[1] J. Karthaus, S. Steenjes, N. Leuning and K. Hameyer: Effect of mechanical stress on different iron loss components up to high frequencies and magnetic flux densities, *COMPEL*, vol. 36, no. 3, 2017.

[2] K. Yoshikawa, Y. Yamada, S. Yamamoto, Y. Mochida, S. Ohashi, Y. Fukunaga, T. Sawai, K. Fujiwara and Y. Ishihara: Evaluation of Magnetic Characteristics of Soft Magnetic Powder Cores under Mechanical Stress, *SEI Technical Review*, no. 67, Oct. 2008.

[3] J. Sievert, The measurement of magnetic properties of electrical sheet steel – survey on methods and situation of standards, *J. of Magnetism and Magnetic Materials*, 2000, vol. 215-216.

[4] G. Modrijan, M. Petkovšek, P. Zajec, D. Vončina: Precision B-H analyser with low THD secondary induced voltage, *IEEE Tr. Ind. Electron*., Jan. 2008, str. 364-370.

[5] M. Petkovšek: Magnetne lastnosti tračnih jeder pri različnih načinih vzbujanja, *Elektrotehniški vestnik*, vol. 83, no. 4, 2016.

Merilna tehnika

Measurement

Kombinirane tehnike merjenja uporabniške izkušnje namiznih spletnih aplikacij s sledenjem kurzorja miške in pogleda uporabnika

Jure Trilar[1], Vid Čermelj[1], Tomaž Čegovnik[2], Emilija Stojmenova Duh[1]

[1]Laboratorij za telekomunikacije, Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, Ljubljana
[2] Laboratorij za informacijske tehnologije, Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, Ljubljana
E-pošta: jure.trilar@ltfe.org

Combined techniques for measuring the user experience of desktop web applications by tracking user's mouse cursor and eye gaze

Abstract. *The paper presents the state of the art and overview of popular approaches to measuring attention and user experience with a special focus on tracking the user's mouse cursor and eye gaze. Special focus is given on the comparison of both techniques in order to present a cost-effective option that would provide useful results along with the greatest possible simplification.*

1 Uvod

Skozi leta so postali spletni uporabniški vmesniki kompleksnejši, z več interaktivnimi elementi raznolikih oblik in formatov v različnih, bolj ali manj standardnih postavitvah, ki tekmujejo za pozornost uporabnika. V tem kontekstu je pozornost uporabnika omejena dobrina. Razumevanje, kako uporabniki pregledujejo strani in spletne aplikacije oz. uporabljajo digitalne uporabniške vmesnike, je pomembno za sklepanje o učinkovitosti in privlačnosti modernih informacijskih rešitev v različnih kontekstih.

Pripravili smo pregled področja in popularnih pristopov k merjenju pozornosti oz. uporabniške izkušnje s posebnim poudarkom na sledenju kurzorja miške in pogleda uporabnika. Osredotočili smo se na primerjavo teh dveh tehnik, z namenom predstaviti stroškovno učinkovito rešitev, ki bi na enostaven način dajala uporabne in kvalitetne rezultate, tudi pri različnih tipih uporabnikov.

2 Pristopi in tehnologije merjenja pozornosti in uporabniške izkušnje

Uporabnost (ang. usability) je s strani Mednarodnega združenja za standarde (International Standards Organisation - ISO 9241-11) [1] definirana, kot mera, ki pokaže, kako posamezen uporabnik dosega določene cilje z učinkovitostjo in zadovoljstvom [2]. Usability Professionals Association (UPA) se pri tem osredotoča na razvojni cikel rešitve, s pristopom, ki se ozira na neposreden uporabniški odziv v razvojnem procesu z namenom nižati stroške razvoja in izdelati končne rešitve, ki zadovoljujejo uporabnikove potrebe.

Mere uporabnosti se nanašajo na ocenjevanje ali merjenje določenih lastnosti produkta (uporabniškega vmesnika), ki jih je mogoče opazovati in kvantificirati [2]. Mere uporabnosti zajemajo različne vidike, kot so, naprimer, uspešnost reševanja nalog, zadovoljstvo uporabnikov in napake. Za določene vidike so izdelani standardizirani vprašalniki. Vprašalnik za merjenje uporabniške izkušnje (ang. *User Experience Questionnaire – UEQ)* se nanaša na 6 dimenzij **uporabniške izkušnje**, ki jih lahko združimo v pragmatične in hedonske lastnosti. **Kognitivno obremenjenost** merimo s pomočjo vprašalnika NASA-TLX. *System Usability Scale (ang. SUS)* ponuja hitro oceno **uporabnosti** vmesnika. *Usability Metric for User Experience* (UMUX) je skrajšan in poenostavljen SUS [3]. Našteti vprašalniki vključujejo uporabnikovo subjektivno dojemanje.

V primeru merjenja funkcionalne uporabnosti (uspešnost reševanja nalog) in pozornosti do elementov in sklopov v grafičnih uporabniških vmesnikih je bolj smiselna uporaba pristopa, ki ni popolnoma odvisen od uporabnikove subjektivne ocene. Takšno merjenje poskuša zajeti obnašanje, ki ga uporabnik ne more zlahka kontrolirati (npr. detektor laži), in ga brez namenske strojne ali programske opreme raziskovalec (opazovalec) ne more dobro opazovati [4]. Ti načini opazovanja delujejo bolj znanstveno in bolj zanesljivo od običajnih merjenj uporabnosti preko vprašalnikov. Potreben pa je pravilen pristop, da je mogoče podatke z dovolj veliko zanesljivostjo interpretirati v uporabne rezultate.

2.1 Izhodišča, zamejitve, predmet, in cilji merjenj

Temeljno izhodišče naše analize oz. članka je predstaviti stroškovno učinkovito rešitev za zanesljivo merjenje uporabniške izkušnje, ki bo razumljiva strokovnjakom v organizacijah, kjer do sedaj še niso implementirali kvantitativnih rešitev za merjenje uporabniške izkušnje, s poudarkom na čim boljšem aplikativnem potencialu za izboljšanje uporabniških informacijskih rešitev.

Določene tehnološke rešitve za merjenje pozornosti oz. usmerjenosti pogleda uporabnika omogočajo spremljanje skozi daljša časovna obdobja. Uporabljamo jih lahko v različnih okoljih, kjer ni nujno, da so prisotni digitalni zasloni ali drugi kanali za informiranje – torej v multimodalnih okoljih [5]. Primer tega je kabina avtomobila, kjer je potrebno zaznavanje več čutnih

kanalov in večopravilno odzivanje [6]. Za potrebe hitrega in učinkovitega merjenja uporabniškega vmesnika se bomo omejili le na eno modalnost, ki bo neposredno vplivala na iterativne izboljšave v skladu z **uporabniško usmerjenim razvojem** (ang. User Centered Design). Omejili se bomo na grafično prezentacijo na zaslonu naprave (praviloma osebnega računalnika), za katero je bil uporabniški vmesnik primarno načrtovan. V spletnem razvoju prepoznavamo različne načine prilagajanja razporeditve elementov grafičnega vmesnika, za potrebe hitrih aplikativnih rezultatov pa smo se pragmatično omejili na eno napravo. Upoštevali smo čas izpolnitve enega uporabniškega scenarija v vmesniku.

2.2 Sledenje kurzorja miške

Za analizo obnašanja uporabnikov v spletnem uporabniškem vmesniku je sledenje kurzorja miške zelo dostopna rešitev [7]. Implementacija tovrstne rešitve običajno ne predstavlja večjega časovnega in stroškovnega vložka. Prednost je, da lahko na daljavo merimo potovanje miškinega kurzorja pravih uporabnikov, ki praviloma ne občutijo tega, da se jih spremlja. V primeru, da se uporabniki zavedajo sodelovanja v eksperimentu, lahko spremenijo obnašanje (Hawthornov efekt) [8]. Chen je leta 2001 pokazal, da je skladanje premikanja miškinega kurzorja s sledenjem pogleda visoko, od 84% do 88% [9]. Zaradi evolucije spletnih uporabniških vmesnikov domnevamo, dat ti izsledki potrebujejo osvežitev.

Razlaga rezultatov je zaradi splošne dostopnosti (spletni servisi) večkrat prepuščena laičnim raziskovalcem, ki so obenem tudi razvijalci aplikacij.

Miškin klik je v večini scenarijev dober indikator uporabnikovega zanimanja, vendar pa nas zanimajo še dodatne značilnosti in signali, ki bi lahko izmerili uporabnost ter tudi napovedali obnašanje, frustracijo in zadovoljstvo uporabnika na spletni strani oz. aplikaciji.

2.3 Sledenje pogledu uporabnika

Sledenje pogledu uporabnika je najbolj natančna metoda za analizo pozornosti uporabnika v spletnih in drugih uporabniških vmesnikih, saj ponuja neposreden vpogled, ki z zelo malo napakami pokaže, kaj in koliko časa uporabnik dejansko opazuje, ter kakšna je pot njegovega pogleda [10]. Ta tehnika se uporablja v kognitivnih študijah, preučevanju interakcije človek-računalnik in marketinških raziskavah [7]. Ne uporablja se samo za merjenje uporabniške izkušnje, temveč tudi za dejansko navigacijo v grafičnih vmesnikih [11].

Ključna prednost je natančnost meritev. Obstajajo pa zadržki glede uporabe tehnologije za sledenje pogleda uporabnika za namene merjenja uporabniške izkušnje oz. pozornosti [12,13]. Bistvena ovira pri tem so višji stroškovni okviri, ki jih organizacije, ki nimajo specializiranih procesov za testiranje uporabniške izkušnje, ne morejo upravičiti. Ti okviri vključujejo nabavo opreme, ali pa najem zunanjih izvajalcev, ki lahko zagotovijo kvalitetno analizo merjenj. Druge zadrege vključujejo manjši vzorec testnih uporabnikov, učinek laboratorijskega okolja in pri nekaterih rešitvah tudi držo, ki ni tipična za normalno uporabo [8]. To se lahko pojavlja ob uporabi sledilnih očal ali pa ob zahtevi, da mora testni uporabnik čimbolj mirovati, kar lahko vpliva na rezultate.

2.4 Pristopi k merjenju in analizi

Kombinacija sledenja premikanja miške in pogleda bi morala podati zadosten uvid v zanimanje uporabnika za določene elemente vmesnika in sklepanje o njihovih uporabnosti, tudi tam kjer uporabniki med dejanskimi testiranji morda ne ubesedijo (metoda »think aloud«) neposredne uporabniške izkušnje [14].

Analiza merjenj sledenja pogleda, kot uveljavljena znanstvena metoda, mora biti podvržena enakim sistematičnim pristopom, kot ostale raziskave uporabniških vmesnikov [4]. V začetku je potrebno prepoznati raziskovalno vprašanje, kot temeljno izhodišče za priprave in merjenje ter interpretacijo meritev. Primeri raziskovalnih vprašanj: ali uporabnik prepozna namen strani samo s pogledom, na podlagi katerih elementov uporabniki izvedejo akcijo (klik), ali uporabniki z lahkoto prepoznajo ključen element, ali uporabniki lahko hitreje opravijo nalogo z različnimi grafičnimi zasnovami? Za uporabnike ustvarimo testne naloge, ki so povezane z raziskovalnimi vprašanji.

Merjenja opravimo na testnih pogledih (ang. *wireframe*), na interaktivnih prototipih, ali že na delujočem, produkcijskem uporabniškem vmesniku.

Po merjenjih smo pri čiščenju (ang. *data clean-up*) podatkov pozorni na primere, kjer podatki niso popolni ali pa preveč izstopajo, saj lahko ključno vplivajo na možnost pravilne razlage uporabnikovega obnašanja.

Pri analizi pogleda uporabnika, kot tudi sledenju miškinega kurzorja, tipično razberemo točke povečane pozornosti (ang. *fixations*). Umeščamo jih v širša področja interesa (ang. *areas of interest*), ki se med seboj ne prekrivajo.

Izložki analitičnih orodij so največkrat posnetki poti kurzorja ali pogleda (ang. *scanpaths*) posameznega uporabnika, toplotni zemljevidi pozornosti več uporabnikov (ang. *heatmap* ali *focus map*) ali drugi vizualizirani rezultati, ki ponujajo splošen uvid v pomembnost elementov v vmesniku z ozirom na zasnovo scenarija uporabnikove naloge oz. širšega procesa uporabe.

Rezultate interpretiramo na podlagi prej postavljenih raziskovalnih vprašanj, ki jih preoblikujemo v hipoteze, obenem pa moramo biti pozorni na tipične napake, povezane z uporabnikovo percepcijo, kot tudi tehničnim okvirom naprav za merjenje in implementacije testnega okolja, tako pri merjenju, kot med samo analizo in interpretacijo podatkov.

3 Različna orodja za spremljanje pogleda in kurzorja miške

Za spremljanje pogleda je na voljo veliko različnih naprav in programskih rešitev [15]. V skladu s prejšnjimi izhodišči želimo predstaviti rešitev, ki je cenovno dostopna in relativno preprosta za uporabo. Zato smo se osredotočili na naprave do 1000 €.
Za spremljanje premikanja kurzorja miške ne potrebujemo namenskih naprav, temveč se to izvaja preko programskih rešitev, ki so večinoma prosto dostopne.

3.1 Strojne rešitve

Strojne rešitve v tem cenovnem razredu so Tobii EyeX [16], Tobii 4C [17] in GazePoint GP3 SD 60 Hz [18]. Napravi podjetja Tobii nista primerni za raziskovalne namene, saj ne omogočata shranjevanja podatkov. Za uporabo teh rešitev v raziskovalne namene je potrebno pridobiti licenco Tobii pro, ki omogoča razčlenjanje in shranjevanje podatkov. Licenco je mogoče pridobiti preko zahtevka na uradni spletni strani podjetja Tobii.
Za raziskovalne institucije je primerna naprava podjetja Pupil Labs. Pupil headset so očala, ki nemoteno spremljajo pogled uporabnika, tudi, če se med uporabo aplikacije premika. Za komercialne uporabnike je naprava dražja in ne spada več v predviden cenovni razred. Programska oprema naprave je odprtokodna in omogoča prilagajanje glede na potrebe testiranja – kar zahteva določeno znanje programiranja.
Trenutno edina še podprta nizkocenovna naprava za spremljanje pogleda je GazePoint GP3. Njena glavna pomanjkljivost je, da podjetje ne nudi programske podpore in jo je potrebno poiskati oz. spisati svoje algoritme za razčlenjevanje podakov.
V ta cenovni okvir spada tudi naprava TheEyeTribe [19] (slika 1), ki je trenutno ni več mogoče kupiti, saj je podjetje TheEyeTribe prekinilo razvoj in prodajo te naprave. Zanjo je prav tako potrebno poiskati programsko opremo na spletu oz. napisati lastno rešitev za razčlenjanje podatkov.



Slika 1. Naprava za sledenje pogledu TheEyeTribe

Pogled lahko spremljamo tudi s pomočjo navadnih spletnih kamer. Ta način je najcenejši izmed vseh naštetih, a obenem tudi najmanj zanesljiv. Primeren je za raziskave, kjer nas ne zanima natančna lokacija pogleda uporabnika.

3.2 Programske rešitve

Večina dražjih orodij ima svojo programsko opremo s katero obdeluje podatke. V naši raziskavi smo se osredotočili predvsem na nizkocenovna orodja, ki je nimajo in jo je potrebno poiskati ali razviti. Pri orodjih, je bilo pomembno, da niso plačljiva in tako vsem dostopna.

Orodjem, ki podpirajo naprave, kot so TheEyeTribe in Gaze Point GP3 je skupno, da je bil njihov razvoj opuščen in trenutno ne prejemajo posodobitev. Orodja, ki ne omogočajo brskanja v novejših različicah spletnih brskalnikov in testiranja modernih aplikacij, so v večini primerov neuporabna.
Eden izmed najboljših odprtokodnih programov za spremljanje pogleda in kurzorja miške je OGAMA (OpenGazeAndMouseAnalyzer) [20]. Program je bil zadnjič posodobljen maja 2015, zato ni primeren za strani, ki vsebujejo moderne elemente in skripte z novejšim JavaScriptom (JS). Glavna prednost tega orodja pred drugimi je, da omogoča razčlenjevanje podatkov prek uporabniškega vmesnika. Program je prirejen posamezni napravi, ki jo povežemo z računalnikom. Prek te povezave se tekom merjenja pošiljamo podatki, ki jih OGAMA prikaže in shrani.
Orodje WebGazer.js [21] je JS knjižnica, ki jo lahko vključimo na spletno stran. To nam omogoča tudi oddaljeno testiranje in posledično ni učinka opazovalca. Za delovanje mora uporabnik dovoliti aplikaciji uporabo spletne kamere. Knjižnica za delovanje ne potrebuje strežnika, saj je celoten video obdelan v uporabnikovem brskalniku. Zaradi varnostnih razlogov se na strežnik pošiljajo le podatki o položaju pogleda. S podatki o zenici in očesih se oblikuje regresijski model, ki se med uporabo ves čas posodablja. Knjižnica za svoje izračune ne uporablja aktivnega prilagajanja položaja glave, ki je procesorsko zahtevno, zato lahko v realnem času deluje v spletnem brskalniku. To omogoča uporabniku prosto premikanje glave, model pa se sproti prilagaja glede na položaj zenic in oči ter koordinat zaslona.
V študiji, ki so jo izvedli razvijalci knjižnice WebGazer.js so s pomočjo najboljše knjižnice za spremljanje pogleda (clmtrackr) [22] in najboljšega regresijskega modela dosegli povprečno napako 104 slikovnih točk, kar je dovolj natančno za splošno testiranje uporabniške izkušnje. Za dosego take natančnosti knjižnica domneva, da je korelacija med pogledom in položajem kurzorja visoka, saj sledenje pogleda dopolnjuje s položajem miškinega kurzorja. Spremljanje pogleda z uporabo knjižnice WebGazer.js je smiselno v primerih, ko nas zanima le približna lokacija pogleda [23].

4 Testiranje izbranih rešitev

Pri testiranju smo se osredotočili na dve različni orodji, ki predstavljajo različne načine merjenja pogleda in kurzorja.
OGAMA je najbolj priročno orodje, saj omogoča rabo različnih naprav za merjenje pogleda. Z OGAMO lahko obdelujemo podatke s TheEyeTribe, Gazpoint GP3, Tobii, hkrati pa nam ponuja tudi možnost sledenja kurzorju miške. Program je preprost za namestitev. Zaradi velikega števila funkcionalnosti ga je potrebno dobro spoznati za optimalno uporabo. Ima vgrajen brskalnik, ki od 2015 ni bil posodobljen. Glavna omejitev pri uporabi orodja je bila, da z njim ni bilo mogoče testirali spletne aplikacije, ki je vključevala moderne elemente. Najlažje je testirati z visoko kakovostnimi prototipi uporabniških vmesnikov. Pri testiranju programa smo uporabili napravo

TheEyeTribe, ki smo jo prek USB-3.0 kabla priklopili v računalnik. Pred začetkom vsakega eksperimenta je napravo potrebno kalibrirati. Osnoven način kalibracije je z 9 točkami, če pa želimo večjo natančnost kalibracije lahko izbiramo še med 12 ali 16 točkami.
Drugo orodje, ki smo ga testirali, je knjižnica WebGazer.js, ki omogoča spremljanje pogleda prek spletne kamere. WebGazer lahko vključimo v kodo spletne strani. Za začetno nastavitev je potrebno imeti nekaj programerskega predznanja. Kalibracija je 9-točkovna in se izvaja v brskalniku. S kurzorjem miške se postavimo na posamezno točko na zaslonu in petkrat pritisnemo na levi miškin gumb. Isti postopek ponovimo na preostalih točkah. Ko je postopek kalibracije končan se nam kakovost le te izpiše na zaslonu. Med samo uporabo se kalibracija ves čas posodablja in izboljšuje.

5 Zaključek

Za iskanje osnovnih vzorcev uporabe uporabniških vmesnikov je merjenje z orodjema EyeTribe in WebGazer.js dovolj učinkovito in cenovno dostopno.
Naša ocena je, da je sledenje miškinemu kurzorju v kombinaciji s sledenjem pogleda s pomočjo spletnih kamer dovolj dober način za testiranje uporabniških vmesnikov informacijskih rešitev z namenom postopnih izboljšav v skladu z uporabniško usmerjenim razvojem.
Nizkocenovne naprave za spremljanje pogleda so primerne za točno določene primere uporabe in ne omogočajo nekaterih funkcij, ki jih imajo dražja orodja. Ta so bolj robustna in omogočajo boljšo kalibracijo ter kvalitetnejše podatke.
V prihodnosti bi radi preverili, ali pri spremljanju miškinega kurzorja in uporabnikovega pogleda v različnih starostnih (ali drugih demografskih in družbenih) skupinah, nastajajo značilne razlike. Pred tem moramo ugotoviti ali predstavljeni pristopi omogočajo dovolj kvalitetno testiranje ali bi za podrobne študije o pogledu uporabnikov še vedno potrebovali dražjo namensko opremo, ki je tudi programsko podprta.

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, 2016, Morgan Kaufmann, Amsterdam

[3] 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.

[4] A. Bojko: Eye Tracking the User Experience a Practical Guide to Research. Rosenfeld Media, 2013.

[5] T. Čegovnik, J. Sodnik: Free-hand human-machine interaction in vehicles, *6th International Conference on Information Society and Technology*, 2016

[6] T. Čegovnik, K. Stojmenova, G. Jakus, J. Sodnik: An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. Appl. Ergon. 2018, 68, 1–11.

[7] U. Sedlar, J. Bešter, A. Kos: Tracking Mouse Movements for Monitoring Users Interaction with Websites: Implementation and Applications, *Electrotehnical review 74*, str. 31-36, 2007, Ljubljana, Slovenija

[8] Eye tracking vs. mouse tracking analytics, https://www.clicktale.com/resources/blog/eye-tracking-vs-mouse-tracking-analytics/

[9] M. C. Chen, J. R. Anderson, M. H. Sohn: What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing, CHI '01 *Extended Abstracts on Human Factors in Computing Systems*, str. 281-282, 2001, Carnegie Mellon University, Pittsburgh, Pennsylvania, ZDA

[10] V. Navalpakkam., L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. *Proceedings of the 22nd International Conference on World Wide Web - WWW 13. 2013*. doi:10.1145/2488388.2488471

[11] S. Manson, L. Kne, K. R. Dyke, J. Shannon, S. Eria: Using Eye-tracking and Mouse Metrics to Test Usability of Web Mapping Navigation, Cartography and Geographic Information Science 39:1, 2013, str. 48-60, DOI: 10.1559/1523040639148

[12] Mouse Tracking vs. Eye Tracking, http://www.picnet.com.au/blogs/guido/2010/02/03/mouse-tracking-vs-eye-tracking/

[13] L. Cooke, Is the Mouse a "Poor Man's Eye Tracker"?, Usability and Information Design, 2006. str. 252-255

[14] A. Olsen, L. Smolentzov, T. Strandvall: Comparing different eye tracking cues when using theretrospective think aloud method in usability testing, 2010

[15] G. Funke, E. Greenlee, M. Carter, A. Dukes, R. Brown, and L. Menke, "Which Eye Tracker Is Right for Your Research? Performance Evaluation of Several Cost Variant Eye Trackers," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 60, no. 1, pp. 1240–1244, 2016.

[16] Specifications for EyeX, https://help.tobii.com/hc/en-us/articles/212818309-Specifications-for-EyeX

[17] Eye Tracker 4C, https://help.tobii.com/hc/en-us/categories/201596489-Eye-Tracker-4C

[18] GP3 Eye Tracker, https://www.gazept.com/product/gazepoint-gp3-eye-tracker/

[19] The Eye Tribe, http://theeyetribe.com/theeyetribe.com/about/index.html

[20] OGAMA – open gaze and mouse analyzer, http://www.ogama.net/

[21] WebGazer.js, https://webgazer.cs.brown.edu/

[22] clmtrackr: Javascript library for precise tracking of facial features via Constrained Local Models. https://github.com/auduno/clmtrackr

[23] A. Papoutsaki, "Scalable Webcam Eye Tracking by Learning from User Interactions," Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA 15, 2015.

Redefinicija mednarodnega sistema enot SI

Gregor Geršak, Janko Drnovšek

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

Redifinition of the SI system of units

In 2019, after some 200 years from its beginning the International system of units SI will be finally entirely defined on physical constants. Thus the SI will evolve from physical-artefacts-based system into a system of units based on fundamental physical constants and properties of the atoms. Definitions of the units have to be independent from technological progress and should not change with new scientific evidences and technological possibilities. New definitions should be such, that general public would not notice any change, that all measurements, performed using old definitions, stay valid within their uncertainties and that values of the kilogram, ampere, kelvin and mol do not change.

Key words: SI sistem enot, redefinicija, osnovne enote, kilogram, amper, kandela, sekunda

1 Uvod

1.1 Mednarodni sistem enot SI

Prvi korak v izgradnji in razvoju današnjega Mednarodnega sistema enot – SI (Systemè International d'Unités) sta bila decimalni metrični sistem, uveden v času Francoske revolucije, in posledična izdelava dveh platinastih etalonov, ki sta predstavljala meter in kilogram leta 1799 [1]. Leta 1832 je Gauss močno propagiral uporabo svojega fizikalnega koherentnega metričnega sistema enot, ki je poleg dolžine in mase vseboval tudi fizikalno veličino čas, ki je bila sicer v uporabi v astronomiji. Gauss je kot prvi izmeril absolutno vrednost zemeljskega magnetnega polja v obliki decimalnega sistema enot, ki je baziral na treh mehanskih enotah – milimetru, gramu in sekundi. Kasneje sta skupaj z Webrom z vključitvijo električne veličine razširila te meritve.

V 60. letih devetnajstega stoletja so potekale številne raziskave na področju elektrike in magnetizma pod vodstvom Maxwella in Thompsona. Skupaj sta sestavila potrebne lastnosti koherentnega sistema enot z osnovnimi in izpeljanimi enotami. Koherentni sistem je sistem, v katerem je mogoče vsako izpeljano enoto izraziti kot linearno kombinacijo osnovnih enot, pri čimer so vsi številski faktorji enaki 1.

Leta 1884 je bil predstavljen CGS sistem enot, koherentni sistem, ki je baziral na treh mehanskih enotah centimetru, gramu in sekundi in za decimalne večkratnike enot uporabljal predpone od mikro do mega.

Na področju elektrike in magnetizma so se velikosti enot CGS pokazale kot neprimerne, zato je bil v 80. letih devetnajstega stoletja predlagan koherentni sistem praktičnih enot. Ta je vseboval tudi enote ohm za električno upornost, amper za električni tok in volt za napetost.

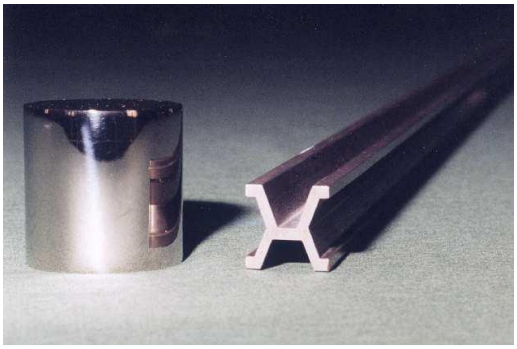
Po podpisu Metrske konvencije se je v okviru CIPM začela izdelava novih prototipov za dolžino in maso. 20. maja 1875 so se pooblaščeni predstavniki 18 držav sestali v Parizu na diplomatski konferenci o metru in podpisali Metrsko konvencijo. Države podpisnice so bile Francija, Avstrija, Nemčija, Belgija, Brazilija, Argentina, Danska, Španija, Združene države Amerike, Italija, Peru, Portugalska, Rusija, Švedska, Norveška, Švica, Otomansko cesarstvo in Venezuela.

Konvencija je definirala enoto dolžine meter in enoto mase kilogram v obliki pramer. Ustanovila je Mednarodni urad za mere in uteži (BIPM) kot stalni znanstveni zavod in Mednarodni komite za mere in uteži (CIPM). S podpisom konvencije o metru leta 1875 se je začelo obširno znanstveno in eksperimentalno delo pri izdelavi pramer metra in kilograma. Trajalo je polnih 14 let, vse do leta 1889. Rezultata sta bila prameter s karakterističnim presekom črke H in prakilogram (mednarodni prototip kilograma – IPK) v obliki valja (slika). Obe enoti sta skupaj z astronomsko sekundo tvorili sistem enot, podoben CGS sistemu, le da so bile osnovne enote meter, kilogram in sekunda. Definiran je bil MKS sistem.

Leta 1901 je Giorgi pokazal, da je mogoče kombinirati mehanske enote MKS sistema s praktičnimi električnimi enotami v obliko koherentnega sistema štirih osnovnih enot – treh mehanskih in ene električne.

Leta 1939 je bil predlagan štiri-dimenzionalni MKSA sistem z enotami meter, kilogram, sekunda in amper. V letu 1954 je CGPM sprejel amper, kelvin in kandelo kot osnovne enote električnega toka, termodinamične temperature in svetilnosti. Leta 1960 je bilo sprejeto tudi ime novega sistema enot Systemè International d'Unités – SI. Mol, kot osnovna enota za množino snovi, je bil v SI sistem dodan leta 1971.

Danes vsebuje SI sistem enot sedem osnovnih enot, in sicer kilogram, meter, sekundo, amper, kelvin, kandelo in mol (slika 2).

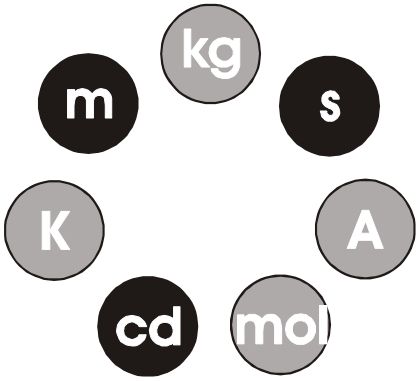


Slika 1. Prakilogram in prameter.

1.2 Definicije enot SI

Za definicije enot SI sistema skrbi Generalna konferenca za mere in uteži - CGPM (Comité International Généralé des Poids et Mesures), ki je organ Mednarodnega urada za mere in uteži - BIPM (Bureau International des Poids et Mesures) s sedežem v Parizu. Pri tem upošteva spodnje zahteve.

Mednarodni sistem enot mora biti dostopen in enak za mednarodno trgovino, visoko-tehnološko proizvodnjo, človekovo zdravje in varnost, zaščito okolja, globalne študije podnebja in temeljne znanost, ki je temelj vsega naštetega. SI enote morajo biti dolgoročno stabilne, interno konsistentne in jih je možno realizirati v praksi, temelječ na trenutnih teoretičnih opisih narave na najvišjem nivoju [2].



Slika 2. Sedem osnovnih enot SI sistema. Štiri enote bodo v 2019 na novo definirane (sivi krogi).

1.3 Zakaj so nove definicije enot SI potrebne?

V splošnem sta izraza enota in fizikalna veličina abstraktna pojma. Da bi se lahko enota uporabljala kot mera, mora biti na razpolago realizacija enot, to je fizikalni etalon. Etalon je lahko otipljiv predstavnik fizikalne veličine, na primer kilogram kot etalonska mera za maso [1]. Lahko pa je definiran tudi s standardiziranim postopkom merjenja in uporabo standardiziranih merilnih metod ter opreme. Takšen primer predstavlja merjenje električnega toka s pomočjo tokovne tehtnice. Tretja možnost je uporaba naravnih pojavov kot fizikalnih etalonov. Etaloni za dolžino, čas in električni potencial so na primer definirani na osnovi atomskih procesov.

Meriti pomeni s poskusom primerjati neznano veličino s primerno enoto, ki je definirana z dogovorom. Naloga nacionalnih metroloških institucij (NMI) je, da preskrbijo etalone, s katerimi je enota uvedena v meritev. Danes etalone, za katere je dogovorjena mednarodno sprejeta definicija (običajno abstrakten opis fizikalne veličine z navedbo vseh njenih bistvenih lastnosti), pridobimo s serijo treh operacij – realizacijo (izvedba mednarodno dogovorjene abstraktne definicije enote), vzdrževanjem (proces vzdrževanja rezultatov realizacije enote s primarnimi etaloni) in diseminacijo etalona (postopek podajanja enote s sledljivostno verigo od primarne realizacije do uporabnika) (slika 2).

Definicije enot SI sistema morajo biti neodvisne od tehnološkega napredka in se ne smejo spreminjati zaradi novih znanstvenih spoznanj.

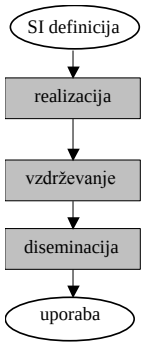
1.4 Nove definicije enot SI

Nove definicije enot so postavljene tako, da omogočajo sledeče:

- nobene opazne spremembe v vsakdanjem življenju ljudi,
- vse meritve, narejene s prejšnjimi definicijami, ostanejo veljavne znotraj svojih merilnih negotovosti,
- velikosti kilograma, ampera, kelvina in mola se ne spremenijo.

Pravzaprav bodo nove definicije vplivale na redkokateri del družbe, večinoma le na aktivnosti nacionalnih meroslovnih laboratorijev [3].

SI sistem je bil po 1960, ko je bil sprejet, dopolnjen nekajkrat. Prvič pa se bo zgodilo, da bodo na novo definiran kar štiri od sedmih osnovnih enot. Prvič po letu 1971, ko je bila med osnovne enote SI sistema vključena enota za množino snovi mol, bodo *na Mednarodni dan meroslovja 20. maja 2019 začele veljati nove definicije osnovnih enot.*



Slika 3. Osnovne operacije metrologije.

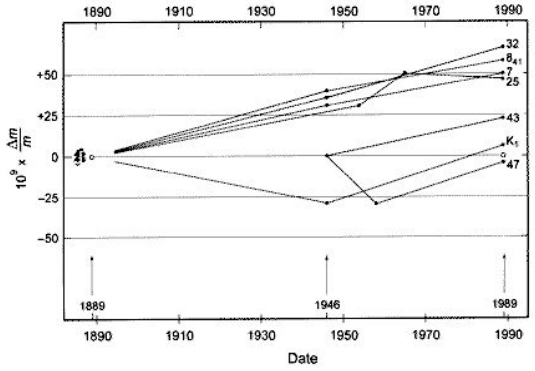
1.5 CODATA 2017

CODATA (The Committee on Data for Science and Technology) je mednarodna organizacija, ki bdi nad dostopnostjo in uporabnostjo znanstvenih podatkov. CODATA oz njeno delovno telo Fundamental Constants periodično objavlja mednarodna priporočila za vrednosti osnovnih konstant in faktorjev v fiziki in kemiji.

CGPM je povabila CODATA, da določi Planckovo konstanto *h*, osnovni naboj elektrona *e*, Boltzmannovo konstanto *k* in Avogardovo število N_A, in to vsako s tolikšnim številom decimalk, ki bi zagotovilo konsistentnost med trenutnim in revidirnaim sistemom SI [4]

1.6 Nova definicija kilograma

6 delovnih kopij mednarodnega prototipa kilograma (IPK) je bilo do sedaj le trikrat primerjano z IPK (leta 1889, 1946 in 1989). Pri tem je bilo ugotovljeno, da masa IPK v stotih letih povprečno raste za okoli 0.5 μg/leto [6].



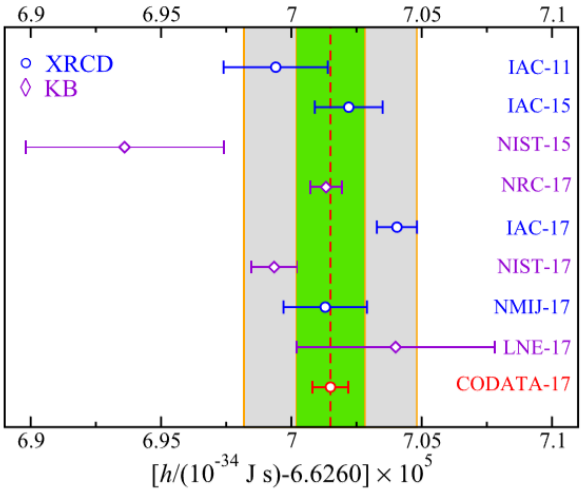
Slika 4. Šest delovnih kopij prakilograma je bilo do sedaj le trikrat primerjano z IPK (leta 1889, 1946 in 1989). Pri tem je bilo ugotovljeno, da masa IPK v stotih letih povprečno raste za okoli 0.5 μg/leto [6].

Da je nova enota vključena v sistem enot, potrebuje vrsto dokazov. Pri zamenjavi enote za maso so se v CGPM dogovorili, da so potrebni vsaj trije neodvisni poskusi z različnimi metodami, ki bodo dosegali relativno negotovost realizacija pod 5.10^{-8}, pri čemer bo vsaj en poskus dosegel negotovosti pod 2.10^{-8} [2]. Dva osnovna principa realizacije kilograma sta Kibblova tehtnica (pred 2016 imenovana tudi vatna tehtnica), ki vključuje povezavo elektromagnetnega in mehanskega vata, in Avogadrov projekt, ki s pomočjo silicijeve krogle in Avogadrovega števila določa maso. Definirane točnosti realizacij so dosegli do leta 2011, ko je skupina International Avogadro Coordination (IAC) dosegla negotovost Avogadrovega števila 3.10^{-8}, hkrati pa je ameriški inštitut NIST objavil svoj dosežek $3{,}6.10^{-8}$. Leta 2012 je EURAMET projekt omogočil navezavo projektov Kibblove tehtnice in silikonske krogle na vrednosti pod 2.10^{-8}.



Slika 5. Osnova Avogardovega projekta je krogla, ki jo izrežejo iz silicijevega monokristala (na sliki levo).

S pomočjo izsledkov teh poskusov bo opravljena kalibracija prakilograma in določitev njegove negotovosti. V praksi se ne bo nič spremenilo, iz definicije pa bo odstranjen fizični artefakt (prakilogram), kar bo omogočilo dolgoročno stabilnost definicije enote za maso (Tabela 1).



Slika 6. Določanje Planckove konstante v različnih študijah – zeleni pas predstavlja ±20 delov v 109 in sivi interval ±50 delov v 109. KB = Kibblova tehtnica (vatna tehtnica), XRCD: x-ray-crystal-density projekt (silikonska krogla). Z rdečo je označena končna CODATA 2017 vrednost Planckove konstante [4].

Tabela 1. CODATA 2017 prilagojene vrednosti fizikalnih konstant [4].

konstanta	vrednost	Relativna standardna negotovost
h	$6{,}626070150(69)\times10^{-34}$ Js	$1{,}0\times10^{-8}$
e	$1{,}6021766341(83)\times10^{-19}$ C	$5{,}2\times10^{-9}$
k	$1{,}38064903(51)\times10^{-23}$ J K^{-1}	$3{,}7\times10^{-7}$
N_A	$6{,}022140758(62)\times10^{23}$ mol^{-1}	$1{,}0\times10^{-8}$

Tabela 2. CODATA 2017 vrednosti fizikalnih konstant za revidirani SI sistem enot [4].

konstanta	vrednost	Relativna standardna negotovost
h	$6{,}62607015\times10^{-34}$ Js	0
e	$1{,}602176634\times10^{-19}$ C	0
k	$1{,}380649\times10^{-23}$ J K^{-1}	0
N_A	$6{,}02214076\times10^{23}$ mol^{-1}	0

Za zagotavljanje kontinuitete novega kilograma, so se odločili, da morajo različni novi postopki pripeljati do istih vrednosti prakilograma IPK.

$$Q = \frac{h}{m(IPK)} = \frac{m(^{28}Si)}{m(IPK)} \qquad (1)$$

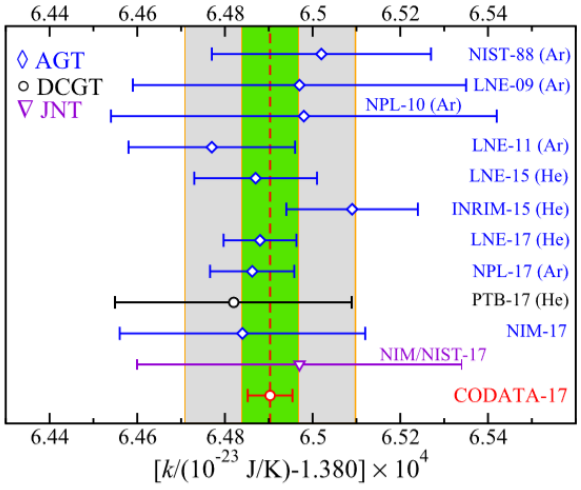
Veličina Q je razmerje, ki je odvisno od mase IPK in od konstante, ki jo dobimo z novim postopkom. V enačbi (1) tako nastopata Planckova konstanta h, ki je osnova Kibblovega poskusa, in Avogardovo število N_A oziroma masa silicijevih atomov $m(^{28}Si)$. Relativna negotovost, s katero je razmerje Q določeno, je u(Q)

Tabela 3. Razmere pred in po redefiniciji kilograma [5]

Tik pred redefinicijo	Takoj po redefiniciji
Masa prakilograma m(IPK) = 1 kg	Masa prakilograma m(IPK) = 1 kg
Relativna negotovost u(IPK) = 0	Relativna negotovost u(IPK) = u(Q)
Relativna negotovost Planckove konstante u(h) = u(Q)	Relativna negotovost Planckove konstante u(h) = 0

1.7 Nova definicija kelvina

V primeru redefinicije kelvina je bilo določeno, da sta potrebna vsaj dva fundamentalno različna pristopa k merjenju Boltzmannove konstante (npr. akustična plinska termometrija ali plinska termometrija z dielektrično konstanto, termometrija z Johnsonovim šumom, Dopplerjeva termometrija, itd), ki dasta enake rezultate in sta točnejša od 10^{-6}. V praksi se bo nova definicija poznala samo v merjenjih pod 20 K in nad 1300 K) [5].



Slika 7. Določanje Boltzmannove konstante v različnih študijah – zeleni pas predstavlja ±5 delov v 107 in sivi interval ±15 delov v 107. AGT = akustična plinska termometrija , DCGT = plinska termometrija z dielektrično konstanto, JNT = termometrija z Johnsonovim šumom. Z rdečo je označena končna CODATA 2017 vrednost Boltzmannove konstante [4].

2 Zaključek

Mednarodni sistem enot SI se počasi razvija iz sistema, ki je baziral na fizičnih artefaktih, v sistem, ki temelji na vrednostih osnovnih fizikalnih konstant in nespremenljivih lastnostih atomov. Stari SI sistem je po tretjem ovrednotenju nacionalnih kilogramov (slika 4) pokazal svoje omejitve, saj je masa definirana na podlagi artefakta. Zato je v zadnjih 30 letih znanost iskala možnosti za nadomeščanje mednarodnega prototipa kilograma (IPK).

Pred redefinicijo Si sistema enot so bile sledeče veličine eksaktno definirane: mednarodni prototip kilograma m(IPK) = 1 kg, magnetna permeabilnost vakuuma $\mu_0 = 4\pi.10^{-7}$ H/m, trojna točka vode T_{TPW} = 273,16 K in molska masa ogljika C 12 $M(^{12}C)$ = 0,012 kg/mol. V redefiniranem sistemu enot so te veličine določene eksperimentalno in so opremljene z negotovostmi. Pri tem je število decimalk za eksaktno numerično vrednost h, e, k in N_A določeno tako, da so numerične vrednosti ostanejo konsistentne s prejšnjimi eksaktnimi vrednostmi m(IPK), μ_0 in $M(^{12}C)$ znotraj negotovosti določene s CODATA 2017. V skladu s tem, je bilo število decimalk za k izbrano tako, da je T_{TPW} enaka 273,16 K znotraj standardne negotovosti, s katero T_{TPW} lahko v praksi tudi realiziramo.

Določitev h v novem SI sistemu bo nadgradila tudi električno metrologijo, saj bo električni tok, trenutno definiran kot I = e f, odslej definiran preko Ohmovega zakona kot razmerje napetosti Josephsonove napetosti in upornosti kvantnega Hall efekta.

Literatura

[1] J. Drnovšek, J. Bojkovski, G. Geršak, I. Pušnik, D. Hudoklin, Metrologija, Fakulteta za elektrotehniko, 2012

[2] Draft Resolution A – 26th meeting of the CGPM (13-16 November 2018), https://www.bipm.org/utils/en/pdf/CGPM/Draft-Resolution-A-EN.pdf

[3] Information for users on the proposed redefinition of the SI, https://www.bipm.org/cc/CCTF/Allowed/Revised_SI/CommonStatement_FinalVersion_June2017.pdf

[4] D B Newell et al, The CODATA 2017 values of h, e, k, and NA for the revision of the SI, Metrologia 55, L13–L16, 2018

[5] M. Kühne, Redefinition of the SI, ITS 9, Los Angeles, 2012

[6] B. Jeckelmann, W. Beer, Hat das Urkilogramm ausgedient?, ftp://www.metas.ch/pdf/Labors/research/wattwage.pdf

Termovizija plemenskih svinj

Tomaž Bogovič [1], Igor Pušnik [1], Marina Štukelj [2]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: igor.pusnik@fe.uni-lj.si
[2]*Veterinarska fakulteta, Univerza v Ljubljani, Gerbičeva ulica 60, 1000 Ljubljana*

Thermal imaging of breeding sows

Abstract.
MMA is a complex syndrome seen in sows shortly after farrowing (12 hours to three days). Mastitis is a bacterial infection of the udder. Metritis is an infection of the uterus. Agalactia is a reduction or total loss of milk production by the sow. MMA leads to increased piglet mortality and reduced weaning weights. Clinical signs include fever, anorexia and constipation. Treatment should be given as soon as MMA is diagnosed. In some pig farms, the MMA represents a very serious problem, therefore expensive veterinary treatment is required. A combination of appropriate criteria is essential to achieve a proper diagnosis and to minimize the use of antibiotics and prevent antibiotic resistance. The aim of experiment was to determine, if the use of thermal imaging compared to rectal temperature measurement could be considered as a reliable and effective preventive action for early detection of MMA and to provide an effective veterinary treatment on time. Unfortunately, thermal imaging compared to rectal temperature measurement could NOT be considered as a reliable tool for effective and preventive action for early detection of MMA, because due to large deviation of results temperatures of the mammary gland can't be uniquely linked to the occurrence of MMA.

1 Uvod

Sindrom mastitisa, metritisa in agalakcije, ki ga poznamo pod kratico MMA, je kompleksen sindrom, ki se pojavlja pri plemenskih svinjah v prvih treh dneh po prasitvi. Mastitis je bakterijska okužba vimena. Lahko je okuženih več vimenskih kompleksov ali pa samo en. Metritis je okužba maternice. Agalakcija je zmanjšanje ali popolna izguba proizvodnje mleka pri svinji, kar privede do lakote in pogina sesnih pujskov.

MMA je gospodarsko pomembna bolezen. Na nekaterih prašičjih farmah predstavlja sindrom MMA velik problem, saj je potrebno veterinarsko zdravljenje drago, poleg tega se pojavlja tudi pogin pujskov v višjem odstotku. Pravočasna diagnoza MMA je zelo pomembna, da se svinj ne zdravi preventivno zaradi povečane rezistence bakterij na antibiotike, da se izognemo izgubam in v dobrobit živali.

Namen raziskave je ugotoviti ali lahko uporabimo termovizijsko merjenje mlečne žleze za pravočasno diagnozo MMA in tako lahko pravočasno ter učinkovito zdravimo svinje še pred izgubo mleka. Poleg tega smo želeli ugotoviti povezavo temperature mlečne žleze z rektalno temperaturo in ugotoviti razliko uporabnosti med boljšo termovizijsko kamero Flir T650sc in slabšo termovizijsko kamero Fluke TiS45 ter izmeriti temperaturo sesnih pujskov na površini do starosti štirih dni. Uporaba termovizijske kamere za diagnozo MMA bi pomenila bolj enostavno meritev za rejca in manj stresno merjenje temperature za svinje.

2 Termografija svinj

Infrardeča termovizijska kamera brezkontaktno meri oddajano infrardeče sevanje merjenca in nato podatke uporabi za ustvarjanje termovizijskih slik oz. termogramov. Termograme lahko s posebno programsko opremo analiziramo, da dobimo podatke o temperaturi na želenih mestih. Termovizijske kamere se uporabljajo za različne praktične aplikacije v industriji, gradbeništvu, policiji, vojski, pri reševalnih operacijah, pri opazovanju človeka in živali, v živinoreji, biologiji in ekologiji, itd.

Pri živih organizmih povečanje ali zmanjšanje temperature na površini telesa povzroča žilni obtok. Na temperaturo kože vpliva tudi izolacija oz. debelina podkožne maščobe, ki izolira telo pred izgubo toplote. Z znanimi temperaturami na določenem predelu lahko ocenimo stanje tega predela. Pri uporabi te metode moramo upoštevati sončno svetlobo, vlago, umazanijo, vremenske razmere, oddaljenost meritve, emisivnost merjene površine.

Glavna prednost termografije je, da ta metoda ne zahteva neposrednega fizičnega stika s svinjo, kar je svinji bolj prijazno ter za rejca lažja in hitrejša meritev temperature od meritve rektalne temperature svinje [1].

3 Sindrom MMA

Prvi pokazatelj MMA je povišanje rektalne temperature svinje nad 39,3 °C od 12 do 18 ur po prasitvi. Prisotnost mastitisa je možno zanesljivo ugotoviti z analizo mleka.

Klinični znaki svinje s sindromom MMA so povišana rektalna temperatura nad 39,3 °C, otekla in boleča mlečna žleza, apatija, anoreksija ter zaprtje. Če svinje ne zdravimo pravočasno, imamo večji pogin in manjši prirast pujskov. Za preživetje pujskov je nujno čim hitrejše zdravljenje.

Prvi ukrep za učinkovito preprečevanje sindroma MMA je dobra higiena hleva in pranje ter razkuževanje svinj pred prasitvijo. Tudi pozna selitev svinje po 110 dnevih brejosti v porodni boks, brejost daljša od 116

dni, višja vlaga in višja temperatura vodijo do stresa svinje in pogostejšega pojava MMA [2].

4 Eksperiment

Meritve smo izvajali od novembra 2017 do februarja 2018 na Zavodu za prestajanje kazni zapora Maribor na odprtem oddelku Rogoza, kjer se ukvarjajo s prašičerejo. Merili smo temperature mlečnih žlez svinj, rektalno temperaturo svinj, temperaturo površine sesnih pujskov, temperaturo zraka, relativno zračno vlažnost in primerjali rezultate dveh termovizijskih kamer različnega cenovnega razreda.

Za merjenje rektalne temperature svinje smo uporabili umerjeni veterinarski alkoholni termometer z razširjeno merilno negotovostjo 0,2 °C. Za merjenje temperature mlečne žleze smo uporabljali dve termovizijski kameri različnega cenovnega razreda. Boljša termovizijska kamera Flir T650sc ima ločljivost 640x480 slikovnih točk, valovno dolžino 7,5 µm do 13,0 µm, 45° lečo in avtomatsko ostrenjem. Slabša termovizijska kamera Fluke TiS45 ima ločljivost 160x120 slikovnih točk, valovno dolžino 7,5 µm do 14,0 µm in ročno ostrenje. Obe termovizijski kameri sta bili umerjeni s posebnim črnim telesom (velika odprtina premera 26 cm) v območju od 10 °C do 70 °C. Razširjena merilna negotovost termovizijske kamere Flir T650sc je 0,4 °C s korekcijsko -1,0 °C. Razširjena merilna negotovost termovizijske kamere Fluke TiS45 je 2,0 °C s korekcijsko -2,2 °C. Za analizo termogramov kamere Flir T650sc smo uporabljali program Flir researchir max 4, za kamero Fluke TiS45 pa smo uporabljali program SmartView 4.3.17.0. Za merjenje temperature zraka smo uporabili umerjene datalogerje Ebro EBI 20 T in za merjenje relativne vlažnosti zraka umerjeni dataloger Ebro EbI 20-TH. Vsi datalogerji so beležili meritve na 5 minut. Za obdelavo podatkov in izris grafov smo uporabili Microsoft Excel.

Meritve smo izvajali na 21-ih brejih plemenskih svinjah v treh skupinah. Skupine so vključevale po 5, 6 in 10 plemenskih svinj. Svinje so bile stare od 10 mesecev do skoraj 5 let. Za lažjo orientacijo in obdelavo podatkov smo plemenske svinje preimenovali oz. oštevilčili po zaporednih številkah od 1 do 21.

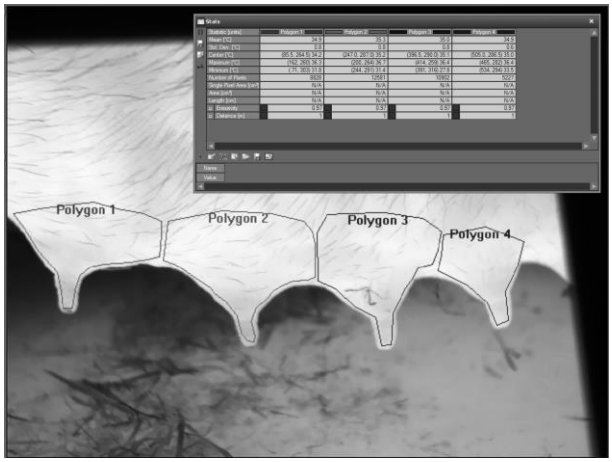
Vsako plemensko svinjo smo začeli spremljati oz. meriti vsaj 5 dni pred rokom prasitve ter merili rektalno temperaturo in temperaturo njenih mlečnih žlez vsaj še 5 dni po prasitvi. Zaradi sočasnega izvajanja meritev na večjem številu svinj v eni skupini, ki so prasile različne dni, smo vsako svinjo tako merili vsaj 15 dni. Meritve smo izvajali na približno 24 ur. Meritve temperature mlečne žleze smo izvajali na razdalji od 0,8 m do 1 m. V analizi smo upoštevali relativno zračno vlažnost zraka, temperaturo zraka, vedno enake pozicijske pogoje za merjenje in emisivnost mlečnih žlez 0,97.

Pred začetkom izvajanja meritev smo opravili pripravo obeh termovizijskih kamer ter zapisali temperaturo zraka in relativno vlažnost zraka v porodnišnici. Kameri smo za opravljanje meritev pripravili tako, da smo jih prižgali in pustili v prostoru, kjer smo izvajali meritve, vsaj 30 min.

Prvo smo vsaki svinji izmerili rektalno temperaturo. Rektalno temperaturo smo merili vsaj eno minuto v rektumu od 3 cm do 5 cm globoko. V primeru premikanja svinje smo meritev ponavljali, dokler nismo izvedli korektne meritve.

Nato smo začeli izvajati meritve temperaturne mlečnih žlez. Kot prvo smo vedno uporabljali termovizijsko kamero Flir T650sc. Uporabljali smo ročno ostrenje, da smo imeli vedno ostro sliko na mlečni žlezi. Zaradi širokega kota leče kamere smo meritve s kamero Flir izvajali na razdalji 0,8 m od svinje in pod približno pravim kotom na svinjo. Ko smo zaključili meritve na svinji s prvo kamero, smo meritve temperatur mlečnih žlez ponovili še z drugo kamero Fluke TiS45. Zaradi ožjega kota leče objektiva smo meritve s kamero Fluke TiS45 izvajali na razdalji 1 m.

Vsako mlečno žlezo smo povezali z lokacijo mlečne žleze na pripadajočem termogramu. V ustreznem programu smo določili temperaturo zraka, relativno vlažnost zraka v prostoru in emisivnost mlečne žleze. V programu smo vsako mlečno žlezo obrisali oz. označili na termogramu, kot prikazuje Slika 1. V izogib morebitnim napakam meritev na robnih točkah smo označili mlečno žlezo nekaj slikovnih oz. merilnih točk od roba. Iz izbranega območja termovizijske slike smo razbrali maksimalno temperaturo posamezne mlečne žleze.



Slika 1: Analiza temperature prvih štirih levih mlečnih žlez v programu Flir researchir max 4

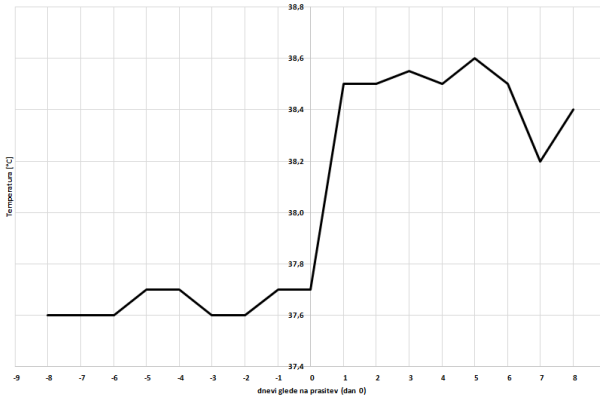
Ker imata velik vpliv na zdravje in počutje živali temperatura prostora in vlažnost zraka, smo merili temperaturo v oddelkih čakališča, porodnišnic in prostoru nad talnim grelcem sesnih pujskov. Relativno zračno vlažnost in temperaturo zraka smo merili približno na sredini prostora, približno 5 m od vseh sten in na višini 1,5 m. Vsi merilniki so bili na tem mestu ves čas in so beležili podatke vsakih 5 minut.

Merili smo tudi temperaturo nad talnim grelcem za sesne pujske. Termometer je bil nameščen ob montažni steni na višini približno 30 cm od talnega grelca in 10 cm pod pokrovom.

S termovizijsko kamero Flir T650sc smo merili tudi temperaturo na površini kože sesnih pujskov v starosti od prasitve do štirih dni.

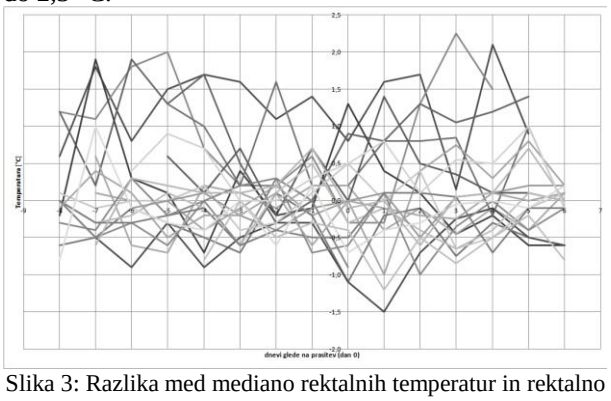
5 Rezultati

Slika 2 prikazuje mediano vseh meritev rektalnih temperatur glede na dan prasitve v obdobju 17-ih dni, 8 dni pred in 8 dni po prasitvi. Iz te slike lahko zaključimo, da se pred prasitvijo rektalna temperatura ne spreminja. V dnevih po prasitvi se rektalna temperatura svinj poviša za 0,8 °C glede na mediano meritev, ki je 38,5 °C.



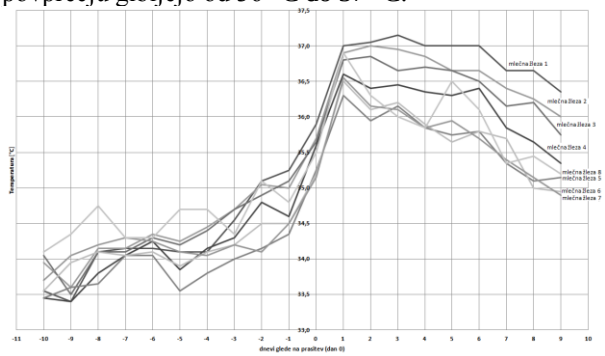
Slika 2: Mediana rektalnih temperatur plemenskih svinj glede na dan prasitve

Razliko med mediano rektalnih temperatur in rektalno temperaturo plemenskih svinj glede na dan prasitve vidimo na Sliki 3. Opazimo odstopanje rektalne temperature od mediane rektalnih temperatur od -1,5 °C do 2,3 °C.



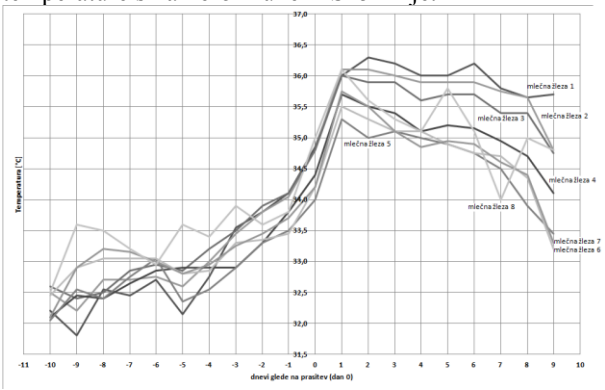
Slika 3: Razlika med mediano rektalnih temperatur in rektalno temperaturo posamezne svinje glede na dan prasitve

Slika 4 prikazuje mediano temperatur posameznih mlečnih žlez izmerjenih s termovizijsko kamero Flir T650sc glede na lokacijo mlečne žleze. Opazimo lahko dokaj enakomerno temperaturo mlečnih žlez pred prasitvijo. Ob začetku proizvodnje mleka imajo prve tri zaporedne mlečne žleze izrazito večji skok temperature, kot mlečne žleze za njimi. Od dneva pred prasitvijo in do dneva po prasitvi narastejo temperature mlečnih žlez od 1 °C do 2 °C. Naslednje dni so temperature mlečnih žlez konstantne glede na prejšnje dni, potem pa začnejo temperature mlečnih žlez zelo počasi padati. Zelo očitno je, da prve mlečne žleze dosegajo najvišjo temperaturo, glede na temperaturo jim sledijo druge in tretje. Mlečne žleze bolj zadaj so v dnevih po prasitvi hladnejše za 1 °C od sprednjih mlečnih žlez. Temperature na površini mlečnih žlez se pred prasitvijo gibljejo okoli 34,5 °C. Po prasitvi se temperature mlečnih žlez v povprečju gibljejo od 36 °C do 37 °C.

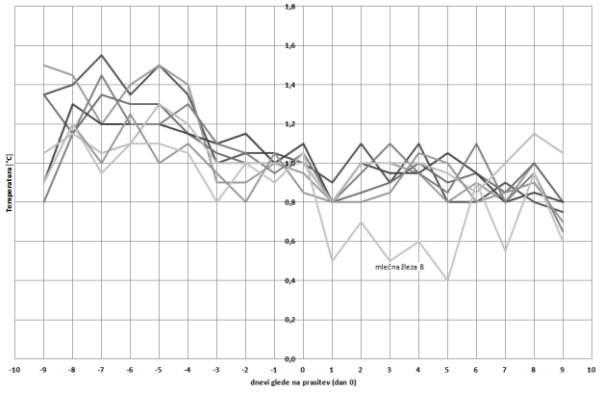


Slika 4: Mediana temperatur posameznih mlečnih žlez izmerjenih s termovizijsko kamero Flir T650sc glede na lokacijo mlečne žleze

Slika 5 prikazuje mediano temperatur posameznih mlečnih žlez izmerjenih s termovizijsko kamero Fluke TiS45 glede na lokacijo mlečne žleze. Opazimo enak trend temperatur kot na Sliki 4, le da so vse izmerjene temperature s kamero Fluke TiS45 nižje.



Slika 5: Mediana temperatur posameznih mlečnih žlez izmerjenih s termovizijsko kamero Fluke TiS45 glede na lokacijo mlečne žleze

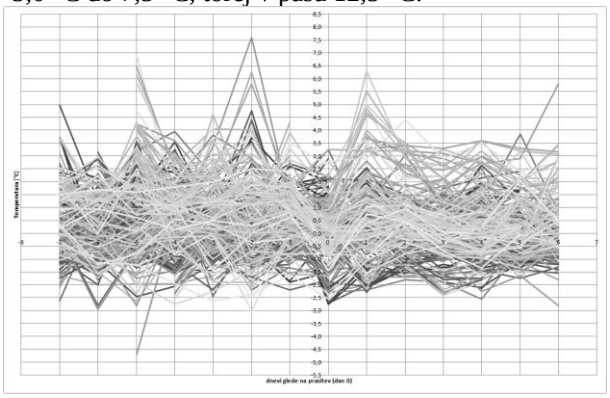


Slika 6: Mediana razlik izmerjenih temperatur posameznih mlečnih žlez izmerjenih s termovizijsko kamero FLIR T650sc in FLUKE TiS45 glede na dan prasitve

Slika 6 prikazuje mediano razlik izmerjenih temperatur posameznih mlečnih žlez izmerjenih z referenčno termovizijsko kamero Flir T650sc in Fluke TiS45 glede na dan prasitve. Opazimo, da smo pred

prasitvijo s termovizijsko kamero Fluke TiS45 merili do 1,5 °C nižjo temperaturo mlečne žleze kot s termovizijsko kamero Flir T650sc. Ob in po prasitvi je bila ta razlika manjša, a še vedno okoli 1 °C.

Slika 7 prikazuje odstopanja oz. raztros temperatur mlečnih žlez od mediane temperatur mlečnih žlez merjenih z obema kamerama glede na dan prasitve. Vse temperature zdravih svinj se gibljejo od -3,0 °C do 4,5 °C od mediane, torej v pasu 7,5 °C. V primeru svinje 5, ki je pred prasitvijo preživela pljučnico, imamo to odstopanje temperatur od mediane še veliko večje. Pri svinji 5 temperature mlečne žleze odstopajo od -5,0 °C do 7,5 °C, torej v pasu 12,5 °C.



Slika 7: Odstopanje oz. raztros temperatur mlečnih žlez od mediane temperatur mlečnih žlez merjenih z obema kamerama glede na dan prasitve

V obdobju od 24. 11. 2017 do 4. 2. 2018 je bilo na oddelku čakališča na 5 min zabeleženih 20710 meritev temperature zraka. Temperature so se gibale od 7,8 °C do 17,4 °C. Povprečna temperatura je bila 13,3 °C.

Relativna vlažnost zraka v obdobju meritev v porodnišnicah se je gibala od 55 % do 91 %. Mediana in srednja vrednost meritev je 70,9 % izmed 19297 meritev.

Na oddelkih porodnišnic, kjer smo izvajali meritve temperatur mlečnih žlez smo zabeležili 13050 meritev temperatur. Mediana in povprečna vrednost temperatur sta 15,8 °C. Najnižjo temperaturo smo zabeležili 10,8 °C in najvišjo 21,0 °C.

Pod pokrovom talnega grelca za sesne pujske se je temperatura zraka gibala med 17,0 °C in 21,5 °C. Največja razlika od te temperature 1 m nad pokrovom grelca je bila 4,8 °C. Mediana in povprečje 3462 meritev kažeta na 2,7 °C višjo temperaturo 10 cm pod pokrovom talnega grelca, kot 1 m nad pokrovom grelca.

Na 69-ih pujskih starih do enega dneva smo izmerili temperature kože od 25,9 °C do 38,4 °C. Povprečna vrednost izmerjenih temperatur je 33,1 °C in mediana teh meritev 33,6 °C.

Na 147-ih pujskih starih od enega do štirih dni smo izmerili temperature kože od 31,4 °C do 39,1 °C. Povprečna vrednost izmerjenih temperatur je 36,0 °C in mediana teh meritev 36,1 °C.

6 Zaključek

V 47-ih dneh meritev na prašičji farmi smo na 21-ih svinjah opravili 352 meritev rektalne temperature. S termovizijskima kamerama Flir T650sc in Fluke TiS45 smo izvedli 4742 meritev temperature mlečne žleze. Skupaj smo analizirali 9484 meritev na 312-ih mlečnih žlezah.

Rektalna temperatura se po prasitvi poviša za približno 0,8 °C. Rektalne temperature so od mediane rektalnih temperatur odstopale v pasu 3,8 °C, kar je fiziološko pogojeno.

Do prasitve so temperature mlečnih žlez dokaj enakomerne. Po prasitvi se temperature mlečnih žlez povišajo za približno 2 °C. Najvišjo temperaturo po prasitvi imata prvi mlečni žlezi, nato sledita drugi in tretji mlečni žlezi. Prvi mlečni žlezi imata običajno vsaj za 0,5 °C višjo temperaturo od četrtih mlečnih žlez. Temperature mlečnih žlez so od mediane temperatur mlečnih žlez glede na dan prasitve odstopale v pasu 7,5 °C v primeru svinje, ki je preživela pljučnico, je bil raztros temperatur mlečnih žlez dvakrat večji. Že sami pujski kažejo na to, da so prve mlečne žleze najkvalitetnejše, saj pri njih sesajo največji in močnejši pujski.

Obe termovizijski kameri sta pokazali podoben trend gibanja temperatur mlečnih žlez. S kamero višjega cenovnega razreda in z manjšo merilno negotovostjo smo vse temperature izmerili kot višje od temperatur nižje cenovne kamere, toda vse temperature so bile v območju merilne negotovosti. Pri analizi termogramov se je izkazalo, da ima slabša kamera večjo merilno napako na robu merjenega objekta.

Pujski se prasijo z enako temperaturo kot jo ima svinja. Nato pa se začnejo hitro ohlajati. Zato je zelo pomembno, da pujske čim hitreje osušimo in jim nudimo dodatno gretje. Malo starejši pujski so toplejši, saj imajo že bolje razvito termoregulacijo. Vendar je kljub temu potrebno še dodatno ogrevanje gnezda, saj v nasprotnem primeru porabljajo lastne zaloge energije za ogrevanje telesa in posledično slabše priraščajo.

V našem primeru pokrov nad talnim grelcem pomaga zadrževati nekaj toplote, saj je zrak nad njim skoraj 3 °C hladnejši.

Termografija se zaradi velikega raztrosa izmerjenih vrednosti temperature mlečnih žlez ni izkazala za zanesljivo oziroma dovolj točno metodo za diagnozo sindroma MMA.

Literatura

[1] J. R. Speakman in S. Ward, „Infrared thermography: principles and applications, “ Department of Zoology, University of Aberdeen, vol. 101, str. 224-232, 1998.

[2] Lek veterina, Mastitis v obdobju okrog prasitve pri svinjah [Online]. Dosegljivo: https://www.lek-veterina.si/blog/mastitis-v-obdobju-okrog-prasitve-pri-svinjah-1. [Dostopano: 10. 5. 2018].

Distributed Signal Processing in Sensor Networks

Gordan Zelinac
Faculty of Electrical Engineering
University of Banja Luka
Banja Luka, Bosnia and Herzegovina
gordan.zelinac@gmail.com

Mitar Simić, Zdenka Babić and Vedran Jovanović
Faculty of Electrical Engineering
University of Banja Luka
Banja Luka, Bosnia and Herzegovina
{mitar.simic, zdenka.babic, vedran.jovanovic}@etf.unibl.org

***Abstract*—Modern sensor networks can have large numbers of sensor nodes. The processing power and the number of features of those sensor nodes are constantly increasing. On the other side, the price per sensor node is decreasing which enables building low-cost sensor networks. In such networks it is mandatory to have effective signal processing schemes to extend life time of the network as well as to reduce maintenance. With the increase of processing power per sensor node, the signal processing can be distributed across all sensor nodes which improves the signal processing power of the complete network. In this paper, we described the theoretical background and the practical implementation of a distributed signal processing in sensor network for visualizing various gas concentration in air.**

***Index Terms*—distributed signal processing, sensor networks, data acquisition.**

I. Introduction

Sensor networks are composed of large numbers of sensor nodes, which perform phenomenon observation. Sensor nodes can be deployed either inside the phenomenon or very close to it. Each sensor node can have various types of sensors elements attached to it. The nodes are usually battery powered. Data acquired from sensor elements is processed on sensor node level. In this way the signal processing is distributed across the complete sensor network. Having each sensor node doing data processing, enables the use of various distributed signal processing techniques and increases the data throughput of the network [1].

To be able to perform data processing, sensor nodes must have certain hardware capabilities which enable them to perform this task. For example, this hardware capabilities can include considerably fast central processing unit (CPU), dedicated digital signal processor (DSP), fast vector processing units or adequate amount of working memory. With the increase in hardware requirements the price per sensor node is also increasing. Since every sensor network needs to be price-justified there is a trade-off between processing power and price per node. Thanks to advances in semiconductor manufacturing and low production costs, it is possible to incorporate considerably powerful hardware into sensor nodes without drastically increasing their price [2].

Sensor networks can have large number of sensor nodes and adequate network topology should be used to define how data is routed between nodes. Sensor nodes share the same transfer medium and therefore use of various medium access control (MAC) protocols is required to avoid collisions and possible data loss. This protocols should be energy efficient so that the life time of a sensor node is long as possible.

In this paper, we described the theoretical background and the practical implementation of a distributed signal processing in sensor network for visualizing various gas concentration in air. Our contribution is to use low-cost devices as sensing nodes with basic signal processing on the node, while the main signal processing is done on the powerful fusion node. Visualization of collected data is implemented using web application, thus the results can be displayed on wide range of devices such as laptops, mobile phones or tablets. The network is designed in such way that it can be easily modified to support various sensor types. The hardware is built using modern semiconductor components with very low price per node.

This paper is organized as follows: Section II describes the theoretical background and basics of sensor networks. In Section III, practical realization of our proposed network architecture is given. Finally, the main results are summarized and directions of future work are discussed in the concluding Section IV.

II. Methods

In this section we described main theoretical parameters of sensor networks with detailed description of solutions suitable for our practical realization of wireless sensor network as widely used concept of connection between sensor nodes.

A. Network topology

Network topology defines how data is routed between sensor nodes. In sensor networks, three most used topologies are [3]:

- parallel topology,
- serial topology, and
- tree topology.

Parallel topology is primarily used in wireless sensor networks because this topology allows adding or removing sensor nodes without interrupting normal network functioning. Fig. 1 shows a block diagram of parallel fusion topology. In this topology one node (fusion node) is responsible for collecting data from all other sensor nodes in the network. Sensor nodes are also called decision makers (DM) because they observe the phenomenon and make local decisions based on that observation. The fusion node is usually more powerful in terms of processing power in contrast to other nodes in network so it

can process fast all incoming data. Main advantages of parallel topology are [3]:

- small amount of data transferred from sensor nodes to the fusion node,
- network will normally operate in case when some sensor nodes fail, and
- sensor nodes can be placed within wide area.

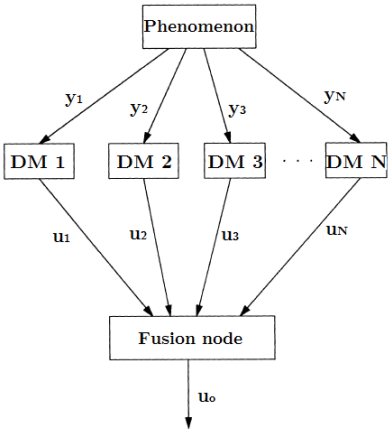


Fig. 1. Block diagram of parallel fusion topology.

In serial topology, first sensor node makes a estimation of the phenomenon and forwards it to next sensor node which combines his observation with received value to create value which forwards to next sensor node, and so on. The last sensor node creates final estimation. Advantage of serial topology is low power consumption and long life time of the network. However, main disadvantage is that sensor nodes located far away from phenomenon can make inaccurate observation or to have very small contribution to the final estimation.

Tree topology is based on unique transmission line for data transfer from all sensor nodes to fusion node. Sensor nodes and fusion node forms directed acyclic graph with fusion nodes placed in the root of the graph. Three topology enables use of advanced routing algorithms to reduce power consumption. However, changes of the current network are not simple because all nodes in the network have to be informed that new node is added or removed. This task can be complex in case of networks with huge number of nodes.

B. Decision rules

In sensor networks, sensor nodes are observing a phenomenon and create their local estimation. This estimations are sent to the fusion node. The fusion node has to decide which of the received estimation is closest to the correct one. To perform this task, decision rules are used. The decision rules can be formed using [3]:

- logical functions as AND or OR, and
- advanced methods provided by statistical decision theory (Bayesian estimation, minimax detection, Neyman-Pearson test, sequential detection, Constant False Alarm Rate, detection of local minimum, etc.)

The use of logical functions to define decision rules has the advantage of simple and fast implementation. The disadvantage of this approach is that the decision rules are not optimal and for very large number of observations this rules become hard to manage. On the other side, using advanced methods provided by the statistical decision theory has the advantage that the decision rules are optimal and applicable to large number of observations. The disadvantage of this approach is complex and time consuming implementation.

C. MAC protocols

In wireless sensor networks, sensor nodes share the same data transfer medium. Collisions occur when two or more nodes try to send data at the same time and this can lead to data loss. MAC protocols are used to resolve every possible collision that may occur during data transmission and to guarantee that no data is lost during that process. This protocols must be energy efficient since the sensor nodes are battery powered. Protocol complexity must also be taken into account since the underlying hardware has various limitations in terms of processing power, available memory and execution time. MAC protocols can be grouped into two categories [4]: collision based and collision free protocols. Collision based protocols allow collisions to occur but they have various mechanism which allow them to recover from this collisions without data loss. Collision free protocols do not allow any collision to occur because they give exclusive access to the data transfer medium for the node which is sending the data. After the data transfer is completed, the data transfer medium is accessible by other nodes.

III. Experimental

In previous section of the paper, we have presented basics of distributed signal processing in sensor networks as well as most important parameters of sensor networks. As an example of practical realization of distributed signal processing in sensor network we have created wireless sensor network for monitoring and visualizing various gas concentration in air in wide indoor area. Based on the given theoretical analysis, the most suitable solution was parallel fusion topology consisted of sensor nodes and fusion node, while for data transfer a collision free MAC protocol was used. Fusion node uses the OR logical function to define decision rules.

A. Sensor node

The hardware of the sensor nodes is based on Arduino Uno development board [5]. At sensor nodes, Arduino Uno executes the main program, coordinates the work of peripheral devices and it is connected with Arduino Ethernet Shield to attached micro SD card used as local data storage device. Used sensors are:

- air quality sensor (MQ-135) suitable for detecting from 10 ppm to 1000 ppm of NH_3, NOx, alcohol, benzene, smoke, CO_2, etc. [6],
- CO gas sensor (MQ-7) suitable for detecting from 20 ppm to 2000 ppm [7],

- methan gas sensor (MQ-4) suitable for detecting from 200 ppm to 10000 ppm [8], and
- air temperature and relative humidity sensor (DHT-11) for 20%-80% relative humidity range and 0°C-50°C temperature range [9].

For wireless communication and data transfer from sensor nodes to the fusion node, a ultra low power 2.4 GHz RF module nRF24L01 [10] from Nordic semiconductor was used. It is suitable for distances up to 100 m outdoor [10], however for reliable indoor use expected maximum distance between modules is about 20 m. We chose sampling time of 15 seconds because it is possible to have significant change of gas concentration during short time period. However, one data entry requires 19 bytes which means that standard SD card with 4 GB capacity can storage data for several years. Electrical schematic of sensor node is shown in the Fig. 2.

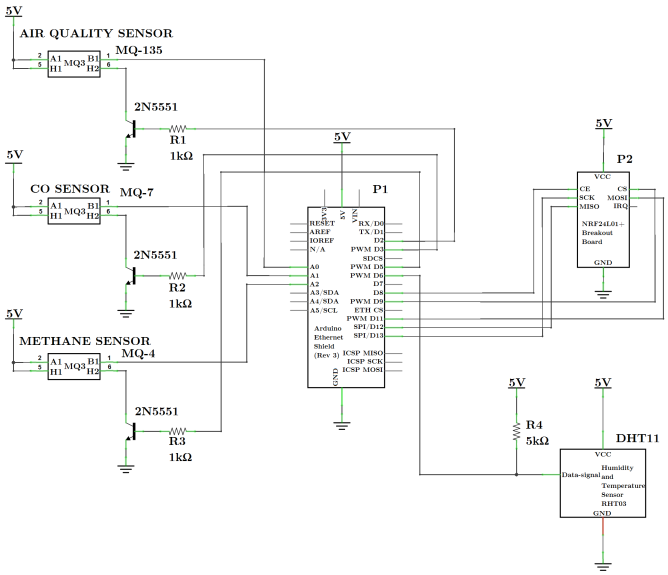


Fig. 2. Electrical schematic of sensor node.

In the Fig. 3 the hardware outcome of the one sensor node is shown.

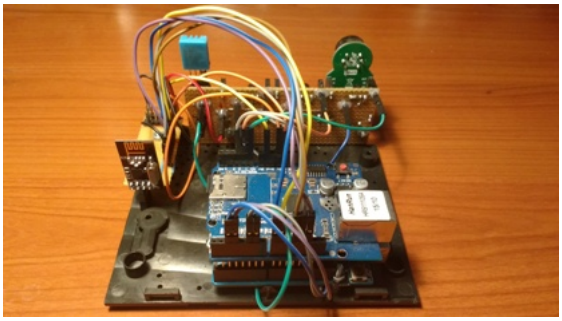


Fig. 3. Practical implementation of sensor node.

B. Fusion node

The hardware of the fusion node is based on Raspberry Pi development board [11]. In our realization, fusion node does not have attached sensors as proposed function of the fusion node is to collect data streams from sensor nodes and to give visualization of measured values. Fusion node has only one peripheral device connected (nRF24L01) for wireless communication with sensor nodes. The node is running Linux operating system with installed MySQL database and a web server. The web server hosts the application which is used to present the measurements to the end user. Electrical schematic of fusion node is shown in the Fig. 4.

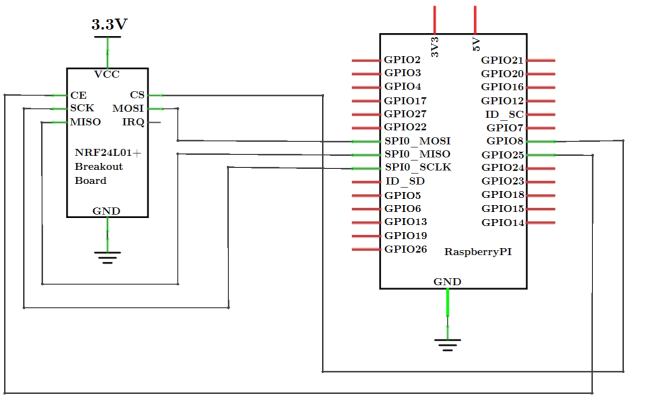


Fig. 4. Electrical schematic of fusion node.

In the Fig. 5 the hardware outcome of fusion node is shown.



Fig. 5. Practical implementation of fusion node.

Fusion node has significantly more complex software in comparison with sensor nodes. Multi layer structure of fusion node's software is presented in the Fig. 6.

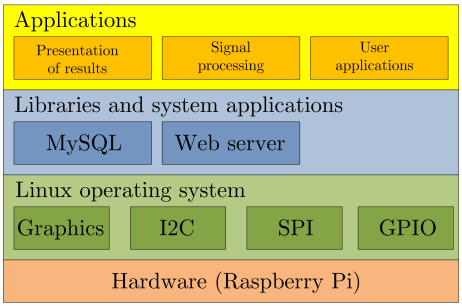


Fig. 6. Multi layer structure of fusion node's software.

Distributed signal processing is performed with two main applications:

- application written in C++ programming language dedicated to the control of attached RF module, processing of data received from sensor nodes and writing of data to the database,
- application for visualization of database entries to the end users who access to it with web browsers.

These applications are running independently which ensures stable and reliable operation of the system. The web application for visualization of database entries acquired by the sensor network, allows to users to select different measurements to be displayed from the database. Every sensor node has its own graph which is dynamically updated as new data arrive on fusion node.

As a part of validation of developed sensor network, we conducted experiment with four sensor nodes during 7 days (approximately 40000 samples per node). Because of limited space, in this paper measurement results from just one sensor node are shown in the Fig. 7.

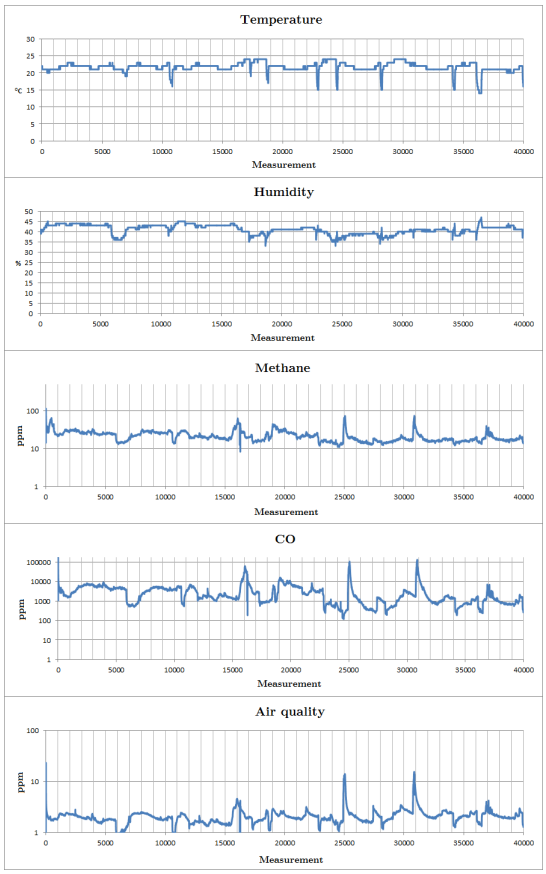


Fig. 7. Measurement results from one sensor node.

IV. Conclusion

In this paper we presented approach in distributed signal processing in sensor networks. In our approach the fusion node, based on the powerful Raspberry Pi, coordinates the work of sensor nodes (based on the low-cost Arduino Uno platform) which are deployed in the field, collects the measured data from sensors of all nodes and presents measured values to end users. Remote access of the end users to measured values is possible from web browser from any location in the world.

Limitations of current realization are based on the Arduino Uno capabilities regarding input pins for sensor connections (14 in total, 6 analog inputs) as well as working memory (2 KB) for implementation of algorithms for more complex signal processing at node. Moreover, security of transferred data requires more attention in future use. Arduino Uno-based sensor nodes are able to create encrypted messages using Advanced Encryption Standard with three different key lengths: 128, 192 and 256 bits without losing performances needed for measurements and signal processing.

Recognized limitations regarding maximum number of attached sensors as well security of transferred data are directions of our future work. Moreover, we would like to investigate what is influence of different decision rules to the final measurement result as well as if changing of network topology would change performances of complete solution.

Possible applications of developed sensor network are in indoor environmental parameters monitoring in office buildings or food/medical storages, etc. Moreover, modular approach in design of sensor architecture allows connection of additional sensors which will provide information of different parameters important for specific applications.

Acknowledgement

The research was supported by Bilateral Collaboration Project of Slovenian Research Agency (ARRS) and Ministry of Civil Affairs of Bosnia and Herzegovina under grant 19/6-020/964-4/16, and by Ministry of Civil Affairs of Bosnia and Herzegovina with project number 8300586.

References

[1] C. Y. Chong, S.P. Kumar, “Sensor networks: Evolution, opportunities and challenges” Proceedings of IEEE, 91 (8), pp.1247-1256, 2003.
[2] P. K. Varshney, “Distributed detection and Data Fusion” Springer-Verlag New York, 1997.
[3] W. Dargie, C. Poellabauer “Fundamentals of Wireless Sensor Networks: Theory and Practice” John Wiley & Sons, 2010
[4] Aguirre E, Lopez-Iturri P, Azpilicueta L, Astrain JJ, Villadangos J, Falcone F., ”Analysis of Wireless Sensor Network Topology and Estimation of Optimal Network Deployment by Deterministic Radio Channel Characterization”, Sensors, 15(2), pp. 3766-3788, 2015.
[5] Arduino - Home”, Arduino.cc, 2018. [Online]. Available: https://www.arduino.cc/. [Accessed: 15.07.2018].
[6] MQ-135, datasheet. Available: https://www.olimex.com/. [Accessed: 15.07.2018].
[7] MQ-7, datasheet. Available: https://www.sparkfun.com/. [Accessed: 15.07.2018].
[8] MQ-4, datasheet. Available: https://www.sparkfun.com/. [Accessed: 15.07.2018].
[9] DHT-11, datasheet. Available: https://www.mouser.com/. [Accessed: 15.07.2018].
[10] nRF24L01, datasheet. Available: https://www.nordicsemi.com/. [Accessed: 15.07.2018].
[11] R. Foundation, ”Raspberry Pi — Teach, Learn, and Make with Raspberry Pi”, Raspberry Pi, 2018. [Online]. Available: https://www.raspberrypi.org/. [Accessed: 15.07.2018].

Vpliv napajalnih virov na meritev segrevanja stikalnih napajalnikov

Tadej Prinčič, Boštjan Glavič, Gaber Begeš

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška c. 25, 1000 Ljubljana
Slovenski institut za kakovost in meroslovje, SIQ Ljubljana, Tržaška c. 2, 1000 Ljubljana
E-pošta: tadej.princic@gmail.com

The impact of the power sources to the temperature measurement of the switching power supplies

Power supply source is important when performing input power measurement and temperature measurement during safety investigation of the product. Different source can result in different measured result.

The goal of this project was to make a comparison measurements on three different power supplies while they are supplied by different power sources. Two power sources were electronic stabilized power sources and one power source was an adjustable transformer connected to public grid. Products under test were switch mode power supplies, two of them with implemented power correction factor (PFC) and one without PFC

We have found out that supply source does not influence much on input power measurements or temperature measurements when we are measuring product with PFC. On contrary, we found out that the type of power source has big impact on measured results when we are measuring product without PFC. The reason is in compensation of reactive power. Measured current and measured temperatures on input filter are much lower when product without PFC is supplied with normal grid

1 Uvod

Javno električno omrežje je podvrženo nihanju napetosti zaradi priklopa različnih porabnikov. Med preskušanjem varnosti električnih proizvodov so pomembni ponovljivi rezultati, zato se uporablja stabilizirane vire. Vse novejše generacije so narejene na osnovi stikalne tehnologije. V laboratoriju za elektroniko na SIQ so prišli do ugotovitve, da vrsta napajalnega vira vpliva na vrednost izmerjenega toka v porabniku. V določenih primerih je tok porabnika tudi do 20 % večji, ko se napaja iz elektronskega stabiliziranega vira v primerjavi s porabo iz javnega omrežja. Odstopanja so odvisna od napajalnega vira ter od vrste napajalnika. S korekcijskim faktorjem (PFC) je bila razlika manjša kot brez PFC.

Naloga je bila, da izvedemo meritve porabe na treh različnih napajalnikih z uporabo treh različnih napajalnih virov ter nato primerjamo izmerjene rezultate.

Za teste smo uporabljali naslednje napajalne vire:

- elektronski stabiliziran vir Elettrotest ID E308,
- elektronski stabiliziran vir Spitzenberger ID E43,
- javno omrežje skupaj z nastavljivim transformatorjem.

Teste smo izvajali na napajalnikih:

- PSU1,
- PSU2,
- PSU3.

2 Elektronski stabilizirani viri

Viri, ki se uporabljajo za preskušanje varnosti v okviru International Standard System of Conformity Assessment Schemes for Electrotehnical Equipment and Components morajo ustrezati parametrom, ki jih specificira operativni dokument IECEE OD-5010:2015 [2].

V njem je točno določeno, kako se mora izvesti validacija vira ter pod kakšnimi pogoji. Biti mora znotraj pogojev, kot je napetostna stabilnost (+/- 3 %), frekvenčna stabilnost (+/- 2 %), harmonično popačenje – THD (< 5 %). Primer rezultatov vira Elettrotest vidimo na spodnji *Tabeli 1*. Meritev posameznega parametra traja približno 1 uro.

Lastnost vira je pomembna za preskušanje varnosti napajalnika. V kolikor napetost med izvajanjem meritev niha, ali je le ta popačena lahko dobimo napačne rezultate.

V okviru seminarske naloge smo meritve opravljali pri dveh napetosti 120 V in 230 V.

Tabela 1. Rezultat operativnega dokumenta IECEE za elektronski vir Elettrotest.

Measured quantity	**Value**
Voltage nominal, V_{norm} =	230 V
Max. open circuit voltage, $V_{m,max}$ =	230,1 V
Min. open circuit voltage, $V_{m,min}$ =	229,9 V
Current loaded, I_{el} =	3,12 A
Max. voltage loaded, $V_{el,max}$ =	229,8 V
Min. voltaga loaded, $V_{el,min}$ =	229,5 V
Max. frequency open circuit, $F_{m,max}$ =	50,01 Hz
Min. frequency open circuit, $F_{m,min}$ =	49,99 Hz
Max. frequency loaded, $F_{el,max}$ =	50,00 Hz
Min. frequency loaded, $F_{el,min}$ =	49,99 Hz
Max. harmonic distortion open circuit, THD_{oc} =	0,050 %
Max. harmonic distortion loaded, THD_{in} =	0,110 %
Reg V_{oc} =	0,061 %
Reg V_{io} =	0,213 %
Reg F_{oc} =	0,018 %
Reg F_{in} =	0,010 %

3 Zahteve standarda IEC/EN 60950-1

Standard IEC/EN 60950-1 pokriva varnost informacijske tehnologije in naprave kot so monitorji, računalniki, telekomunikacijske naprave ter napajalnike, ki se uporabljajo s to opremo.

Vsak proizvod mora biti varen preden gre na trg, da ne pride do električnega udara, požara, opeklin zaradi prevročih dostopnih delov ali mehanskih poškodb. Vključeni so tudi nenormalni pogoji, kjer uporabnik uporabi napravo izven specifikacij ali pride do odpovedi kakšne od komponent [3].

Proizvajalec mora specificirat, kje se naprava lahko varno uporablja. Primeri takšnih parametrov so temperatura okolice, stopnja onesnaženosti, nadmorska višina, prenapetostna kategorija itd. V vsakem primeru mora biti proizvod varen. V standardu je predvidena dvojna zaščita uporabnika. Prva zaščita deluje med normalnim delovanjem naprave, druga deluje v primeru nenormalnih pogojev, ko prva zaščita odpove.

Določena je meja za varno dostopno napetost, meja za temperaturno dostopnih delov, meja za gibljive dele naprave itd. Glavni testi, ki se izvajajo na napajalniku v okviru tega standarda so; meritev porabe, temperaturni test, test zračne in površinske razdalje, test prebojne trdnosti, test odvodnih tokov, test nenormalnih pogojev in test pogojev odpovedi komponent.

3.1. Meritev porabe napajalnika

Pri meritvi preverjamo kakšno porabo ima napajalnik in jo primerjamo z nazivno, ki jo deklarira proizvajalec na napisni tablici. Paziti moramo, da v stanju dinamičnega ravnovesja vhodni tok ne presega nazivnega za več kot 10 %.

Tabela 2. Meritev porabe na napajalniku PSU 1 z Elettrotest virom.

	Input Condition	Input Condition, A	Average Power	PF
Operating Condition	Volta / Hz	Ralted / Measured	Walts	
12,8 V / 9,5 A	90 / 50	_ / 2,96	156 / 267	0,584
12,8 V / 9,5 A	100 / 50	9,50 / 2, 68	154 / 268	0,575
12,8 V / 9,5 A	240 / 50	9,50 / 1,23	150 / 297	0,504
12,8 V / 9,5 A	264 / 50	_ / 1,16	150 / 306	0,490

Tabela 3. Meritev porabe na napajalniku PSU 1 s Spitzenbeger virom.

	Input Condition	Input Condition, A	Average Power	PF
Operating Condition	Volta / Hz	Ralted / Measured	Walts	
12,8 V / 9,5 A	90 / 50	_ / 2,81	155 / 253	0,613
12,9 V / 9,5 A	100 / 50	9,50 / 2,53	154 / 253	0,607
12,9 V / 9,5 A	240 / 50	9,50 / 1,21	150 / 290	0,516
12,8 V / 9,5 A	264 / 50	_ / 1,13	150 / 298	0,502

Tabela 4. Meritev porabe na napajalniku PSU 1 pri napajanju preko omrežja s transformatorjem.

	Input Condition	Input Condit, A	Average Power	PF
Operating Condition	Volta / Hz	Ralted / Measured	Walts	
12,8 V / 9,5 A	90 / 50	_ / 2,74	154 / 247	0,625
12,9 V / 9,5 A	100 / 50	9,50 / 2,40	153 / 241	0,634
12,9 V / 9,5 A	240 / 50	9,50 / 1,18	149 / 273	0,545
12,9 V / 9,5 A	264 / 50	_ / 1,04	149 / 275	0,542

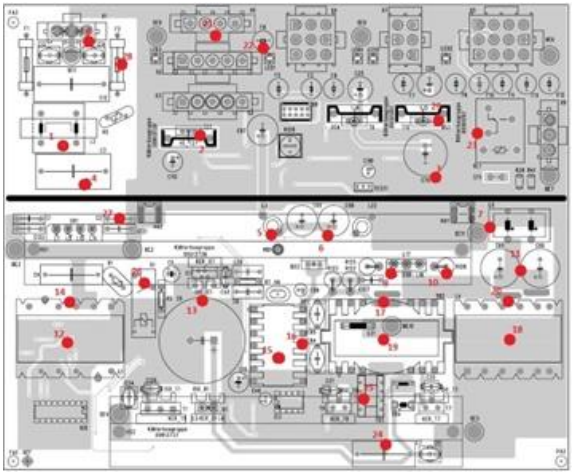
Meritev smo opravili pri štirih različnih napetostih. Napajalnik smo na izhodu obremenili z nazivno napetostjo, z merilnikom toka in moči pa merili

vhodni tok in moč. Rezultate za napajalnik PSU 1 pri napajalni napetosti z vsemi tremi viri vidimo v Tabelah 2, 3 in 4. Dva napajalnika, ki imata PFC nista imela bistveno odstopanje pri različnih virih. Vidno odstopanje je imel napajalnik PSU 1.

Največjo razliko opazimo pri meritvi z nestabiliziranim virom, kajti napajalnik ne more zgladiti napetost. Omenjen test, so vsi trije stikalni napajalniki prestali, saj izmerjeni vhodni tokovi niso presegali 110 % vrednosti nazivnega toka.

3.2. Segrevanje napajalnika (temp. test)

Segrevanje napajalnika izvajamo v komori kjer vzdržujemo maksimalno temperaturo okolice, ki jo specificira proizvajalec.

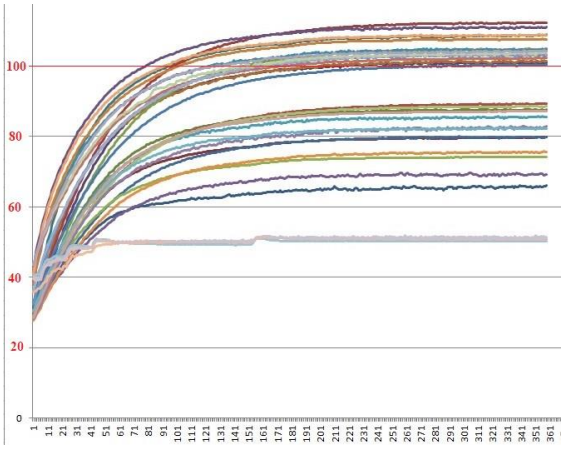


Slika 1. Prikaz sheme PSU 3 napajalnika, kjer so postavljeni termočleni.

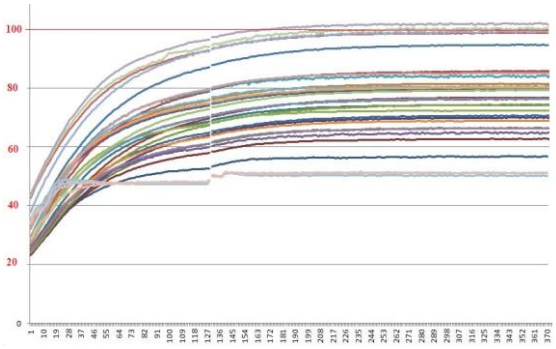
Pred pričetkom meritev, moramo na proizvod namestiti termočlene, s katerim merimo temperaturo. V standardu je predpisano katere komponente so kritične za varnost, zlasti tisti deli kateri so dostopni [2]. S pomočjo termične kamere ugotovimo kateri deli ohišja so podvrženi najvišji temperaturi. Za merjenje uporabljamo termočlene tipa K, ki jih tudi priporoča mednarodna IECEE CB shema. S pomočjo posebnega lepila jih pritrdimo na določene točke. Vidne so na sliki (1).

Najbolj kritična meritev za varnost temperature je na transformatorju. Le ta ločuje nevarno napetost od varne. Maksimalna dovoljena temperatura transformatorja je odvisna od vrste izolacije in se giblje od 100 °C za razred A do 165 °C za razred H. Proizvajalci tipično uporabljajo razred B, katerega je meja 110 °C ali razred F za katerega je meja 130 °C.

V našem primeru smo uporabljali merilnik temperature Agilent 34980, ki omogoča priklop do 180 termočlenov ter omogoča sprotno spremljanje rezultatov na računalniku.



Slika 2. Prikaz meritev temperature na napajalniku PSU 2 pri napajalni napetosti 90 V.



Slika 3. Prikaz meritev temperature na napajalniku PSU 2 pri napajalni napetosti 264 V.

Meritev traja dokler se temperature ne ustalijo. V povprečju traja minimalno dve uri. Na sliki (2) je viden temperaturni odziv PSU 2. Opazimo da so pri nižji napajalni napetosti bistveno višje temperature. Razlike so na sliki (3) z večjo napetostjo. Temperatura je za 10 °C višja, saj je pri nižji napetosti tok večji, s tem pa se elementi bolj segrevajo.

4 Rezultati

Primerjavo rezultatov meritev porabe smo izvedeli na treh stikalnih napajalnikih. Eden izmed teh ni imel PFC. Ravno ta faktor je povezan s kompenzacijo jalove energije, katera pride iz električnega omrežja v omenjeni vlogi moči. Napajalnik PSU 1 ima faktor moči manj kot 0,95. To je meja, ki je še dopustna. Izračuna se jo po enačbi *1*.

$$PF = \frac{P\,[W]}{U \cdot I} \qquad (1)$$

S podatkom faktorja moči je opazna razlika pri porabi napajalnika. Ta je znašala okoli 10 mA. Pri drugih dveh napajalnikih, ki imata PFC, pa bistvenih razlik ni opaziti, poleg tega imata faktor moči nad oz. okoli 0,95. Podatke o moči, smo razbrali iz merilnika moči, tako za delovno kot za navidezno moč.

Drugi del naloge je bila meritev temperature na napajalniku. Opravljali smo jih v komori pri temperaturi 50 °C. En napajalnik je bil zmožen delovati pri tej temperaturi, in sicer PSU 2. Ostala dva napajalnika sta pri tem testu šla v pretemperaturno zaščito. Zaradi tega, smo zmanjšali temperaturo za 10 °C. Napajalnik PSU 1 se ni več izklapljal. Pri napajalniku PSU 3 je bilo potrebno upoštevati navodila proizvajalca glede montaže, da je bil zagotovljen minimalen zračni pretok.

Rezultati so pokazali, da napajalnika s PFC nimata večjih odstopanj med različnimi viri. Temperature se pri elementih bistveno ne razlikujejo. Navedeni so v Tabeli 5. Očitna razlika se pojavi pri napajalniku PSU 1. Te so znašale skoraj 10 °C. Razlog je predvsem, zaradi nestabilnosti napetosti in različnega vhodnega toka glede na uporabljen vir. Napajalnik nima glajenja energije in zato pride do odstopanj.

5 Zaključek

Pri uporabi različnih elektronskih stabiliziranih virov ne pride do bistvenih razlik. Opazne pa so med izmerjenimi rezultati pri uporabi nestabiliziranega vira (javno omrežje) v primerjavi z rezultati uporabe stabiliziranega elektronskega vira. Potemtakem lahko sklepamo, da je vir napajanja pomemben pri merjenju vhodne moči in temperature med varnostno preiskavo izdelka.

Posebej je potrebno paziti kakšen napajalnik imamo. Z meritvami smo ugotovili, da pride do večjih odstopanj prav pri napajalnikih brez PFC. V nadalje bi priporočali, da se pri meritvah na napajalnikih brez PFC, prej odloči glede vira. Definitivno je boljši elektronski stabiliziran vir. Poleg rečenega bi opozorili na pomembnost proizvajalčevih zahtev. Med te je navedena njegova postavitev, katero je potrebno upoštevati.

Tabela 5. Temperature posameznih elementov napajalnika PSU 1 pri različnih virih.

Supply volt. (Vdc)	90	264	90	264	90	264
Frequency (Hz)	50	50	50	50	50	50
Ambient T (°C)	40	40	40	40	40	40
Max. measured temp.	Maximum measured temperature (°C)					
	Elettrotest		Trafo		Spitzenberg	
1). CN1	62,5	55,1	59,7	55,2	61,7	54,9
2). F1	74,8	63,5	70,0	62,5	72,9	62,6
3). FL1	83,7	63,3	76,2	61,9	81,6	62,8
4). CX1	74,2	62,3	69,1	61,2	72,0	61,4
5). FL2	96,8	68,5	86,1	65,1	93,9	66,9
6). BD1	97,6	81,2	84,1	79,0	87,5	74,7
7). NTC1	94,7	82,1	90,0	77,9	93,3	80,4
8). C1	77,3	65,2	75,2	63,4	76,3	64,1
9). CY5	70,7	66,3	68,9	65,7	69,7	65,1
10). C7	87,4	73,7	84,7	71,6	86,0	72,5
11). T1 core	100,7	96,9	98,4	96,0	99,6	95,6
12). T1 winding	100,7	99,3	99,0	97,9	100,1	97,5
13).HS2	79,0	73,9	76,6	73,4	77,7	73,0
14). Q371 /Q370	83,1	78,5	80,8	78,0	82,0	77,5
15). C10C	90,4	86,2	88,6	85,9	89,5	85,6
16). C103	80,9	77,7	79,2	77,3	80,2	76,7
17). L101	79,2	76,4	77,4	75,8	78,3	75,3
18). C105	74,9	71,7	73,2	71,5	74,2	71,0
19). Right	77,5	73,9	76,2	73,8	76,4	73,3
20). Top	74,9	66,7	72,8	65,0	74,7	66,2

Literatura

[1] (ETI Elektroelement d.d. , 2017)

[2] (IECEE, 2016)

[3] (International standard IEC 60950 -1, 2005)

[4] (Wikipedia, 2017)

Akustika in elektroakustika

Acoustics and Electroacoustics

Uporaba mobilnih naprav za oktavno analizo zvoka

Metod Celestina[1], **Andrej Trost**[2]

[1]*EA LAB d.o.o., Cesta Ljubljanske brigade 23A, Ljubljana*
[2]*Fakulteta za elektrotehniko UL, Tržaska 25, Ljubljana*
E-pošta: metod.celestina@ea-lab.eu

Using mobile devices as an octave frequency analyzer

Excessive noise is a major public health issue and affects many millions of poeple world-wide. The noise measurements are done by using a dedicated sound level meter, but in many cases a mobile device (phone) has a potential to provide means of an adequate noise measurement. While several studies were conducted to determine the adequacy of mobile devices for noise measurement, none of the studies examined sound analysis beyond the basic parameters. In the paper we present the results of a study of octave analysis accucracy based on an in-house measurement app.

1 Uvod

Prekomeren hrup predstavlja problem za javno zdravje in pri prekomerno izpostavljenih osebah povzroča številne zdravstvene probleme, kot so okvara sluha, kardiovaskularne bolezni, povišan krvni tlak, motnje spanja in druge psihološke in sociološke probleme. Svetovna zdravstvena organizacija WHO (angl. World Health Organization) ocenjuje, da ima 466 milijonov ljudi sluh okvarjen do te mere, da je poslabšana njihova kvaliteta življenja [1], inštitut NIOSH (National Institute for Occupational Safety and Health) iz ZDA pa ocenjuje, da je približno 22 milijonov ameriških delavcev izpostavljenih škodljivemu hrupu [3].

Meritve hrupa v okolju in delovnih mestih so urejene z zakonodajo in zahtevajo uporabo merilnikov zvoka, ki so skladni z mednarodnim standardom IEC 61672 [10]. Medtem ko so namenski merilniki zvoka dostopni le profesionalnim uporabnikom, so dragi in pogosto zahtevajo lastniški program za obdelavo podatkov, je bilo v zadnjih letih izvedenih več projektov merjenja hrupa v okolju, kjer je bil za meritve hrupa uporabljen mobilni telefon, kot cenejša in dostopnejša alternativa namenskemu merilniku zvoka. Dostopnost in vseprisotnost mobilnih telefonov odpira možnost za nove načine merjenja hrupa, kot je množičenje (angl. crowd-sourcing) meritev [6, 7].

Število uporabnikov mobilnih naprav raste in bo doseglo 2,87 milijarde do leta 2020 [5]. Računska zmogljivost modernih mobilnih naprav je primerjiva z zmogljivostjo namiznih računalnikov. Po podatkih primerjalnega preizkusa Geekbench je zmogljivost mobilnega telefona Apple iPhone 7 približno enaka zmogljivosti računalnika, ki ga poganja procesor Intel Core i5 [2]. Takšna zmogljivost je zadostna za opravljanje zahtevnejše digitalne obdelave signalov. Sodobne mobilne naprave imajo tudi funkcije, ki v konvencionalnih merilnikih zvoka niso prisotne ali pa so prisotne le izjemoma. Pozicioniranje z uporabo sistema GPS (angl. Global Positioning System), stalna povezanost z internetom, kamera in bogat ter uporabniku prijazen uporabniški vmesnik je le nekaj primerov.

2 Mobilna naprava kot merilnik zvoka

Uporaba mobilne naprave za izvajanje resnih meritev hrupa je odvisna predvsem od zagotavljanja pravilnega merjenja in ustrezne natančnosti. Na tem področju je bilo narejeno več študij. Kardous in Shaw z ameriškega inštituta NIOSH sta izvedla pilotno študijo [11], v kateri sta ugotavljala ustreznost aplikacij za merjenje zvoka za meritve hrupa v delovnem okolju. Rezultati so pokazali, da nekatere aplikacije izkazujejo dovolj veliko natančnost (napaka manjša od 2 dB) in bi bile lahko uporabljane kot adekvaten merilnik zvoka oz. za namene presejalnih meritev. Ista avtorja sta v nadaljnih raziskavah [12, 14] ugotavljala natančnost pri uporabi zunanjega mikrofona. Raziskava je bila zožana za štiri mobilne aplikacije, ki so izkazovale največjo natančnost v prvi raziskavi. Rezultati so pokazali napako merjenja manjšo od 1 dB, kar je boljše od merilne negotovosti merilnika rezreda 2 (skladno z IEC 61260), ki znaša 1,9 dB. Avtorja Murphy in King sta v svoji raziskavi [13] razširila delo avtorjev Kardous in Shaw [11] s tem, da sta uporabila veliko stevilo različnih mobilnih naprav in aplikacij namesto izbrane množice. Rezultati so pokazali veliko variacijo med posameznimi napravami.

Vse meritve hrupa so na tak ali drugačen način standardizirane, uporabljena merilna oprema pa mora izpolnjevati zahteve določenih standardov. Za širše sprejetje mobilnih naprav za merjenje hrupa je nujno ovrednotiti skladnost meritve z relevantnimi standardi. V raziskavi [8] smo se osredotočili na skladnost merilnika zvoka, ki je bil sestavljen iz mobilnega telefona ter zunanjega mikrofona, z mednarodnim standardom IEC 61672 [10]. Merilnik smo izpostavili standardiziranim testom, ki se uporabljajo pri periodični kalibraciji profesionalnih merilnikov zvoka in ugotovili skladnost z zahtevami za Razred 2. Nadaljevanje tega raziskovalnega dela pa je ovrednotenje

skladnosti oktavne analize v mobilni aplikaciji za meritev zvoka z mednarodnim standardom IEC 61260 [9].

3 Oktavna analiza

Oktavna analiza je način frekvenčne analize, pri katerem frekvenčno vsebino signala ugotavljamo tako, da frekvenčno območje, ki nas zanima, razdelimo na frekvenčne pasove, katerih srednje frekvence si sledijo v logaritemskem zaporedju. Za vsak pas uporabimo pasovno prepustni filter, ki prepušča samo signal iz tega pasu, kateremu srednjo frekvenco f_m, in obe mejni frekvenci f_1 in f_2 določimo skladno z enačbami 1 2 in 3 (f_r je referenčna frekvenca 1000 Hz, $x \in \{..., -2, -1, 0, 1, 2, 3, ...\}$ indeks okravnega pasu, b število frekvenčnih pasov na oktavo in G 2 ali $10^{0,3} \approx 1,99526$).

$$f_m = f_r \cdot G^{\frac{x}{b}} \tag{1}$$

$$f_1 = f_r \cdot G^{-\frac{2}{b}} \tag{2}$$

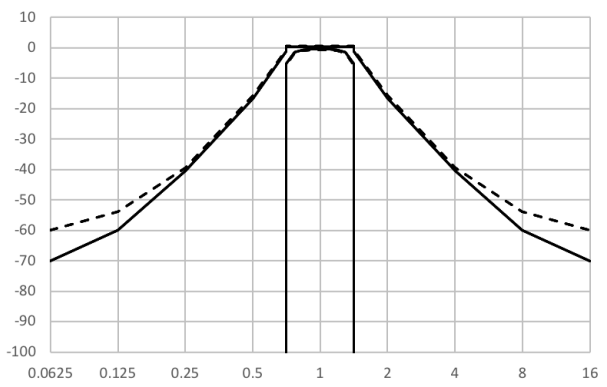
$$f_2 = f_r \cdot G^{+\frac{1}{2b}} \tag{3}$$

Tak način frekvenčne analize je zelo pogost v akustiki, kjer so širine frekvenčnih pasov tipično 1 oktava ali 1/3 oktave.

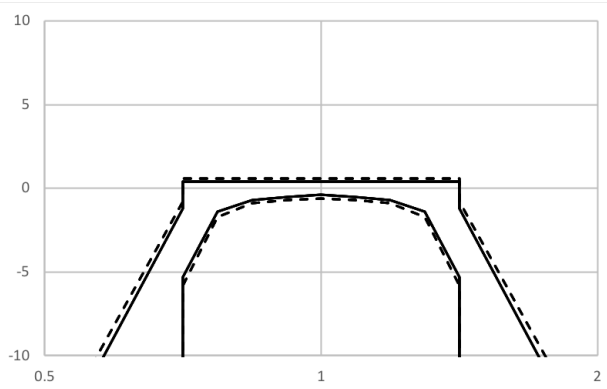
Zahteve za 1/n-oktavno analizo so podane v mednarodnem standardu IEC 61260 [9]. Standard predpisuje srednje frekvence oktavnih pasov, gabarite za prevajalno funkcijo pasovno prepustnih filtrov ter še nekatere dodatne zahteve, katerim morajo ustrezati oktavni filtri. Standard definira dva točnostna Razreda, Razred 1 in Razred 2, kjer Razred 1 zagotavlja večjo točnost. V raziskavi smo ugotavljali skladnost z zahtevami za prevajalno funkcijo pasovno prepustnih filtrov za vsak oktavni pas, efektivno pasovno širino vsakega od pasov, linearnost ter seštevanje signalov.

3.1 Prevajalna funkcija

Prevajalna funkcija vsakega izmed oktavnih pasov mora biti znotraj toleranc, ki so prikazane na slikah 1 in 2. Tolerance za Razred 1 so prikazane s polno, za Razred 2 pa prekinjeno črto. Frekvenčna skala je relativna in je normalizirana za srednjo frekvenco filtra.



Slika 1: Tolerance za Razred 1 in Razred 2.



Slika 2: Tolerance za Razred 1 in Razred 2 v propustnem pasu.

3.2 Efektivna pasovna širina

Efektivna pasovna širina ΔB je merilo, koliko ploščina pod prevajalno funkcijo filtra odstopa od prevajalne funkcije idealnega pasovno prepustnega filtra in je definirana kot

$$\Delta B = 10 \cdot \log \frac{\int_o^\infty \frac{1}{\Omega} \cdot 10^{-0,1 \cdot \Delta A(\Omega)}}{\frac{1}{b} \ln G} \tag{4}$$

Integral v števcu enačbe 4 se v praksi izračuna numerično s kvadraturno formulo. Za izpolnjevanje zahtev standarda, mora biti ΔB manjši od 0,4 dB (Razred 1) ali manjši od 0,6 dB (Razred 2).

3.3 Seštevanje signalov

V primeru idealnih filtrov bi bil seštevek RMS vrednosti signalov na izhodu vsakega izmed oktavnih filtrov enak RMS vrednosti vhodnega signala. Realna prevajalna funkcija filtrov pa je taka, da obstaja delno prekrivanje med sosednjimi pasovi, zato je seštevek na izhodu filtrov običajno nekoliko večji od vrednosti na vhodu. Za izpolnjevanje zahtev standarda ne sme seštevek na izhodu odstopati od vhodne RMS vrednosti za več kot +0,8/-1,8 dB (Razred 1) oz. +1,8/-3,8 dB (Razred 2) pri katerikoli frekvenci, ki leži med srednjima frekvencama dveh sosednjih frekvencčnih pasov.

3.4 Linearnost

Linearno območje delovanja se meri pri srednji frekvenci vsakega izmed pasovno prepustnih filtrov. V linernem območju amplituda izhodnega signala sledi amplitudi vhodnega in od pričakovane vrednosti ne odstopa več kot 0,5 dB (Razred 1) oz. 0,6 dB (Razred 2) v zgornjih 40 dB linearnega območja. V preostanku linearnega območja je dovoljeno odstopanje 0,7 dB (Razred 1) oz. 0,9 dB (Razred 2). Zahtevano območje za Razred 1 je 60 dB, za Razred 2 pa 50 dB.

4 Metoda testiranja

V raziskavi smo se omejili na izpolnjevanje zahtev pri uporabi zunanjega mikrofona. Zunanji merilni mikrofon zagotavlja boljši zajem signala in uporabo akustičnega kalibratorja za izvajanje kalibracije. Test je bil opravljen

na lastni mobilni aplikaciji NoiSee [4] na mobilnem telefonu iPhone X. Aplikacija omogoča frekvenčno analizo s širino pasu 1 oktava, srednje frekvence oktavnih pasov pa se ratezajo od 31,5 Hz do 16 kHz. Za namene testiranja je bila mobilna aplikacija modificirana tako, da je omogočala branje oktavnega spektra preko omrežne povezave in protokola REST.

Za test oktavnih filtrov smo nadomestili zunanji mikrofon z električnim nadomestnim vezjem, ki ima enako impedanco kot merilni mikrofon in omogoča vzbujanje akustičnega vhoda mobilnega telefona z električnim testnim signalom. Celotna testna postavitev je prikazana na sliki 3.



Slika 3: Testna postavitev za testiranje oktavnih filtrov.

Za generiranje testnega signala je bil uporabljen generator Stanford DS360, ki generira sinusni signal z zelo nizkim popačenjem. Generator omogoča nadzor preko vmesnika GPIB, kar nam je omogočilo avtomatsko izvajanje testov. Osrednji del, ki je skrbel za avtomatizacijo testnega postopka, je bil realiziran v obliki skripte v jeziku Python. Naloga skripte je bila nastavljanje amplitude in frekvence signala na generatorju, branje oktavnega spektra iz mobilne aplikacije ter zapisovanje vrednosti v Excelovo preglednico.

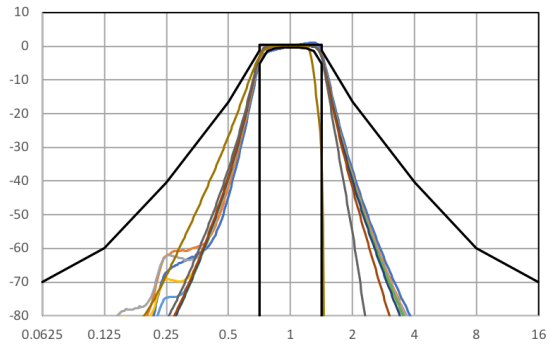
Za testiranje prevajalne funkcije filtrov, efektivne pasovne širine in seštevanje signalov je bil odziv vseh filtrov pomerjen z gostoto 48 točk na oktavo v frekvenčnem območju od 1 Hz do 24 kHz.

5 Rezultati

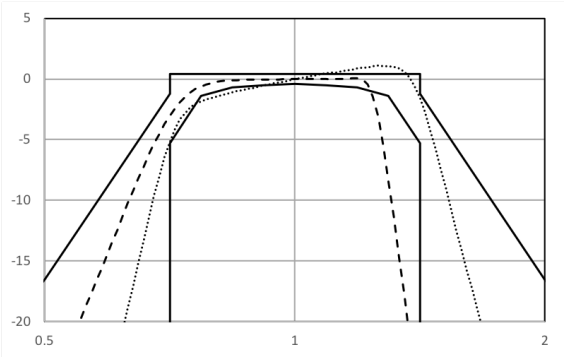
5.1 Prevajalna funkcija

Na sliki 4 so prikazani odzivi vseh filtrov, skupaj s tolerancami za Razred 1. Rezultati testa kažejo, da je odziv vseh oktavnih filtrov z izjemo prvega (srednja frekvenca 31,5 Hz) in zadnjega (srednja frekvence 16 kHz) skladen z zahtevami za Razred 1. Slika 5 prikazuje odstopanje prvega in zadnjega oktavnega pasu. Na vseh slikah predstavlja horizontalna os relativno frekvenco, ki je dejanska frekvenca, normalizirana za srednjo frekvenco vsakega izmed pasov.

Odstopanje prvega pasu (31,25 Hz) je posledica neidealnega analognega avdio vhoda na mobilnem telefonu, ki ni zasnovan za zajem tako nizkih frekvenc. Odstopanje zadnjega pasu (16 kHz) pa je artefakt digitalnega filtra. V delu, kjer odziv filtra ne sledi gabaritu, se namreč frekvenca vhodnega signala približuje Nyquistovi frekvenci.



Slika 4: Odziv vseh filtrov.



Slika 5: Odziva filtrov s srednjo frekvenco 31,25 Hz (črtkano) in 16 kHz (pikčasto).

5.2 Efektivna pasovna širina

Tabela 1 prikazuje rezultate efektivne pasovne širine. Skladnost z zahtevami Razreda 1 je izpolnjena pri štirih oktavnih pasovih, pri ostalih pa je izpolnjena skladnost z zahtevami Razreda 2.

Tabela 1: Rezultati testiranja efektivne pasovne širine.

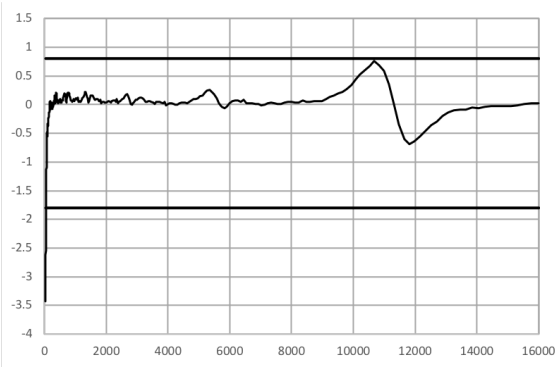
f [Hz]	ΔB **[dB]**	**f [Hz]**	ΔB **[dB]**
31.25	0,249	1000	0,421
62.5	0,281	2000	0,442
125	0,367	4000	0,461
250	0,419	8000	0,430
500	0,414	16000	-0,169

5.3 Seštevanje signalov

Napaka pri seštevanju signalov v odvisnosti od frekvence vhodnega signala je prikazana na sliki 6. Rezultati kažejo, da je v frekvenčnem območju od 31,25 Hz do 16 kHz napaka seštevanja znotraj predpisanih toleranc za Razred 1 (prikazane na sliki), razen pri najnižjem oktavnem pasu.

5.4 Linearnost

Rezultati testa linearnosti so prikazani v tabeli 2. Meje linearnega območja so določene glede na tolerance za Razred 1, spodnja (SM) in zgornja meja (ZM) linearnega območja pa sta izraženi v decibelih zvočnega tlaka.



Slika 6: Napaka seštevanja signalov v odvisnosti od frekvence.

Tabela 2: Rezultati testiranja linearnosti.

f [Hz]	ZM [dB]	SM [dB]	Lin. obm. [dB]
31.25	116,5	33,8	82,2
62.5	118,8	29,2	89,6
125	119,6	30,1	89,5
250	119,8	30,4	89,4
500	119,9	29,2	90,7
1000	119,9	28,2	91,7
2000	119,9	29,5	90,4
4000	119,8	30,5	89,3
8000	119,8	35,2	84,6
16000	119,8	37,2	82,6

Pri vseh oktavnih pasovih je izpolnjen kriterij za skladnost z Razredom 1 skladno z IEC 61260. Razlog za odstopanje pri najnižjem frekvenčnem pasu je isti kot pri prevajalni funkciji. Pri nizkih frekvencah je vhodni signal zadušen zaradi karakteristike avdio vhoda, zaradi česar je odziv oktavnega filtra prenizek. Negativno odstopanje v tem frekvenčnem območju je tako pričakovano.

6 Zaključek

Rezultati oktavnega analizatorja realiziranega v mobilni aplikaciji za merjenje hrupa so pokazali, da je skladnost z Razredom 1, kot je definiran v standardu IEC 61260, možna pri večini zahtev, pri vseh pa so izpolnjene zahteve za Razred 2. Odstopanje prevajalne funkcije pri nižjih frekvencah lahko pojasnimo z odzivom analognega dela avdio vhoda, ki pri tako nizkih frekvencah že duši vhodni signal. Najvišji oktavni pas pa se na zgornjem robu približa Nyquistovi frekvenci, kar popači njegov odziv do te mere, da ne ustreza več zahtevam.

Znanstvene raziskave v zadnjih letih so v veliki meri pokazale ustreznost mobilnih naprav za meritve hrupa in nove načine izvajanja meritev, kot je množičenje, vendar so se osredotočale le na posamezne parametre hrupa, ne pa na frekvenčno analizo, ki se praviloma uporablja pri meritvah hrupa v okolju in na delovnih mestih. Raziskava predstavlja prvi korak pri ugotavljanju skladnosti s standardnimi zahtevami za oktavno analizo.

Glavna pomanjkljivost raziskave je, da je bila omejena na eno mobilno napravo, zaradi česar rezultatov ne moremo z gotovostjo posplošiti na vse mobilne naprave. Nadaljnje raziskovalno delo bo temeljilo na testih več različnih mobilnih naprav in testu 1/3-oktavne analize.

Literatura

[1] Deafness and hearing loss. `http://www.who.int/mediacentre/factsheets/fs300/en/`. Accessed March 2018.

[2] iPhone, iPad, and iPod Benchmarks. `https://browser.geekbench.com/ios-benchmarks/`. Accessed July 2017.

[3] Noise and hearing loss prevention. `https://www.cdc.gov/niosh/topics/noise/`. Accessed March 2018.

[4] NoiSee. `http://www.ea-lab.eu/noisee/`. Accessed March 2018.

[5] Number of smartphone users worldwide from 2014 to 2020. `https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/`, June 2016. Accessed July 2017.

[6] I. Aspuru, I. García, K. Herranz, and A. Santander. CITI-SENSE: methods and tools for empowering citizens to observe acoustic comfort in outdoor public spaces. *Noise Mapping*, 3(1), jan 2016.

[7] P. Aumond, C. Lavandier, C. Ribeiro, E. G. Boix, K. Kambona, E. D'Hondt, and P. Delaitre. A study of the accuracy of mobile technology for measuring urban noise pollution in large scale participatory sensing campaigns. *Applied Acoustics*, 117:219–226, feb 2017.

[8] M. Celestina, J. Hrovat, and C. A. Kardous. Smartphone-based sound level measurement apps: Evaluation of compliance with international sound level meter standards. *Applied Acoustics*, 139:119–128, oct 2018.

[9] IEC 61260-1, Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications. Standard, International Electrotechical Commission, 2014.

[10] IEC 61672-1, Electroacoustics – Sound level meters, Part 1: Specifications. Standard, International Electrotechical Commission, 2013.

[11] C. A. Kardous and P. B. Shaw. Evaluation of smartphone sound measurement applications. *The Journal of the Acoustical Society of America*, 135(4):EL186–EL192, 2014.

[12] C. A. Kardous and P. B. Shaw. Evaluation of smartphone sound measurement applications (apps) using external microphones—a follow-up study. *The Journal of the Acoustical Society of America*, 140(4):EL327–EL333, oct 2016.

[13] E. Murphy and E. A. King. Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise. *Applied Acoustics*, 106:16–22, may 2016.

[14] B. Roberts, C. Kardous, and R. Neitzel. Improving the accuracy of smart devices to measure noise exposure. *Journal of Occupational and Environmental Hygiene*, 13(11):840–846, sep 2016.

Pomen natančnih meritev prostorskih modalnih frekvenc

Franc Policardi-Antoncich[1]

[1] *LUCAMI, Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1000 Ljubljana, Slovenija*

franc.policardi@fe.uni-lj.si

About the importance of exact room modal frequencies measurements

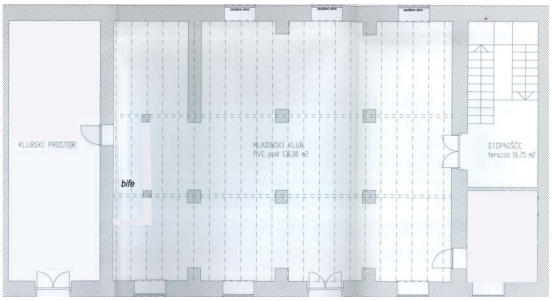
Sound propagation is theoretically a simple matter but room acoustics deals with reality complexity, multiple phenomena interconnection and mutual influence. Room acoustics is even today problematic because of theory and practice difficult matching, in most of the cases caused by difficulties in exact room modeling.
A good sound impression for listeners should take into account an occupied hall sound behaviour and in this paper we will focus on room standing modal frequencies a) from a theoretical point of view, b) measure real empty room behaviour c) measure real room with listeners and d) compare results, demonstrating the extreme importance of room exact modal frequency analisys before acoustic treatment.

1 Uvod

Propagacija zvoka je teoretično enostavna, se vendar prostorska akustika sooča s kompleksno realnostjo, s povezavo več fenomenov in z medsebojnimi vplivi. Prispevek predpostavi osnove prostorske akustike in se osredotoča na specifične negativne akustične pojave imenovane *resonance*. Prispevek obravnava le jakost posameznih frekvenc, ne pa tudi sprememb jakosti v daljšem časovnem obdobju.
Za doseganje dobrega zvočnega vtisa na poslušalce moramo upoštevati obnašanje zvoka v polni dvorani in v prispevku se bomo osredotočili na arhitektonsko analizo stavbe in dvorane, na prostorske stoječe modalne frekvence a) s teoretičnega vidika, b) izmerili bomo obnašanje prazne dvorane, c) ponovno bomo izmerili obnašanje dvorane s poslušalci in d) primerjali rezultate. Dokazali bomo pomen točnosti analize modalnih frekvenc obstoječega prostora pred akustično obdelavo.
V tem kontekstu smo uporabili dve orodji za akustične meritve in izračune, AcousticCalculator in RoomEQ Wizard. Meritve frekvenčnega odziva smo opravili v glasbenem klubu Ilirske Bistrice in natančno analizirali rezultate za nadaljnje sanacijsko delo.

2 Arhitektonsko stanje

Velikost stavbe znaša 20m x 12m x 11m; temelji, klet ter prvo nadstropje so sezidani iz kamna in vsi nosilni zidovi so debeli 80cm; stavba ima poleg zunanjih zidov tudi dodatne kamnite podporne stebre ter stene debele 60cm [1]. Med vsakim nadstropjem je strop, narejen iz lesenih gredi ter desk; poleg notranjih zidov dvorano v prvem nadstropju nosijo štiri po širini postavljene glavne gredi, ki so vzidane v podporne stebre v kleti. Na glavnih gredeh slonijo manjše gredi, postavljene po dolžini stavbe. Mladinski klub Nade Žagar [1] že več desetletij redno organizira glasbene dogodke in glede na današnja pričakovanja poslušalcev je izboljšava akustičnega obnašanja notranjega prostora nujna. MKNŽ je le rahlo preuredil notranjost kleti, kjer so danes koncertna dvorana za do 150 obiskovalcev, bife in klubski prostor.



Slika 1: Tloris kleti po zadnjih spremembah leta 1980

Stene kluba so do višine 1,5m oblečene v lesene deske in preostanek vidnih zidov je prepleskan. Les pripomore k boljši akustiki zaradi delne absorpcije in posledično zmanjšuje količino odbite energije, tako da energija odbojev hitreje pade pod mejo slišnosti in po naši meritvi v praznem prostoru RT_{60} znaša približno 1.5s. V polni dvorani se RT_{60} zmanjša na 0,8s, kar pripomore k dobri razumljivosti (STI).
Koncertna dvorana je pravokotne oblike mer 8 x 10m z višino 2.6m. Tla so betonska s 5mm debelo talno oblogo iz linoleja in strop je v celoti iz masivnega lesa: gredi velikosti približno 25cm x 25cm so postavljene po celotni dolžini hiše. Na gredeh je postavljen lesen pod, ki služi kot tla v zgornjem nadstropju. Tla so ravna in vzporedna s stropom, vendar zaradi oblike stropa ne pride do negativnih akustičnih pojavov. Akustično obnašanje lesenega stropa spominja na zelo velik difuzor, ki zvok razpršuje v različne smeri in hkrati preprečuje negativne akustične pojave med vzporednimi površinami.
Več kot 1.000 izvedenih koncertov in posebno arhitektonsko stanje dvorane sta avtorja spodbudila za natančno analizo modalnih resonančnih frekvenc (MRF), kajti meritve RT_{60} ter S/N se nista pokazali kot problematični.

3 Prostorske modalne frekvence

Znano je, da negativni akustični pojavi slabo vplivajo na razumljivost in natančnost prenašanja informaciji preko zvočnega signala. T.i. pojav *modal ringing* oz. modalno zvonjenje nastopi, ko se v prostoru pojavijo stoječi valovi modalnih resonančnih frekvenc (MRF) pri frekvencah nekaj pod 200Hz [2]. Glede na valovno dolžino in na določen prostor so ti ponekod močnejši in ponekod tišji, ker se med seboj algebraično seštevajo. Modalno zvonjenje se navadno pojavi v prostorih, kjer so glavne površine vzporedne in sicer pri frekvencah, pri katerih je celotna valovna dolžina, polovica, četrtina itn. enaka razmerju stranic prostora.

Izračun modalnih resonančnih frekvenc (MRF)

Za izračun MRF moramo izmeriti širino, dolžino in višino prostora, ker jih vedno računamo med dvema vzporednima površinama s pomočjo Rayleighove enačbe

$$f = \frac{c}{2L}, \quad (1)$$

pri čemer je *f* frekvenca, *c* hitrost zvoka in *L* dolžina stranice. MRF v glavnem delimo na aksialne, tangencialne in poševne [3]. Poleg osnovnih MRF izračunamo še višje harmonske frekvence, za katere smo uporabljali program AcousticCalculator [4].

V tabeli 1 so rezultati izračuna MRF do 200Hz pri aksialnih MRF po širini, dolžini in višini prostora označene z različnimi barvami z namenom prepoznavanja osnovne frekvence ter večkratnikov osnovne frekvence in frekvenc, ki se ponavljajo ter so zaradi tega bolj moteče. Označene frekvence imajo zelo velik vpliv na akustiko prostora in smo jih primerjali s frekvencami rezultat akustičnih meritev.

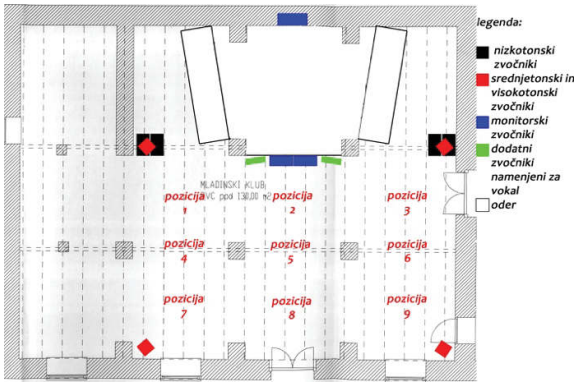
Tabela 1. Izračunane aksialne modalne frekvence do 200Hz

Aksialne			Tangencialne			Poševne
Po dolžini	Po širini	Po višini	Dolžina - širina	Dolžina - višina	Širina - višina	
17.02	21.27	6.54	27.24	18.23	22.25	28.01
34.03	42.54	13.09	45.81	18.23	24.97	30.22
51.05	63.81	19.63	66.04	21.47	28.94	46.28
68.06	85.08	26.18	86.76	25.98	33.73	40.66
85.08	106.34	32.72	40.13	31.22	43.04	54.87
102.09	127.62	39.27	54.47	34.65	44.51	42.21
119.11	148.88	45.81	72.31	36.46	46.85	47.65
136.12	170.15	52.35	91.63	39.29	49.95	56.02
153.14	191.43	58.9	55.3	42.93	64.14	
			66.45	51.46	65.13	
			81.71	53.69	66.76	
			99.21	57.37	68.97	
			71.31	68.37	85.33	
			80.26	69.31	86.08	

4 Meritev MRF v prazni dvorani

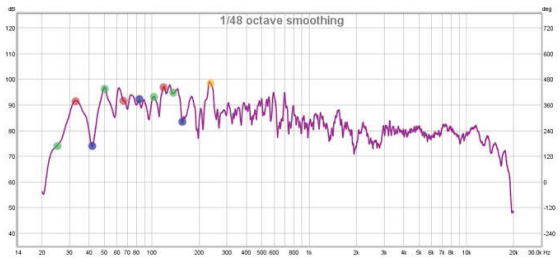
Meritve frekvenčnega odziva prostora smo opravili preko prenosnega računalnika, zvočne kartice Focusrite Scarlett 2i2, kondenzatorskega vsesmernega merilnega mikrofona DBX M2 RTA [5] in z računalniško programsko opremo Room EQ Wizard V5.1 [6], katera omogoča meritve RT do 10s in grafično prikazuje močnostno ter časovno obnašanje zvoka.

Zvočno kartico in merilni mikrofon smo natančno kalibrirali, saj REW V5.1 omogoča vključitev kalibracijske datoteke merilnih mikrofonov. Tloris koncertne dvorane smo razdelili na 9 pravokotnikov in opravili meritev z merilnim mikrofonom v 9 točkah, označenih na sliki 2.



Slika 2: Tloris koncertne dvorane z označenimi pozicijami meritev

Mikrofon je zajemal zvok na višini 1.6m, kar je povprečna višina človeškega ušesa. Opravili smo meritve celotnega človeškega slišnega frekvenčnega spektra s sinusnim preletom [7] med 20Hz in 20kHz pri jakosti 12dB FS dolžine 23.8s, 2-krat ponavljano. Graf na sliki 3 prikazuje primer "*distorzije*" v prostoru in omogoča točno določanje glasnosti posameznih frekvenc.



Slika 3: Primer: graf jakost vs. frekvenca

Pri izvajanju meritev smo ugotovili, da frekvenc nižjih od 30Hz naša elektroakustična oprema ne proizvede dovolj glasno; glasnost frekvenc nad 30Hz rahlo narašča in ko odziv nizkih frekvenc primerjamo z izračunanimi MRF ugotovimo, kakšen vpliv imajo resne MRF med koncerti. Največje razlike v glasnosti glede na merilne pozicije se nahajajo točno v območju izračunanih MRF. Odboji teh frekvenc se v nekaterih delih prostora seštevajo, v drugih odštevajo in fenomen se na grafu prikaže v obliki vrhov in dolin. Iz grafov je za vsako pozicijo razvidno, da največ problemov povzročajo nizke frekvence do 200Hz z ogromno razliko tudi do 35dB. Med 200Hz in 1kHz so razlike velike do 15dB; nad 1kHz človeški slušni sistem MRF težko razloči. Pri nizkih frekvencah in ne glede na pozicijo v prostoru opazimo dominantne MRF pri 51Hz, 63Hz, 70Hz, 85Hz, 100Hz, 136Hz, 150Hz, 170Hz, 190Hz in še okoli 245Hz.

Glede na izračunane in izmerjene MRF je razvidna razlika med teorijo akustike [2] in prakso [3]: prisotna je zelo močna MRF okoli 245Hz.

S tako ugotovitvijo smo še enkrat izračunali MRF do 300Hz in pridobili še 2 dominantni vrednosti pri aksialnih MRF, po širini na 245,24Hz in dolžini na 238,22Hz. S tem smo ugotovili, da MRF pri 119,11Hz,

katera ni izgledala pomembna, postane zelo izrazita pri svoji 1. harmonski frekvenci.

Tabela 2: Izračunane modalne frekvence do 300Hz.

Aksialne			Tangencialne			Poševne
Po dolžini	Po širini	Po višini	Dolžina - širina	Dolžina - višina	Širina - višina	
17.02	21.27	6.54	27.24	18.23	22.25	28.01
34.03	42.54	13.09	45.81	18.23	24.97	30.22
51.05	63.81	19.63	66.04	21.47	28.94	46.28
68.06	85.08	26.18	86.76	25.98	33.73	40.66
85.08	106.34	32.72	40.13	31.22	43.04	54.87
102.09	127.62	39.27	54.47	34.65	44.51	42.21
119.11	148.88	45.81	72.31	36.46	46.85	47.65
136.12	170.15	52.35	91.63	39.29	49.95	56.02
153.14	191.43	58.9	55.3	42.93	64.14	
238,22	245.24		66.45	51.46	65.13	
			81.71	53.69	66.76	
			99.21	57.37	68.97	
			71.31	68.37	85.33	
			80.26	69.31	86.08	

Sklep meritev v prazni dvorani:
1) različne pozicije v realnem prostoru prikazujejo različne rezultate glede jakosti in vedno potrdijo MRF;
2) po pregledovanju realnih grafov smo ugotovili, da se teorija [2] in praksa [3] v naši dvorani ne ujemata popolnoma. Še enkrat smo izračunali MRF in ugotovili, da je v našem primeru potrebno povečati Schröderjevo frekvenco vsaj do 300Hz;
4) na splošno je pri vseh meritvah jakost MRF v prazni dvorani velika, kar je bilo tudi pričakovano zaradi popolnoma vzporednih površin, vendar se poleg označenih frekvenc pojavijo tudi nekatere druge MRF pri 40Hz, 48Hz, 55Hz itd. do 250Hz;
5) velik problem v prazni dvorani predstavlja frekvenčni spekter med 60Hz in 200Hz, kjer je akustična energija prevelika za približno 5 do 10dB;
6) frekvenčni odziv pri vseh meritvah je dovolj linearen;
7) skoraj vse meritve potrdijo luknjo pri frekvenci 2kHz do 15dB;
8) visoke frekvence nad 800Hz bi lahko linearizirali preko izenačevalnika za dodatnih 8dB.

5 Meritev MRF v polni dvorani

Enake meritve smo opravili na 9 pozicijah tudi v polni dvorani, ker se glasbene oz. zvočne aktivnosti dogajajo v realni situaciji. Upoštevati moramo tudi, da se bodo glasbene vaje izvajale v prazni dvorani in zato je tako pomembno primerjati akustično obnašanje dvorane v prazni in v polni zasedenosti. Glede na veliko število grafov in podatkov bomo prikazali le sklep vseh meritev v polni dvorani:
1) jakost celotnega spektra je od prazne dvorane manjša za približno 10dB; na splošno je pri vseh meritvah jakost MRF v polni dvorani manjša, kar je bilo tudi pričakovano, vendar se poleg označenih MRF pojavijo tudi nekatere druge MRF pri 40Hz, 48Hz, 55Hz itd. do 250Hz;
2) velik problem v polni dvorani predstavlja frekvenčni spekter med 240Hz in 250Hz, kar se ujema z razširjenim teoretičnim izračunom MRF;
3) frekvenčni odziv je pri vseh meritvah v polni dvorani nekoliko bolj linearen;
4) različne pozicije v realnem prostoru prikazujejo različne rezultate glede jakosti in vedno potrdijo MRF;
5) po analizi realnih odzivov smo ugotovili, da se teorija [2] in praksa [3] tudi v polni dvorani ne ujemata popolnoma in tudi v tej situaciji smo preverili frekvence do 300Hz oz. nad teoretično Schröderjevo frekvenco do 200Hz preko (2)

$$2000 \times \sqrt{\frac{RT60}{V}}, \qquad (2)$$

6) problem v polni dvorani predstavlja frekvenčni spekter od 60 Hz do 90Hz, kjer je akustična energija višja za približno 5dB;
7) frekvenčni odziv je v polni dvorani pri skoraj vseh meritvah dovolj linearen, ker je akustični odziv absorpcije s strani publike jasno viden in globalno gledano je energija pod 200Hz skoraj naravno poravnana;
8) v polni dvorani se pri okoli 700 do 900Hz velikokrat pojavi vrh ali luknja za 6dB;
9) v polni dvorani skoraj vse meritve potrdijo luknjo na 2kHz tudi do -15dB;
10) v polni dvorani bi verjetno visoke frekvence nad 800Hz lahko linearizirali preko izenačevalnika za dodatnih 8dB.

6 Primerjava rezultatov v prazni in polni dvorani

Rezultate akustičnih meritev lahko ponazorimo na različne načine in zaradi lažje predstave smo v prejšnjih poglavjih označene problematične MRF ponazorili z različnimi intenzitetami barve: nad 90dB, med 80 in 89dB, med 70 in 79dB, med 60 in 69dB.
Prikazovanje rezultatov smo razdelili na 9 merilnih pozicij; te smo označili z jakostmi posameznih modalnih resonančnih frekvenc (Hz) v dB.

Tabela 3: Primerjava skale jakosti zvoka v dB in barve: jakost modalne frekvence 34 Hz na 9 različnih pozicijah

a) prazna dvorana			b) polna dvorana		
92dB	94dB	90dB	75dB	75dB	78dB
87dB	83dB	84dB	73dB	78dB	72dB
78dB	88dB	78dB	67dB	72dB	61dB

Tabela 3 prikazuje primerjavo med prazno in polno dvorano za 34Hz. Za vse druge glavne MRF smo uporabili isti sistem prikazovanja.
Jakosti MRF se v prazni in polni dvorani pri 34Hz, 68Hz in 136Hz bistveno razlikujejo na vsaki od 9-ih merilnih pozicijah.
Iz grafov in iz zadnje tabele je razvidno, da je višja koncentracija jakosti modalnih frekvenc v sprednjem delu dvorane, da je nižja v srednjem delu in kar zadovoljiva v zadnjem delu, še posebej v polni dvorani.

Primerjava po posameznih pozicijah nam pri frekvenci 34Hz predstavi naslednje razlike: pozicija 1 za -17dB, pozicija 2 za -19dB, pozicija 3 za -12dB, pozicija 4 za -14dB, pozicija 5 za -5dB, pozicija 6 za -12dB, pozicija 7za -21dB, pozicija 8 za -16dB in na poziciji 9 za -17dB.

Primerjava po posameznih pozicijah nam pri frekvenci 68Hz predstavi naslednje razlike: pozicija 1 za -9dB, pozicija 2 za -8dB, pozicija 3 za -14dB, pozicija 4 za -9dB, pozicija 5 za -18dB, pozicija 6 za -22dB, pozicija 7 za -7dB, pozicija 8 za -15dB in na poziciji 9 za -13dB.
Primerjava po posameznih pozicijah nam pri frekvenci 136Hz predstavi naslednje razlike: pozicija 1 za -12dB, pozicija 2 za -21dB, pozicija 3 za -35dB, pozicija 4 za -13dB, pozicija 5 za -20dB, pozicija 6 za -9dB, pozicija 7 za -11dB, pozicija 8 za -18dB in na poziciji 9 za -14dB.
Primerjava po posameznih pozicijah nam pri frekvenci 136Hz predstavi naslednje razlike: pozicija 1 za -12dB, pozicija 2 za -21dB, pozicija 3 za -35dB, pozicija 4 za -13dB, pozicija 5 za -20dB, pozicija 6 za -9dB, pozicija 7 za -11dB, pozicija 8 za -18dB in na poziciji 9 za -14dB.

7 Ugotovitve

V prispevku smo teoretično in praktično analizirali MRF določene dvorane namenjene glasbenih dogodkov.
Pravokotna oblika male dvorane večkrat povzroča modalne frekvence in negativne akustične pojave, ki jih brez konkretnih akustičnih rešitev težko rešimo.

S pomočjo računalniških orodij in realnih meritev smo v konkretnem prostoru spoznali naslednje ugotovitve:
1) jakost MRF je v polni dvorani bistveno manjša kot v prazni dvorani;
2) izračunane modalne frekvence smo potrdili preko dovolj natančnih meritev v različnih predoločenih pozicijah, najprej v prazni dvorani in po tem v polni;
3) na splošno je skoraj pri vseh meritvah jakost modalnih frekvenc v prazni in polni dvorani velika, kar je bilo tudi pričakovano, zaradi popolnoma vzporednih vseh 6-ih površin;
4) različne pozicije v realnem prostoru prikazujejo različne rezultate glede na jakosti posameznih frekvenc in da vedno potrdijo izračunane resonančne frekvence;
5) jakost celotnega slišnega spektra je v polni dvorani manjša za približno 10dB in frekvenčni odzivi so v polni dvorani bolj linearni, kar je posledica absorpcije akustične energije s strani publike;
6) pri skoraj vseh grafih je prisotna resonančna frekvenca okoli 240-250Hz, tako da smo še enkrat izračunali modalne frekvence in s tem ugotovili, da sta prisotni dve 1. harmonski komponenti osnovnih modalnih frekvenc 119,22Hz in 122,62Hz, kateri močno vplivata na nelinearnost frekvenčnega odziva;
7) ker sta 1. harmonski komponenti zgornjih modalnih frekvenc dovolj blizu smo ugotovili, da se v grafu pojavita kot vrh z “odprtim faktorjem Q”;
8) skoraj pri vseh meritvah v prazni in polni dvorani opazimo manjši vrh ali luknjo v frekvenčnem spektru od 700 do 900Hz;
9) skoraj pri vseh meritvah v prazni in polni dvorani opazimo luknjo pri frekvenci okoli 2kHz tudi do -15dB;
10) glede na rezultate akustičnih meritev bi verjetno visoke frekvence nad 800Hz lahko linearizirali s pomočjo izenačevalnika za dodatnih 8dB.

Na podlagi predstavljenih akustičnih osnov in dovolj natančnih akustičnih meritev smo ugotovili, da je kakovost akustike koncertnega prostora MKNŽ dovolj slaba in da je realizacija podanih rešitev nujno potrebna, a hkrati lahko trdimo, da je glede na veliko število koncertov v živo, ki smo jih obiskali v drugih prostorih, naša koncertna dvorana akustično in elektroakustično bolj primerna od nekaterih drugih slovenskih koncertnih prostorov.
S strani komentarjev med tonskimi tehniki, ki so v klubu delali, je prostor akustično stabilen in "*nima prevelikih problemov*", kar potrjuje, da se tonski tehniki ne zavedajo vseh možnosti izboljšanja akustičnega stanja koncertnih dvoran. Nastopajoči so nad zvokom glede na druge, podobne prostore navdušeni. Tudi tukaj z veliko verjetnostjo sklepamo, da glasbeniki niso preveč pozorni na to, kar predstavljajo poslušalcem.
Kot nadaljnje delo bomo projektirali izboljšave v MKNŽ.

Zahvala

Autor se zahvaljuje svojemu bivšemu študentu in danes dipl. ing. Multimedijskih Komunikacijah VSŠ g. Martinu Stadlerju za delo pri opravljanju meritev in elaboracija podatkov.

Literatura

[1] A. Brancelj, M. Filak, Ž. Gorišek, *Sokolski dom v Ilirski Bistrici - Konservatorski načrt* http://www.sneznik.net/wp-content/uploads/2012/06/SEMINARSKA-sokolski-dom-kon%C4%8Dna-verzija.pdf 11. 9. 2014

[2] M. R. Schröder, *New method of measuring Reverberation Time*, JASA, 37, pp 409-412, 1965

[3] A. Žnidarčič, F. Policardi, *Eksperimentalno preverjanje akustike bivalnega prostora ERK 2013*

[4] AcousticCalculator http://www.mh-audio.nl/user/acoustic%20calculator.asp

[5] Merilni mikrofon DBX model M2 RTA http://www.diyaudio.com/forums/equipment-tools/257292-dbx-rta-m2-condenser-mic.html, 15. 1. 2015

[6] Room EQ Wizard V5.1 http://www.roomeqwizard.com, 10. 1. 2015

[7] F. Policardi, MLS and *“Sine-Sweep technique comparison in room-acoustic measurement”,* Elektrotehniški Vestnik vol. n.78, n.3, pp 91-95, Ljubljana, 2011 ISSN 2232-3228

Metode merjenja odmevnega časa v prostoru

Rok Prislan

Univerza v Ljubljani, Fakulteta za matematiko in fiziko, Jadranska 19, 1000 Ljubljana, Slovenija
rok@prislan.net

Reverberation time measurement methods

Reverberation time is the central room acoustic parameter introduced by Mr. W. C. Sabine in 1885. As part of the acoustic design it is used to define the target acoustic properties of a room and is computationally predicted. Furthermore, reverberation time can be measured as presented in the paper. As such, it is important for the supervision of the acoustic design and the fit out works. Two main reverberation time measurement methods are presented, i.e. the interrupted noise method and integrated impulse response method. The measuring procedure and the required signal processing for both methods is briefly explained and schematically presented.

I. Uvod

Na področju akustike je relevantnih več tipov meritev, ki obsegajo relativno preproste meritve fizikalnih količin (npr. zvočnega tlaka [2] ali moči [3]), kot tudi meritve akustičnih parametrov in akustičnih odzivov. Naprednejše meritve zahtevajo sočasno meritev v več legah, med katerimi gre našteti meritev medušesne korelacijske funkcije (ang. *inter-aural cross correlation function* — IACF) [4], uporabo mikrofonskih polj [5] za lokalizacijo zvočnih virov, ali večmikrofonsko merilno metodo [6, 7], ki omogoča vizualizacijo modalnih oblik.

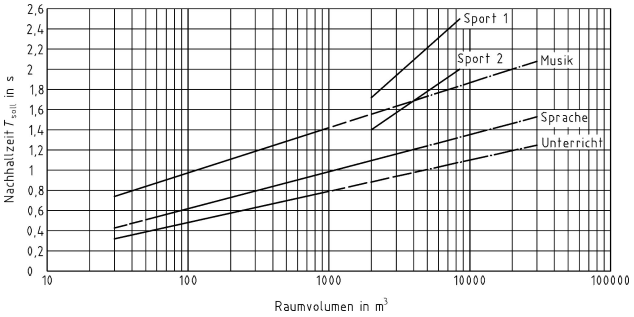
Meritev odmevnega časa natančno določa mednarodni merilni standard SIST EN ISO 3382-2 (A*kustika - Merjenje parametrov prostorske akustike - 2. del: Odmevni čas v običajnih prostorih (ISO 3382-2:2008)*). Ta podaja zahteve za uporabljeno merilno opremo (neusmerjen zvočni izvor, velikost mikrofonske diafragme), lege mikrofona in zvočnega izvora, za število oseb, prisotnih v prostoru, in uporabljene oktavne in terčne filtre (skladne s standardom IEC 61260). Pri tem lahko uporabimo različne merilne metode, ki se razlikujejo po potrebni obdelavi zajetih signalov in tudi točnosti izmerka. Najpogosteje uporabljeni metodi sta metoda prekinjenega šuma in metoda integriranega impulznega odziva, ki sta v nadaljevanju ločeno predstavljeni.

A. Odmevni čas

Odmevni čas T je čas, ki je potreben, da raven zvočnega tlaka v prostoru po izklopu zvočnega vira pade za $60\,\mathrm{dB}$. Vrednost odmevnega časa se v območju pod Schröderjevo frekvenco med različnimi točkami v prostoru močno spreminja, tako da praviloma kot odmevni čas razumemo njegovo prostorsko povprečje. Odmevni čas je povezan s subjektivno zaznavo **odmevnosti** prostora.

V luči Sabinove enačbe za odmevni čas, ta podaja razmerje med volumnom in absorpcijsko površino prostora. Tako ob večanju velikosti prostora in brez spreminjanja absorpcijskih lastnosti mejnih površin odmevni čas raste. Zato smo vajeni, da imajo večji prostori daljše odmevne čase.

Odmevni čas je tudi osrednji akustični parameter prostora, ki ga je že leta 1885 uvedel W. C. Sabine. Danes so izoblikovana priporočila za njegovo vrednost in kot primer podajam graf iz nemškega standarda DIN180154 [8] (glej sliko 1), po katerem ciljno vrednost odmevnega časa določimo glede na velikost in namembnost prostora[1].



Slika 1: Priporočene vrednosti odmevnega časa glede na namebnost prostora po standardu DIN180154 [8]. Na abscisi je volumen prostora, na ordinati pa vrednost odmevnega časa. Daljice povezujejo priporočeno vrednost in volumen za različne namembnosti: Sport (šport), Musik (glasba), Sprache (govor) in Uterricht (pouk).

II. Metoda prekinjenega šuma

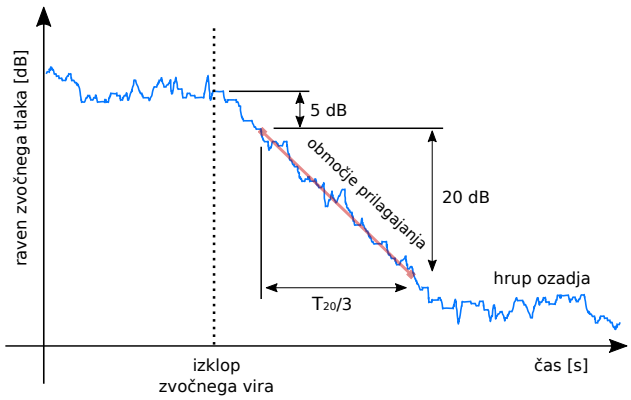
Kot navaja meritveni standard ISO3382 [4], lahko meritev odmevnega časa opravimo na podlagi dveh merilnih metod. **Metoda prekinjenega šuma** direktno sledi njegovi definiciji in je zato prikladna za njegovo

[1] Poleg same vrednosti odmevnega časa je pomembna tudi njegova frekvenčna odvisnost, ki se ravno tako spreminja z namembnostjo prostora.

ponazoritev. **Metoda integriranega impulznega odziva** pa odpravlja statistično odstopanje izmerkov, a zahteva obsežnejšo obdelavo signalov.

Izpostavljam, da obe merilni metodi sledita padanju ravni zvočnega tlaka, torej bi po definiciji odmevnega časa morali zajemati dinamično območje 60 dB. Običajno pa zajemamo manjše dinamične razpone 30, 20 in 10 dB, s čimer dobimo T_{20}, T_{30} in EDT (zgodnji čas padanja - ang. *early decay time*). Kljub obravnavi manjšega dinamičnega razpona, pa so vse količine podaljšane na čas, ki bi ga dobili ob opazovanju 60 dB območja.

Metoda prekinjenega šuma (ang. *interrupted noise method)* temelji na predvajanju šuma, ki ga prekinjamo. Pri tem merimo raven zvočnega tlaka, in sicer v frekvenčnih pasovih, če nas zanima frekvenčno odvisno obnašanje odmevnega časa. Ko se v prostoru vzpostavi stacionarna raven zvočnega tlaka, vir prekinemo, nato pa spremljamo padanje ravni, kot je shematsko prikazano na sliki 2. Iz te je razvidno, kako npr. določimo T_{20} kot tri-kratnik časa, v katerem pade glasnost za 20 dB (z -5 dB na -25 dB) glede na referenčno raven 0 dB, ki je vzpostavljena ob dlje trajajočem delovanju zvočnega vira.



Slika 2: Shemtaski prikaz principa merjenja odmevnega časa T_{20} z metodo prekinjenega šuma. Graf prikazuje izmerjeno raven zvočnega tlaka ob izklopu zvočnega vira in območje linearnega prilagajanja, s katerim določimo vrednost parametra T_{20}.

Predvajan signal pri merjenju z metodo prekinjenega šuma je najpogosteje naključni šum roza spektra. Ta namreč zagotavlja enakomerne zvočne ravni v vseh frekvenčnih pasovih, kar je pomembno za doseganje ugodnega razmerja med signalom in šumom. V praksi odmevni čas tudi opazujemo v posameznem frekvenčnem pasu, saj je praviloma frekvenčno odvisen.

III. Metoda integriranega impulznega odziva

Metodo integriranega impulznega odziva je leta 1965 predstavil nemški akustik Schröder [9], ki je izpostavil težave merilne metode prekinjenega šuma. Pri tej je namreč padanje ravni zvočnega tlaka zaradi uporabe naključnega signala odvisno od faze, s katero je vzbujen posamezni tlačni način v prostoru. Zato moramo meritev ponavljati in rezultate povprečiti, pri čemer je konvergenca še posebej počasna v območju nizkih frekvenc. Schröder je pri tem pokazal, da povprečenje vodi k odmevnemu času, ki ga dobimo z metodo integriranega impulznega odziva.

V smislu meritve moramo torej izmeriti impulzni odziv, z obdelavo katerega nato določimo odmevni čas.

A. Meritev odziva

Ustaljen postopek merjenja akustičnega odziva prostora temelji na spektrih korelacije [10, 11] predvajanega $x(t)$ in zajetega signala $y(t)$. Kot $x(t)$ preko zvočniškega sistema predvajamo poljuben signal, ki pa mora imeti zadostno energijo v celotnem frekvenčnem območju merjenja, v katerem se mora vzpostaviti zadostno razmerje med signalom in šumom oz. hrupom ozadja. Med predvajanjem zvočnega signala merimo zvočni tlak $y(t)$ z mikrofonom.

$y(t)$ je konvolucija predvajanega signala z impulznim odzivom prostora

$$y(t) = x(t) * h(t). \tag{1}$$

Konvolucijo v enačbi (1) v frekvenčnem prostoru nadomestimo z množenjem

$$Y(f) = X(f) \cdot H(f), \tag{2}$$

kjer so $Y(f)$, $X(f)$ in $H(f)$ Fourierove transformiranke signalov $y(t)$, $x(t)$ in $h(t)$. $H(f)$ je torej frekvenčni odziv prostora, ki bi ga lahko dobili na osnovi enačbe (2) kot

$$H(f) = \frac{Y(f)}{X(f)}. \tag{3}$$

To postopanje pa ni ustrezno, ko želimo meritev ponoviti, zajete signale povprečiti in s tem izboljšati razmerje med signalom in šumom. V praksi tako frekvenčni odziv izračunamo kot [11]

$$H(f) = \frac{S_{xy}(f)}{S_{xx}(f)}, \tag{4}$$

kjer je $S_{xy}(f)$ spekter korelacije signalov $x(t)$ in $y(t)$ (ang. *cross correlation spectral density*), $S_{xx}(f)$ pa spekter avtokorelacije predvajanega signala $x(t)$. Ob večkratnem zajemu signala $y(t)$ s povprečenjem spektra korelacije signalov $S_{xy}(f)$ torej izboljšamo razmerje signal/šum meritve.

Kot kontrolo kvalitete dobljenih rezultatov ob vsaki ponovitvi meritve izračunamo tudi koherenco [11]

$$C(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)}, \tag{5}$$

ki je pokazatelj nelinearnosti sistema ali prisotnosti šuma. Tako med meritvijo praviloma spremljamo, ali je vrednost koherence ves čas zadovoljivo blizu 1.

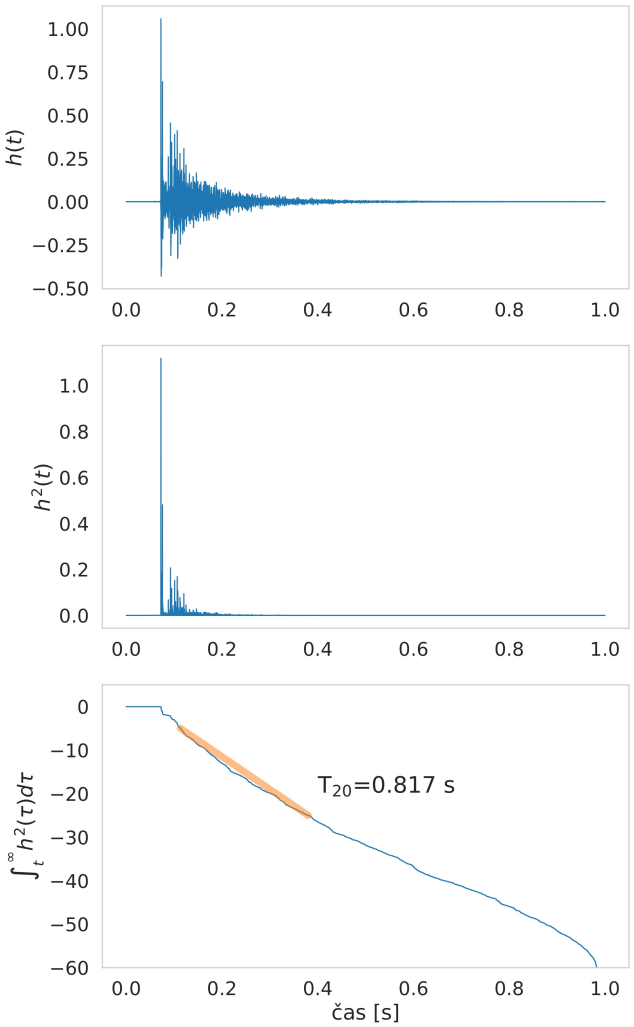
B. Določitev odmevnega časa

Shematski prikaz določanja odmevnega časa z metodo integriranega impulznega odziva je na sliki 3. Izmerjen impulzni odziv prostora $h(t)$ v prvem koraku filtriramo po frekvenčnih pasovih, če želimo opazovati frekvenčno

odvisnost odmevnega časa. Filtriran signal $h'(t)$ v naslednjem koraku kvadriramo, od koder dobim **krivuljo padanja ravni zvočnega tlaka**[2] z integracijo

$$\int_t^\infty h'^2(\tau)\mathrm{d}\tau. \tag{6}$$

V zadnjem koraku na izbranem dinamičnem območju krivulje padanja ravni zvočnega tlaka izvedemo prilagajanje linearne funkcije, iz koeficienta katere razberemo odmevni čas.



Slika 3: Določanje odmevnega časa po metodi integriranega impulznega odziva. Gornji graf prikazuje izmerjen impulzni odziv, srednji graf njegov kvadrat, spodnji graf pa krivuljo padanja ravni zvočnega tlaka. Na slednjem je z odebeljeno črto označeno optimalno linearno prilagajanje (med $-5\,\mathrm{dB}$ in $-25\,\mathrm{dB}$), na podlagi katerega določimo odmevni čas T_{20}. Za potrebe prikaza je uporabljen impulzni odziv, izmerjen v dvorani Kulturnega doma Bukovica decembra 2016.

IV. Zaključki

Odmevni čas je temeljni parameter prostora, na podlagi katerega praviloma določimo projektne/ciljne akustične lastnosti prostora. Odmevni čas tudi računsko ali modelsko napovemo in je torej vodilo pri izbiri materialov in obodnih površin prostora. Kot sem predstavil, lahko odmevni čas tudi merimo, tako da z njim spremljamo uspešnost akustičnega projektiranja oz. izvedbe.

Predstavil sem dve temeljni metodi merjenja odmevnega časa - metodo prekinjenega šuma in metodo integriranega impulznega odziva. Za obe metodi sem predstavil zahtevano obdelavo signalov in shematsko prikazal potek določitve odmevnega časa. Pri tem sem povzel nekatere pomembne razlike med obema metodama.

[2] V strokovni literaturi je krivulja padanja ravni zvočnega tlaka imenovana Schröderjeva krivulja padanja (ang. *Schröder decay curve*).

Literatura

[1] R. Prislan and D. Svenšek, "Room mode shape visualization using a multi-microphone measurement technique," in *Proceedings of the 7th AAAA Congress on Sound and Vibration*, 2016.

[2] ISO 1996-2:2017, *Acoustics – Description, measurement and assessment of environmental noise – Part 2: Determination of sound pressure levels*. International Organization for Standardization, Geneva, Switzerland, 2017.

[3] ISO 9614-1, *Acoustics – Determination of Sound Power Levels of Noise Sources Using Sound Intensity—Part 1: Measurement at Discrete Points*. International Organization for Standardization, Geneva, Switzerland, 1993.

[4] International Organization for Standardization, Geneva, Switzerland, "ISO 3382-1:2003 - Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces," 2003.

[5] R. Prislan and D. Fefer, "Principi lokalizacije zvočnih izvorov," in *Zbornik šestindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2012*, 2012.

[6] R. Prislan and D. Svenšek, "Room mode shape visualization using a multi-microphone measurement technique," in *Proceedings of the 7th AAAA Congress on Sound and Vibration*, 2016, pp. 125–130.

[7] ——, "Kako z meritvijo prepoznati modalne oblike v prostoru?" in *Zbornik šestindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2017*, 2017, pp. 146–150.

[8] DIN 180154:2004, *Hörsamkeit in kleinen bis mittelgroßen Räumen*. DIN Deutsches Institut für Normung, 2008.

[9] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am*, vol. 37, pp. 409–412, 1965.

[10] C. Sujatha, *Vibration And Acoustics*. McGraw-Hill Education (India), 2010.

[11] F. Jacobsen, *An Elementary Introduction to Applied Signal Analysis - note number 7001*. Technical University of Denmark, Acoustic Technology, Department of Electrical Engineering, 2010.

Validacija zvočne zaznavno-odzivne naloge (ZON) za ocenjevanje pozornostnih učinkov kognitivne obremenitve

Kristina Stojmenova[1] & Jaka Sodnik[1]

[1]Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana
E-pošta: kristina.stojmenova@fe.uni-lj.si, jaka.sodnik@fe.uni-lj.si

Abstract.
The Detection-response task (DRT) is a method for the assessment of the attentional effects of driver's cognitive load [1]. With this method the driver is presented with different stimuli in 3-5 s intervals and has to respond to the presented stimuli by pressing a button attached on his/her finger. Response times and hit rates (success of detection of stimuli) are considered as indicators of increased cognitive load. Since driving is primarily visual-manual task, we believe the use of an alternative stimuli modality could be more appropriate. Sound stimuli for example have very good properties as it is possible to influence the response times and attention allocation by manipulating the signal's frequency and/or intensity.
In this paper, we present a study, which validates a type of auditory signal to be used as DRT stimulus determined in our previous work, and evaluate the proposed auditory version of this method by comparing it with the standardized visual and tactile version.

1 Uvod

Metoda zaznavno-odzivne naloge (ZON) je standardizirana metoda za posredno merjenje kognitivne obremenitve voznika in njenega vpliva na voznikovo pozornost s pomočjo sekundarne naloge [1]. Temelji na zaznavanju dražljajev, ki se pojavijo na vsakih 3 do 5 sekund. Standard priporoča uporabo vizualnih ali taktilnih dražljajev, njihova oblika, intenziteta, postavitev in trajanje pa so zelo natančno definirani, saj je zaznavanje in odziv na dražljaje odvisen od vseh teh parametrov[1]. Metoda zahteva, da se voznik odzove na dražljaje s pritiskanjem gumba ob volan, najbolj pogosto pripetega na njegov/njen kazalec. Odzivni čas in uspešnost zaznavanja dražljajev sta definirana kot kazalnika povečane kognitivne obremenitve in njenih učinkov na voznikovo pozornost. *Odzivni čas* (ang. response time) je definiran kot čas od trenutka sprožitve dražljaja do trenutka, ko se voznik odzove na dražljaj s pritiskom na gumb ob volan. Kazalnik *uspešnost zaznavanja dražljajev* (ang. hit rate) je razmerje med številom uspešno zaznanih dražljajev in vseh predstavljenih dražljajev. Dražljaj je bil uspešno zaznan, v kolikor je bil odzivni čas voznika med 100 in 2500 ms po nastopu dražljaja, ostali dražljaji pa se štejejo kot zgrešeni.

Vožnja je primarno vizualno-manualna naloga, ki jo voznik v zelo veliki meri opravlja s pomočjo čutil vida in dotika. Komunikacija z večino sodobnih IK-sistemov prav tako poteka prek teh dveh čutil, saj so informacije v vozilu najpogosteje prikazane vizualno, voznik pa vnaša ukaze v informacijskih sistem ročno. Uporaba vizualnih ali taktilnih dražljajev oziroma dražljajev, ki jih voznik zaznava z enakimi čutili kot opravlja vožnjo in komunikacira z IK-sistemi med vožnjo, z vidika teorije o več virih ni najbolj primerna [2]. Ta namreč predpostavlja, da je količina pozornosti (sestavljena iz več pozornostnih virov glede na čutila za sprejem informacij), ki jo ima človek v danem trenutku na voljo, omejena. In čeprav je človek sposoben opravljati več nalog, ki zahtevajo isti pozornostni vir sočasno, je pri tem uspešen le do trenutka, ko pride do zasičenja uporabljenega pozornostnega vira. Posledično je uporaba dražljajev druge modalnosti, ki ne uporabljajo že zasedenih pozornostnih virov pri vožnji, lahko veliko primernejša: Tak primer je npr. zvočni dražljaj. Zvočni dražljaji imajo zelo dobre lastnosti, saj je s pravo izbiro frekvence in glasnosti možno vplivati na odzivnost voznika in prioriteto pri selektivni pozornosti (na primer, z izbiro alarmantnih zvokov).

2 Predhodne raziskave

Če prav so v standardu podane priporočene specifikacije le za vizualne in taktilne dražljaje, so bile narejene raziskave tudi z uporabo zvočnih dražljajev. Zvočna različica metode ZON je bila uporabljena v več študijah [3]-[5], vendar je bil v vseh študijah uporabljen drugačen zvok kot ZON dražljaj. Zato na podlagi preteklih študij ni možno narediti direktne primerjave občutljivosti zvočne metode s standardiziranimi različicami vizualne in taktilne ZON za zaznavanje učinke kognitivne obremenitve. Ker je odzivni čas odvisen od intenzitete in oblike signala, je zelo pomembno tudi točno določiti specifikacije zvočnega dražljaja metode ZON. Le ob uporabi enakega dražljaja v različnih študijah je možno delati primerjave med njimi, saj bodo spremembe v odzivnih časih posledica sprememb v kognitivni obremenitvi voznika in ne kot rezultat lastnosti zvoka, ki je uporabljen kot dražljaj.

Stojmenova in drugi so naredili študijo v kateri so želeli identificirati najbolj primeren zvoka za zvočno različico metode ZON [6]. Rezultati študije so pokazali 1-sekundne čiste tone s frekvenco 4 kHz in 8 kHz kot najbolj obetavne. Avtorji so vseeno predlagali uporabo 4 kHz signala kot zvočnega dražljaja za ZON, saj se zgornja meja frekvence sluha zmanjšuje s starostjo, kar

[1] Priporočene specifikacije so zaščitene s strani Mednarodne organizacije za standardizacijo (ISO) in njihova reprodukcija in javno objavljanje ni dovoljeno.

bi ob uporabi 8 kHz signala omejilo uporabo metode pri starejših voznikih.

Omenjena študija se je osredotočala izključno na primerjavo različnih zvočnih signalov in ni primerjala rezultatov neposredno z vizualno ali taktilno različico ZON. Tako je izbran signal pokazal najboljše rezultate le v primerjavi z drugimi zvočnimi signali, primerjava z vizualno in taktilno različico pa ni bila narejena.

S študijo, ki je predstavljena v tem prispevku, smo želeli izbrani signal primerjati tudi s standardiziranimi različicami. Primerjali smo vse tri različice metode ZON tako, da smo sledili priporočenim specifikacijam ISO standarda za izvajanje vizualne in taktilne ZON [1] ter da smo sledili rezultatom Stojmenove in drugih za izvajanje zvočne različice ZON [6]. Vse tri različice DRT so bile primerjane v istem preskusnem okolju in v istih pogojih vožnje.

3 Metodologija

Študija je bila izvedena v simulatorju vožnje *NERVteh* [7]. Simulator je sestavljen iz treh ukrivljenih 48-palčnih televizorjev, programske opreme *AV simulations*, dirkalnega sedeža z volanom *Fanatec* in tridelnega seta pedal za upravljanje vozila z avtomatskim menjalnikom (slika 1). Vozniki so vozili v voznem scenariju, ki je simuliral dvopasovno avtocesto (slika 2), omejitev je bila 130 km/h, veljala pa so prometna pravila za slovenske ceste.



Slika 1. NERVteh simulator vožnje.



Slika 2. Simulacijsko okolje – dvopasovna avtocesta.

V študiji je sodelovalo 24 udeležencev (12 žensk), M=30,42, SD=6,94. Vsi vozniki so imeli veljavno vozniško dovoljenje.

3.1 Naloge

Vsak udeleženec je opravil osem dvominutnih voženj, med katerimi je izvajal različne naloge: vožnjo, odgovarjanje na dražljaje ZON in izvajanje kognitivne naloge.

Primarna naloga udeležencev je bila varna vožnja. Pred začetkom študije kot tudi pred vsako novo vožnjo smo voznikom poudarili, da je njihova primarna naloga varna vožnja ter da imata nalogi ZON in n-nazaj sekundarno prioriteto.

Udeleženci so odgovarjali na vizualne, taktilne in zvočne ZON dražljaje. Vizualni in taktilni dražljaji so bili predvajani po priporočenih specifikacijah ISO standarda 17488:2016 [1]. Zvočni dražljaji so sledili priporočilom študije [6]. Za predvajanje vseh dražljajev ter beleženje odzivov nanje uporabili mikrokontrolno ploščo Arudio Mega in odprto kodo raziskovalne skupine Krause idr. [8].

Pri izbiri kognitivne naloge, ki simulira uporabo IK-sistema, smo izhajali iz objavljenih študij o evalvaciji vizualne in taktilne metode ZON. Pri tem je bilo možno zaslediti več različnih sekundarnih nalog, vsem pa je skupna možnost večstopenjske zahtevnosti. Ena izmed najbolj pogosto uporabljenih nalog je bila čista kognitivna naloga n-nazaj [9], zato smo se odločili tudi mi uporabiti to nalogo za simulacijo interakcije z IK sistemom in kontrolirano povečevanje kognitivne obremenitve.

Naloga *n-nazaj* (ang. n-back task ali Delayed digit recall task) je preprosta naloga s števili, kjer se vozniku vsakih 5 sekund zvočno predstavi naključno izbrano enomestno število, naloga voznika pa je ponoviti slišano število po vnaprej določenem zaporedju [9]. Naloga predvideva 4 težavnostne stopnje, kjer je 4. stopnja oziroma 3-nazaj najzahtevnejša stopnja. Prva oziroma najmanj zahtevna stopnja pa zahteva od voznika, da ponovi številko takoj, ko jo sliši (stopnja 0-nazaj). Pri naslednjih stopnjah si mora voznik zapomniti vedno več številk in jih ponoviti v točno določenem vrstnem redu. V predstavljeni študiji smo uporabili stopnjo 2-nazaj (primer je predstavljen v tabeli 1).

Tabela 1. Primer vrstnega reda številk pri nalogi n-nazaj, nivo zahtevnosti 2-nazaj.

	Zaporedje števk							
Predstavljeno število	0	7	8	4	9	1	7	8
Odgovor voznika	-	-	0	7	8	4	9	1

Nivo kognitivne obremenitve in vrsta dražljajev ZON sta bili definirani kot neodvisni spremenljivki. Kot odvisne spremenljivke smo opazovali odzivne čase ZON in uspešnost zaznavanja dražljajev in uspešnost izvajanja naloge n-nazaj.

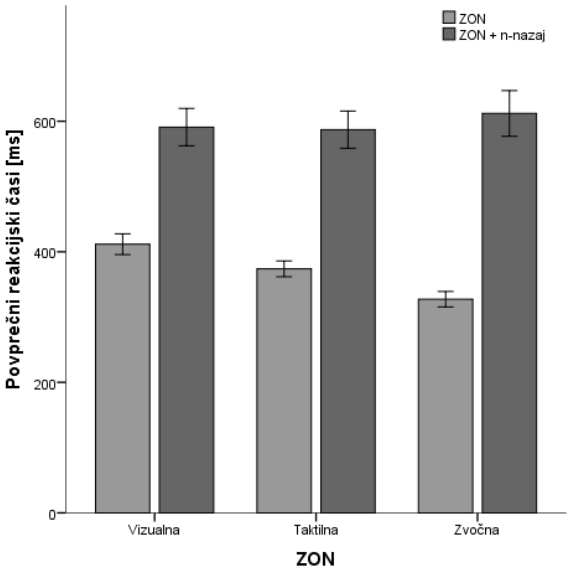
4 Rezultati

Za preverjanje normalnosti porazdelitve podatkov smo uporabili Shapiro-Wilkov test za testne vzorce z manj kot 2000 podatkov in Kolmogorov-Smirnov test za vzorce z več kot 2000 podatkov, nivo tveganja 5 %.

Za preverjanje homogenosti varianc smo uporabili Levenov test, nivo tveganja 5%. Glede na rezultate teh testov smo nato uporabili ustrezne parametrične in neparametrične teste.

4.1 Odzivni časi

Neparametrični Friedmanov test je pokazal statistično pomembne razlike v odzivnih časih na dražljaje ZON med vožnjami brez in s kognitivno nalogo n-nazaj, $\chi2(5) = 568{,}383$, $p<0{,}001$. Wilcoxonov post-hoc test vsote rangov z Bonferronijevo korekcijo ($p=0{,}017$) je pokazal, da so bili odzivni časi za vse tri oblike dražljajev (vizualni, taktilni in zvočni) statistično pomembno daljši pri vožnjah, kjer so udeleženci vozili, odgovarjali na ZON dražljaje in opravljali nalogo n-nazaj v primerjavi z vožnjami, kjer so samo vozili in odgovarjali na dražljaje ZON, $p<0{,}001$ (slika 3).

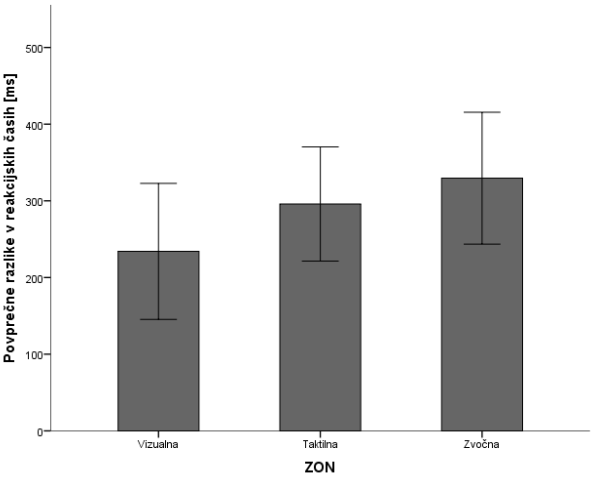


Slika 3. Odzivni časi na vizualne, taktilne in zvočne dražljaje pri vožnjah brez in s kognitivno nalogo n-nazaj.

Razlike v odzivnih časih med vožnjami brez in s kognitivno nalogo odražajo zahtevnost oziroma kognitivno obremenitev, ki jo naloga naloži vozniku. Večje kot so te razlike, bolj kognitivno zahtevna je naloga. Ker smo v študiji uporabili enako zahtevnost naloge v vseh vožnjah, razlike med odzivnimi časi na različne dražljaje ZON odražajo občutljivost posamezne različice za zaznavanje sprememb v kognitivni obremenitvi voznika.

Rezultati so pokazali, da so se odzivni časi v povprečju podaljšali za vizualni ZON za M=180,95 ms (29.68 %), za taktilni ZON M=213,32 ms (40.92 %), in za zvočni ZON M= 285.04 ms (46.24 %) pri vožnjah brez v primerjavi z vožnjami s kognitivno nalogo (slika 4).

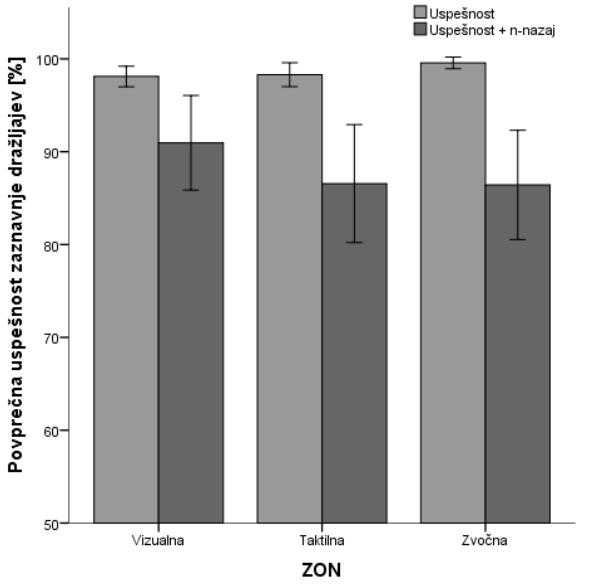
Wilcoxonov post-hoc test vsote rangov z Bonferonnijevo korekcijo ($p=0{,}017$) je pokazal, da je razlika v odzivnih časih za zvočno različico ZON statistično pomembno večja v primerjavi z razlikami v odzivnih časov za vizualno ($p<0{,}001$) in taktilno ($p=0.014$) različico.



Slika 4. Razlike v odzivnih časih na različne dražljaje ZON pri vožnjah brez in s kognitivno nalogo n-nazaj.

4.2 Uspešnost zaznavanje dražljajev

Neparametrični Friedmanov test je pokazal statistično pomembne razlike v uspešnosti zaznavanja dražljajev ZON med vožnjami brez in s kognitivno nalogo n-nazaj, $\chi2(5) = 33{,}709$, $p<0{,}001$. Wilcoxonov post-hoc test vsote rangov z Bonferronijevo korekcijo ($p=0{,}017$) je pokazal, da je bila uspešnost zaznavanje dražljajev statistično pomembno boljša pri vožnjah brez v primerjavi z vožnjami s kognitivno nalogo (slika 5).



Slika 5. Povprečna uspešnost zaznavanja dražljajev ZON pri vožnjah brez in s kognitivno nalogo n-nazaj.

Ob primerjavi uspešnosti zaznavanje dražljajev med različnimi modalnosti ZON (vizualni, taktilni in zvočni) ni bilo prisotnih statistično pomembnih razlik med različicami metode ($p>0{,}017$).

4.3 Uspešnost opravljanja dodatne kognitivne naloge n-nazaj

Uspešnost opravljanja naloge 2 nazaj (število točno ponovljenih številk) je bila za vizualno ZON M=92,8%,

taktilno ZON M=88,41% in zvočno ZON M=87,44% pri vožnjah z ZON in kognitivno nalogo. Fridmanov test je pokazal, da ni bilo statistično pomembnih razlik v uspešnosti opravljanja dodatne naloge 2-nazaj $\chi 2(3) = 4{,}984$, $p = 0{,}173$ med vožnjami brez in z ZON dražljaji.

5 Diskusija in zaključek

S predstavljeno študijo smo želeli preveriti ali je z zvočno različico ZON možno zaznavati spremembe v kognitivni obremenitvi voznika s podobno stopnjo občutljivosti kot z vizualno in taktilno različico ZON. V ta namen smo za vse različice uporabili enak nivo kognitivne obremenitve, razlike v reakcijskih časih in uspešnosti zaznavanja dražljajev pa interpretirali kot razlike v občutljivosti različic ZON na spremembe v kognitivni obremenitvi voznika. Rezultati študije so pokazali statistično pomembne razlike v reakcijskih časih za vse tri različice metode, kar pomeni, da je tudi z zvočno različico ZON možno zaznavati kognitivno obremenitev voznika na podoben način kot z vizualno in taktilno različico. Poleg tega so post-hoc analize pokazale, da so bile te razlike največje za zvočno ZON (slika 4). To lahko pomeni, da je testirana ZON v primerjavi s standardiziranima različicama tudi bolj občutljiva za zaznavanje kognitivne obremenitve voznika.

Rezultati o uspešnosti zaznavanja dražljajev so pokazali podobne razlike za vse tri metode, kar potrjuje, da je občutljivost zvočne metode primerljiva z vizualno in taktilno različico.

Visoka uspešnost opravljanja dodatne kognitivne naloge pomeni, da so udeleženci aktivno sodelovali v študiji in skušali uspešno opraviti nalogo n-nazaj. Dejstvo, da ni statistično pomembnih razlik v uspešnost opravljanja naloge med ponovitvami z različnimi dražljaji ZON, je dodatno potrdilo, da je bil nivo kognitivne obremenitve podoben za vsako različico ZON.

Na podlagi teh rezultatov lahko sklepamo, da je zvočna različica ZON, pri kateri se uporablja zvočni signal določen v [6], občutljiva za učinke kognitivne obremenitve na pozornost voznika in je za zvočno-govorne dodatne kognitivne naloge n-nazaj statistično pomembno boljša od vizualne in taktilne različice. Z nastankom nove različice metode, sta se področje uporabe in občutljivost ZON za zaznavanje kognitivne obremenitve tudi razširila in izboljšala.

Zahvala

Avtorja se zahvaljujeta Javni agenciji za raziskovalno dejavnost Republike Slovenije za sofinanciranje raziskave iz državnega proračuna.

Literatura

[1] ISO, 2016. Road vehicles -- Transport information and control systems -- Detection-Response Task (DRT) for assessing attentional effects of cognitive load in driving. *ISO*. 2016, Geneva, Switzerland. 17488:2016.

[2] Wickens, C. D., 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), pp.159-177.

[3] Merat, N., Jamson, A. and Leeds, U., 2007. Multisensory signal detection: How does driving and IVIS management affect performance? In *Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pp.351-357.

[4] Merat, N. and Jamson, A.H., 2008. The effect of stimulus modality on signal detection: Implications for assessing the safety of in-vehicle technology. *Human factors*, *50*(1), pp. 145-158.

[5] Stojmenova, K., Jakus, G. and Sodnik, J., 2017. Sensitivity evaluation of the visual, tactile, and auditory detection response task method while driving. *Traffic injury prevention*, *18*(4), pp.431-436.

[6] Stojmenova, K., Policardi, F. and Sodnik, J., 2018a. On the selection of stimulus for the auditory variant of the detection response task method for driving experiments. *Traffic injury prevention*, *19*(1), pp.23-27.

[7] NERVteh. Road Safety organization. Dostopno na: https://www.nerv-teh.com/

[8] Krause M., Conti A., Späth M., Bengler K., 2014. Testing Open-Source Implementations for Detection Response Tasks. In *Proceedings of the XV International Conference on Human Computer Interaction (Interacción '14)*. ACM, New York, NY, USA.

[9] Mehler, B., Reimer, B. and Dusek, J.A., 2011. MIT AgeLab delayed digit recall task (n-back). *Massachusetts Institute of Technology, Cambridge, MA*, pp.1-23.

Digitalno krmiljenje analognega studijskega izenačevalnika

Nejc Kirn[1], Franc Policardi-Antoncich[2]

[1] *DISTOPIK d.o.o., Tobačna ulica 5, SI-1000 Ljubljana, Slovenija*
[2] *LUCAMI, Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1000 Ljubljana, Slovenija*

nejc.kirn47@gmail.com - franc.policardi@fe.uni-lj.si

Audio industry has greatly developed in the last century and digital audio has been a standard for forty years already. In this article we demonstrate the process of creating a software program controlling a custom designed analog studio processor i.e. an equalizer. First we briefly overview why, despite this technology is decades old, it remains popular in recording studios and the problems that users face. Then we propose the digital control solution for the custom circuit boards. The finished system uses proprietary potentiometers and switches allowing users to both control the parameters by using the instrument's front panel, and to read, save and control the unit from a computer program. This article follows the ERK 2017 [1].

1 Uvod

Kljub relativni starosti tehnologije je analogna avdio oprema v zvočnih snemalnih studiih še vedno priljubljena. Glavni razlog za njeno uporabo je edinstven zvočni podpis, ki ga prinesejo fizične komponente. Čelne plošče s fizičnimi stikali, potenciometri, tipkami in ostalimi komponentami omogočajo direkten pristop k oblikovanju zvoka in so dodana vrednost pri uporabniški izkušnji. Klasična izvedba takih procesorjev ima tudi nekaj pomanjkljivosti, tako da uporaba digitalnega kontrolnega sistema za krmiljenje parametrov v signalni poti analognega avdio procesorja prinese veliko prednosti; možnost primerjave nastavitev in shranjevanje teh pomenijo za studijskega inženirja lažje, hitrejše in optimizirano delo ter kakovostnejši končni izdelek.
V prispevku bomo predstavili proces izdelave programske opreme, digitalnega krmiljenja in elektronskega vezja novega visokokakovostnega zvočnega izenačevalnika Distopik Precision Series Mastering EQ. Ostale podrobnosti so poslovne skrivnosti podjetja Distopik d. o. o. [2], kateremu se vnaprej zahvaljujemo.

2 Analogni studijski procesorji

Analogni studijski procesorji so, zaradi stranskih učinkov v elektronskih vezjih in elementih, še danes priljubljeni v kontekstu snemanja in obdelave zvoka. Dober primer so filtri za oblikovanje specifičnih frekvenc, t.i. *equalizers* (izenačevalniki), dinamični procesorji, osnovani na fotocelicah ali vakuumskih ceveh ter magnetni trak.

Prvi problem uporabe analognih procesorjev je njihov uporabniški vmesnik z nastavitvami, kjer so stikala, tipke, gumbe in ostale komponente, ki delujejo kot spremenljiv del signalne poti vezja. S spreminjanjem pozicije v vezje vstavljamo elektronske komponente kot so npr. upori in kondenzatorji in s tem spreminjamo vpliv vezja na zvok z vsemi "zaželenimi" distorzijami originalnega signala. Shranjevanje nastavitev na analognih napravah izvajamo ročno z uporabo slik oz. zapisovanja številk, na katere kažejo indikatorji čelne plošče. Tako shranjevanje nastavitev je dokaj težavno pri zveznih potenciometrih, zato se za zahtevnejše aplikacije kot je npr. mastering uporabljajo koračni potenciometri z diskretnimi nastavitvami, ki zvišajo zahtevnost izdelave in ceno.
Drugo problematiko predstavlja kalibracija sistema v celoti oz. zmanjšanje nezaželenih vplivov motečih signalov. Eden najpogostejših je električno napajanje, saj se iz tega velikokrat v sistem prenese šum okoli 50Hz in njegove harmonske frekvence do *fn* ter s tem poslabša S/N. To je značilnost analognega obratovanja; na ta šum moramo biti pozorni, ker za določene frekvence lahko preseže 30dB. Šum pripomore k intermodulacijski distorziji v vsaki opremi v verigi in lahko večkrat zaokroži skozi verigo, tako da postane, kljub vsem poskusom, praktično neizogiben. Le tri možnosti zmanjševanja šuma imamo na voljo: zelo kakovosten vir napajanja, visoko kakovostnen UPS sistem ali baterijsko napajanje. V primeru neupoštevanja tega dejavnika se bomo od začetka snemalne verige do končnih post-produkcijskih del vedno soočali s t.i. zmanjšanjem "jasnosti" izdelka. Tudi zaradi tega vzroka v digitalnih sistemih produkcije pogosto uporabljamo t.i. *plug-in* (*p-i*) oz. vtičnike, ki posnemajo rokovanje s čelno ploščo z nastavljivimi parametri, kateri vplivajo na delovanje določenih algoritmov. Za razliko od analogne opreme se nastavitve teh parametrov lahko shranijo v računalniškem spominu, na katerem se izvaja Digital Audio Workstation (DAW) in jih kasneje lahko dokaj enostavno prikličemo kot prednastavitev. *P-i*-i omogočajo tudi avtomatizacijo parametrov, ki so v programu zapisani kot dodatna sled parametra, katero je kasneje mogoče spremeniti. Dodatno se s takojšnjo uporabo *p-i*-ov izognemo motečim električnim šumov, ki jih omejimo na najkrajšo verigo mikrofon-kabel-predojačevalnik-A/D pretvornik. Te tri lastnosti se izkažejo za glavne razloge, zaradi katerih dandanes vse več poklicnih avdio inženirjev svoje delo opravlja preko *p-i*, saj jim ti prihranijo veliko truda in časa.

3 Digitalno krmiljenje izenačevalnika

Omogočanje digitalnega krmiljenja izenačevalnika je pomenilo zamenjavo določenih elementov iz signalne poti in njihovo krmljenje s pomočjo mikrokontrolerja. V izvirnih vezjih je signal potoval od glavne avdio plošče do potenciometrov na čelni plošči, tako da smo s spreminjanjem lege potenciometra v signalno pot spremenili vrednost upornosti in s tem sredinsko frekvenco, obliko in ojačenje. Čelno in avdio ploščo smo ločili in med njiju vstavili vezje z mikrokontrolerjem. Iz čelne plošče sedaj beremo nastavitve potenciometrov in te informacije uporabimo za preklop stikala na po meri izdelanem tiskanem vezju. Avdio signal tako potuje po enaki poti, le da imamo sedaj možnost digitalnega krmiljenja. Želimo tudi, da ima vezje z mikrokontrolerjem možnost omrežne povezave, kar bi nam omogočilo pošiljanje ukazov iz računalnika brez uporabe fizične prednje plošče. Idejna oblika vezja za digitalno krmiljenje algoritma za nastavljanje izenačevalnika poteka sledeče: a) pozicije potenciometrov oz. enkoderjev se berejo v ekspanderje, b) ekspanderji informacije spremenijo v format SPI in jih pošljejo mikrokontrolerju, c) mikrokontroler podatke obdela in jih pošlje na multiplekserje, d) multiplekserji izberejo ustrezen ekspander na glavni plošči, e) ekspander krmili releje in čipe ter tako nastavi filtre na želeno vrednost; kadar krmilimo izenačevalnik iz računalnika pa tako: a) ob spremembi parametrov na zaslonu *p-i* pošlje sporočilo preko mreže na prednastavljen IP-naslov in vrata v formatu OSC, b) mikrokontroler to sporočilo prevede v ukaz, ki ga multiplekserji in ekspanderji razumejo, c) naprej je postopek enak, kot od zgornje 4. točke.

Za obliko filtra smo uporabili koračne potenciometre Lorlin z 12-imi koraki, medtem ko smo za frekvence uporabili 24-koračne, od katerih smo porabili le 22 korakov. Pini so vezani na ekspanderje, ki trenutne nastavitve ob ustreznem proženju sestavijo v dva bajta podatkov in jih pošljejo mikrokontrolerju.

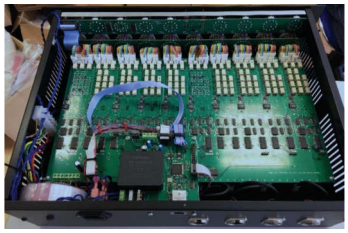
Za nastavljanje ojačenja smo uporabili pretvornik LTC1596-1 D/A oz. MDAC; ta nam omogoča, da s 16-bitno natančnostjo nastavljamo izhod med 0V in referenčno napetostjo, v tem primeru našim avdio signalom. Ker je enkoder le 8-biten, se s čelne plošče ne da izkoristiti polne ločljivosti; to nam dopušča *p-i* oz. krmiljenje iz računalnika.

Mikrokontroler se z računalniškim *p-i*-om sporazumeva preko kabla ethernet in protokola OSC, tako se vsaka sprememba na čelni plošči ustrezno odraža na uporabniškem vmesniku *p-i*-a. Obratno to ne deluje, saj bi v tem primeru morali biti potenciometri in stikala motorizirani. Spremembe se v *p-i*-u vseeno odražajo v dejanskih nastavitvah na kontrolni plošči oz. v relejih.

Elektronska vezja

Naštetim modelom pripadajo posamezna fizična tiskana vezja. Kakor je prikazano na sliki 1, so v ohišju izenačevalnika štiri vezja:

- analogni del na dnu ohišja
- vezje z mikrokontrolerjem
- glavna plošča, ki je povezana z analognim delom
- čelna plošča



Slika 1: Od vrha slike navzdol: čelna plošča, glavna plošča in vezje z mikrokontrolerjem

Pomembnejše komponente na vezju z mikrokontrolerjem so sam mikrokontroler, stikalni napajalnik in pri ključek za ethernet. Preko tega vezja se izvaja komunikacija z protokoloma OSC za računalnik in SPI za vse ostale komponente, kot so ekspanderji in multiplekserji ter preko njih tudi z releji (bele kvadraste komponente na sliki 1), stikali, statusnimi LED diodami itn..

Na glavni plošči je večji del kontrolnega vezja: štiri multiplekserji in za vsak filter po dva 16-bitna in en 8-bitni ekspander, en MDAC ter do 36 relejev. Z njimi nastavljamo sredinsko frekvenco in kakovost zvonastih filtrov, jih vklapljamo in menjamo med pozitivnim in negativnim ojačenjem na MDAC. Ta plošča je na vezje z mikrokontrolerjem povezana z oklopljenim kablom. Čelna plošča je sestavljena iz treh tiskanih vezij, ki so med sabo povezana z tračnim kablom. Na vsaki od treh najdemo dva 16-bitna ekspanderja in ustrezne povezave za komponente, kot so stikala, LED diode, potenciometre itn., s katerimi uporabnik nastavlja izenačevalnik.

4 Izdelava programske opreme

Orodja

Pri programiranju krmilnega programa izenačevalnika smo uporabljali naslednja orodja: 1) programator ST-LINK/V2 za mikrokontrolerje STM8 in STM32; 2) Digilent Analog Discovery za merjenje, analizo, vizualizacijo in snemanj dogajanja na vezjih; skupaj s 3) programsko opremo Digilent WaveForms nam omogoča preverjanje delovanja SPI komunikacije, ker vsebuje dvokanalni osciloskop, dvokanalni funkcijski generator, 16 digitalnih vhodov in izhodov ter dva izhoda za napajanje.

Programska oprema

Za programiranje mikrokontrolerja STM32 v jeziku C++ smo uporabili okolje Eclipse.

Za razvoj *p-i*-a, ki skrbi za komunikacijo med računalnikom in mikroprocesorjem v izenačevalniku smo uporabili Microsoft Visual Studio 2013. *P-i* je programiran v jeziku C++ in uporablja knjižnice JUCE za razvoj namiznih in mobilnih aplikacij v namen avdia. Microsoft Visual Studio 2013 vsebuje metode, razrede itn., ki so optimizirani za izvajanje v realnem času in take, ki olajšajo izdelavo uporabniškega vmesnika. Glavni namen tega skupka knjižnic je, da olajša razvoj avdio aplikacij, saj ista izvorna koda lahko teče na različnih namiznih in mobilnih operacijskih sistemih.

Za merjenje delovanja izenačevalnika smo uporabili program Reaper DAW, ki nudi snemanje več kanalov zvoka in/ali MIDI naenkrat ter program Fuzzmeasure, s katerim smo kalibrirali izenačevalnik: ta nudi natančnej

ši prikaz rezultatov, ki je zaradi uporabe sinusnega preleta namesto belega šuma vseeno počasnejši in zato za zagotavljanje kakovosti pri izenačevalniku s tako velikim številom parametrov manj uporaben.
Zvočna kartica Audient iD4 in programska oprema iD Audient sta nam omogočala predvajanje in zajemanje avdio signala med računalnikom in izenačevalnikom.

Modeli

Ključnega pomena za uspešen razvoj kode so t.i. "*modeli*" izenačevalnika, čelne plošče in *p-i*-a, pri osnovanju katerih smo se poslužili tehnike »*domain driven design*«-u, s katero smo morali, za uspešno reševanje neke problematike z računalniškim programom, poznati njeno domeno in poznati vrhnjo arhitekturo, delovanje in dinamiko znotraj naše domene. Dobro načrtovan sistem je zmanjšal število konceptualnih razlik med delovanjem domene v praksi in delovanjem abstraktnega programa, ki to domeno opisuje ter hkrati zmanjšal število kasnejših zahtev za spremembe in popravila. Glavni pomembni modeli so trije: a) model izenačevalnika, ki predstavlja stanje na glavni avdio plošči, b) model čelne plošče, ki predstavlja stanje nastavitev na čelni plošči in c) model *p-i*-a, ki predstavlja nastavitve na računalniškem uporabniškem vmesniku. Razdelili smo jih zato, ker moramo omogočiti smiselno delovanje pri kombiniranem upravljanju s čelno ploščo oz. *p-i*-om. Zahteve so bile: a) sprememba na čelni plošči se mora poznati tako na zvoku kot tudi na *p-i*-u, b) sprememba na *p-i*-u se mora poznati na zvoku, pri čemer se trenutno stanje na čelni plošči ignorira in c) če uporabnik odpre projekt, v katerem teče *p-i*, kjer so zapisane nastavitve parametrov, se le-te ne smejo prepisati s tistimi, ki so trenutno na čelni plošči razen, če prepis nastavitev uporabnik sam sproži preko za to namenjene tipke.
Z uporabo treh modelov lahko vsem trem zahtevam zadostimo dokaj enostavno, ker v vsakem posebej hranimo nastavitve (glavne avdio plošče oz. čelne plošče oz. *p-i*-a): dejansko stanje na relejih in MDAC-ih oz. nastavitve, ki jih uporabnik sliši, so vedno shranjene v modelu izenačevalnika. Da modelu glavne plošče lahko posredujemo ustrezno informacijo, moramo v modelih čelne plošče in *p-i*-a hraniti ločene spremenljivke za vse nastavitve. Metoda, ki nam je omogočila hiter prepis parametrov v *p-i*-u z nastavitvami fizične čelne plošče je invalidFrontPanel. Kadar spreminjamo nastavitve v *p-i*-u izgubimo sinhronizacijo med čelno ploščo in *p-i*-om. Zaradi velikega števila parametrov bi bilo ročno nastavljanje za uporabnika zamudno in zato smo dodali tipko Load From Panel, ki naenkrat prekopira vse vrednosti iz modela čelne plošče v modelu izenačevalnika in *p-i*-a. S tem smo s pritiskom ene tipke omogočili ponastavitev približno 50 parametrov.

Osnovna logika programa

Določene dele kode, ki so se uporabljali pri vseh modelih, je bilo smiselno zapisati v svojo knjižnico. V njej smo med drugim definirali metode in razrede, ki so namenjeni neposrednemu pisanju in branju pinov ekspanderjev na čelni oz. glavni plošči. Prvi pomemben del so metode, ki preberejo izbran pin oz. bit iz ekspanderja v primeru čelne plošče oz. zapišejo ustrezno vrednost v ekspander, ki krmili releje v primeru glavne plošče. Metodo postavi n-ti bit na 1 in ostale na 0, sestavili smo jo iz osnovnejših, ki so na primer postavile n-ti bit na 0 oz. 1.

Glavna plošča

Na glavni plošči se nahajajo multiplekserji, ekspanderji, releji in MDAC-i, ki vplivajo na nastavitve frekvenc, oblike itn. filtrov izenačevalnika.
Ko po SPI pošljemo vrednost enega parametra filtra, se ta vrednost pošlje na vse filtre, kar ni zaželeno. Zato smo morali poskrbeti, da le določeni ekspanderji dobijo ta podatek in to smo izvedli z metodo selectBlock, kateri na podlagi podanega parametra »i« ugotovi kombinacijo vhodov ABCD na multiplekserjih, ki ustreza na i-temu filtru; nato ekspanderjem na tem filtru pošlje signal Chip Select.
InitEQ in initBlock sta metodi, ki ob zagonu programa oz. vklopu izenačevalnika inicializirata ekspanderje na glavni plošči. Ob klicu initEQ se sproži for zanka, ki eno za drugim naredi inicializacijo na vseh filtrih.

Čelna plošča

Čelna plošča je sestavljena iz treh posameznih tiskanih vezij, ki so med sabo povezane z tračnim kablom. Na plošče so pritrjeni koračni potenciometri, enkoderji, stikala in LED diode, ki služijo kot fizični uporabniški vmesnik. Iz vseh teh komponent, nameščeni na vseh treh ploščah, ekspanderji berejo nastavitve in jih posredujejo mikrokontrolerju. Vsak ekspander bere en tip parametra z izjemo frekvence, kjer je zaradi 22 možnih nastavitev treba uporabiti dva 16-bitna ekspanderja.
V razredu FrontPanelExpanders najdemo metodo initExpanders, ki sama kliče metodo initInput oz. initOutput iz prejšnje knjižnice CommonIO.hpp. Z njeno pomočjo se ekspanderji pravilno inicializirajo in pripravijo na delovanje. Naslednja metoda je scanFrontPanel, kateri podamo številko od 0-9. Glede na podano številko se najprej izbere ustrezen filter in nato preberejo nastavitve vseh potenciometrov in stikal na čelni plošči. Za branje nastavitev smo uporabili metode *write* in *read*, ki so definirane v CommonIO.hpp. S tema dvema metodama zagotovimo povezavo med knjižnico FrontPanelIO.hpp, ki se ukvarja z domeno na čelni plošči in CommonIO.hpp, ki se ukvarja z delovanjem samih čipov na nižjem nivoju. Koda postane s tem bolj berljiva.
FPLowFilter, FPMidFilter in FPHighFilter se ukvarjajo z nastavitvijo oblik posebej nizkih, srednjih in visokih filtrov, saj imajo različno funkcionalnost. Nizki filter mora pri določeni nastavitvi iz vezja izklopiti potenciometer in vklopiti ločen del vezja, ker s tem omogoči preklop iz filtra oblike zvonca v obliko police.
Za nastavljanje ojačenja smo uporabili enkoder Bourns ACE 128. Na osmih podatkovnih pinih generira število, ki ga moramo preko tabele pretvoriti tako, da pri vrtenju v smeri urinega kazalca dobimo števila od 0 do 127. Programsko vrednosti pretvorimo v število s plavajočo vejico z razponom od 0 do 1, ki ga uporabimo za shranjevanje vrednosti ojačenja, saj ga prek 16-bitnega MDAC lahko nastavljamo skoraj zvezno.

***P-i* in komunikacija**

Za komunikacijo med računalnikom in izenačevalnikom uporabimo protokol OSC, razvit za upravljanje in

komunikacijo med računalniki, sintetizatorji zvoka, aplikacijami itd. preko mreže za namen glasbenega nastopanja. Protokol OSC ima podoben namen kot protokol MIDI, vendar omogoča bolj podrobno, razširjeno in brezžično komunikacijo.
Za namen komunikacije *p-i*-a in izenačevalnika smo opisali metode modela izenačevalnika in *p-i*-a, ki sestavljajo in pošiljajo OSC sporočila na ustrezna vrata na mreži. OSC sporočilo je sestavljeno iz naslova, tipa argumenta in argumenta [3]. Naslov nam pove, kateri parameter izenačevalnika spreminjamo, tip argumenta pri nas loči med 32-bitnim integer-jem in 32-bitnim float-om, argument pa je nastavitev parametra. Pri OSC moramo biti pozorni na število znakov; na konec naslova, tipa in argumenta moramo dodati toliko ničel, da je število znakov v OSC sporočilu deljivo s 4 in se konča z 0. Napisali smo avtomat stanj tako, da je izenačevalnik uporaben tudi brez priklopa na mrežo; stanja so Disconnected, Initialized, NoEthernet, ConnectedPending in Connected. Del komunikacije smo spisali tudi v kodi, ki opisuje delovanje *p-i*-a. Ta nam ponuja grafični uporabniški vmesnik (Slika 4), s katerim lahko nastavitve izenačevalnika spreminjamo iz DAW.

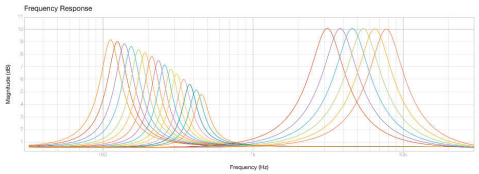


Slika 4: Uporabniški vmesnik plug-in-a

Skupina knjižnic JUCE, po definiranju imen parametrov in njihove povezave na ustrezne elemente v grafičnem vmesniku, sama poskrbi za delovanje. Dodali smo še metode za pošiljanje in sprejemanje OSC sporočil ter tipke Load From Panel, Preset A oz. B in Load Preset. Prva vzame nastavitve iz čelne plošče in povozi trenutno nastavljene v *p-i*-u, zadnja deluje obratno. Tipki A in B omogočata priklic dveh prednastavitev za hitro in enostavno primerjavo. *P-i* ima tudi gumb za ojačenje signala za +-24dB.

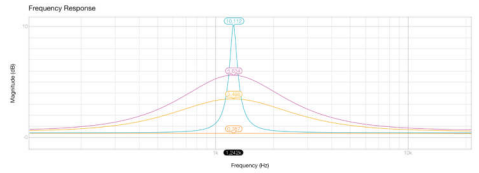
5 Preizkus delovanja

Verifikacija krmiljenja parametrov je potekala z uporabo več orodij. Prvi je bil logični analizator Analog Discovery, s katerim smo že ob programiranju knjižnic MainboardIO.hpp, MEQmain.hpp, FrontPanelIO.hpp in CommonIO.hpp sproti preverjali delovanje SPI komunikacije. Z uporabo programa Reaper in Fuzzmeasure smo preverili, ali se z izbiranjem številk napisanih na čelni plošči filtri res nastavijo na ustrezno sredinsko frekvenco. Iz prvega smo preko zvočne kartice v izenačevalnik pošiljali beli šum in ga nato analizirali s *p-i* FFT analizatorjem; preverili smo kombinacije vseh nastavitev frekvenc in kakovost filtrov tako pri pozitivnem kot pri negativnem ojačenju. Naleteli smo na nekaj nepravilnosti v kodi kot tudi v strojni opremi. Primera takih napak sta bila metoda za izbiro frekvence, ki je le-te izbirala v obratnem vrstnem redu in ni pravilno upoštevala vezav na čelni in glavni plošči, zaradi česar sta bila ob izboru frekvence 1 in 22 dejansko izbrani 2 in 1. Naleteli smo tudi na pregoreli upor na enem izmed kanalov, ki je onemogočal pravilno delovanje uporovne lestve za izbiro frekvenc (Slika 5).



Slika 5: Pokvarjen upor povzroči preskok frekvenc

Izkazalo se je tudi, da ob izbiri +3dB ojačenja na čelni plošči le-tega nismo dobili na zvoku in da niti krivulja ojačenja ni bila pravilna. Za te meritve smo uporabili program Fuzzmeasure, ki namesto belega šuma uporablja sinusni prelet. S tem smo si zagotovili bolj natančne meritve za iskanje aproksimacijske funkcije. To funkcijo smo ustrezno vstavili v kodo in s tem poskrbeli, da vse številske nastavitve ojačenja na čelni plošči ustrezajo ojačenju, ki vpliva na avdio (Slika 6).



Slika 6: Meritev ojačenja

6 Sklepne ugotovitve

V prispevku smo predstavili izdelavo programske opreme, digitalno krmiljenje in elektronsko vezje novega zvočnega izenačevalnika. Tekom našega dela smo razdelili in postopoma rešili analogne izzive povezane z električnim napajanjem, potenciometri in stikali, vezjem z mikrokontrolerjem, programiranjem na različnih nivojih, komunikacijo med komponentami kot npr. glavno in čelno ploščo in z digitalnim shranjevanjem nastavitev in računalniškim krmiljenjem parametrov. Naš končni izenačevalnik omogoča klasično ročno nastavljanje parametrov preko potenciometrov in stikal do čelne plošče in istočasno branje, shranjevanje ter spreminjanje vseh parametrov preko računalniškega programa. Naprava sama poskrbi, da so parametri na pravilnih vrednostih; takoj, ko uporabnik odpre projekt v DAW lahko nadaljuje z delom in s tem prihrani čas.
Menimo, da bo sčasoma na trgu prišlo do večje razširjenosti digitalnega krmiljenja procesorjev, kar bo znižalo cene in povečalo dostopnost.

Literatura

[1] N. Kirn, F. Policardi Antoncich, *Optimizacija uporabniške izkušnje in krmilnih sistemov v modernih studijskih avdio procesorjih*, ERK 2017;

[2] mix:analog [Online]. Dosegljivo mixanalog.com. [Dostopano: 9.1.2018];

[3] The Open Sound Control Specification [Online]. Dosegljivo: opensoundcontrol.org/spec- 1_0. [Dostopano: 9.1.2018].

Računalništvo in informatika

Computer and Information Science

Vizualizacija trkov osnovnih delcev v fiziki visokih energij na spletu

Sebastien Strban, Ciril Bohak, Matija Marolt

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko
E-pošta: ss4774@student.uni-lj.si, {ciril.bohak, matija.marolt}@fri.uni-lj.si

Web-based visualization of high-energy physics particle collisions

In this paper we present a visualization system for particle collisions in high-energy physics. During the experiments in particle physics a large amount of data is acquired, used for event reconstruction. For better understanding suitable visualization is needed. This poses a big challenge due to large amount of data, especially when we want to achieve real-time interactive functionality. For wider accessibility we decided to implement the visualization in web-based technologies. We developed a visualization system using WebGL, optimized for fast and efficient use with GPUs. We focused on minimization of draw calls and minimization of data transfer to the the GPU. This proofs as an efficient solution confirmed by performance analysis of the developed system.

Povzetek

V članku predstavljamo sistem za vizualizacijo trkov delcev v fiziki visokih energij. Pri eksperimentih v fiziki osnovnih delcev raziskovalci zajamejo ogromno količino podatkov, na podlagi katerih poskušajo rekonstruirati dogajanje. Za boljše razumevanje dogajanja je potrebna ustrezna vizualizacija, kar v primeru velike količine podatkov predstavlja svojevrstni izziv. Še posebej v primeru, ko želimo takšno količino podatkov vizualizirati v realnem času in z visoko stopnjo interaktivnosti. Zaradi želje po čim širši dostopnosti smo se odločili za implementacijo v spletnih tehnologijah. V ta namen smo razvili sistem za vizualizacijo opisanih eksperimentov v spletnem brskalniku z uporabo tehnologije WebGL. Sistem je optimiziran za čim hitrejše in učinkovitejše delovanje na grafičnih procesorskih enotah. Pri tem smo se osredotočili na minimizacijo klicev za izris, in minimizacijo prenosa podatkov na grafično kartico. To se je izkazalo za zelo učinkovito rešitev, kar potrjuje performančna analiza razvitega sistema.

1 Uvod

Vizualizacija predstavlja zelo pomembno vlogo v znanosti. S primerno vizualizacijo omogočimo lažje razumevanje obravnavanega problema, spremljamo obnašanje določenega sistema, ali morda celo opazimo določene lastnosti, ki iz surovih podatkov niso opazne. Predvsem pomembna je vizualizacija na področju fizike, natančneje pri opazovanju dogajanja v kvantni mehaniki, kjer stvari postanejo hitro kontraintuitivne. Primerna vizualizacija omogoči pravilnejše, natančneje in boljše razumevanje delovanja tovrstnih sistemov in izpopolnjevanje znanstvene teorije.

Z napredkom v tehnologiji je vizualizacija postala zelo pomembna pri trkih osnovnih delcev. Tovrstni eksperimenti so nadzorovani z uporabo množice različnih detektorjev prilagojenih za zaznavanje izbranih tipov delcev. Pri trkih delcev z visokimi energijami pride do razpada v še bolj osnovne delce, ki pa jih želimo zaznati in zaradi boljšega razumevanja tudi vizualizirati.

Posamezni dogodek zajema trke okoli 50 parov delcev, ki razpadejo na množico osnovnejših delcev. Tisti, ki nosijo naboj se po trku pod vplivom močnega magnetnega polja odklonijo in zaidejo skozi množico detektorjev, kjer se izmerijo njihove lastnosti, kot so položaj, gibalna količina in energija. Takšen dogodek v eksperimentu je predstavljen z množico točk v prostoru, kjer so bili delci zaznani in njihove izmerjene lastnosti v teh točkah. Za nadaljnje študije je potrebno rekonstruirati poti vseh nastalih delcev, kar predstavlja zelo težak problem. Da pri rekonstrukciji lažje opredelimo pravilnost posamezne poti in morebitne napake (npr. zaradi prisotnosti šuma) potrebujemo primerno vizualizacijo.

Upodabljanje tovrstnih podatkov razdelimo na interaktivne vizualizacije podatkov dogodka (prikazi dogodkov), vizualizacije statističnih podatkov za namene analize podatkov in prostorsko neodvisne vizualizacije podatkov. Tovrstne vizualizacije so predstavljene v [2]. Zajemajo rešitve na namiznih napravah z uporabo vmesnika OpenGL [6] in prenosnih napravah z uporabo vmesnika WegbGL [4]. Primere statističnih 2D in 3D vizualizacij predstavlja spletno orodje JSRoot [1].

Vizualizacija takšnih eksperimentov lahko hitro postane problematična zaradi velikega števila elementov, še posebej kadar želimo zagotoviti visoko stopnjo interaktivnosti. Pri razvoju takšnega sistema želimo razrešiti sledeče izzive:

- intuitivna vizualizacija prikaza podatkov osnovnih delcev in njihovih trajektorij v prostoru,
- realizacija platformno neodvisnega prikaza,
- visoka stopnja odzivnosti vizualizacije na interakcijo in

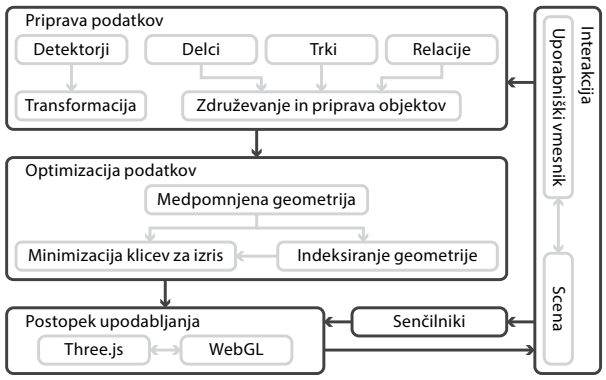
- možnost uporabe na vsakdanjih (tudi mobilnih) sistemih.

Cilj je izdelava sistema, ki omogoča interaktivno vizualizacijo geometrije eksperimenta in vizualizacijo dogodkov interakcije delcev s posameznim detektorjem, kar predstavlja veliko količine delcev znotraj različnih delov geometrije eksperimenta in rekonstruiranih trajektorij delcev. Hkrati želimo za posamezni delec intuitivno vizualizirati izbrane parametre. Z uporabo upodabljanja v brskalniku (z uporabo WebGL-a) želimo realizirati platformno neodvisni prikaz in zagotoviti možnost uporabe na vsakdanjih sistemih.

Takšen sistem predstavimo v naslednjem poglavju 2. V poglavju 3 sistem analiziramo in ovrednotimo. V zadnjem poglavju podamo tudi sklepne ugotovitve in možna izhodišča z nadaljnje delo.

2 Sistem za vizualizacijo trkov

Razviti sistem temelji na upodabljanju v spletnem brskalniku z uporabo tehnologije [4]. Zaradi enostavnosti smo uporabili programsko ogrodje Three.js [3], namenjeno upodabljanju interaktivnih 3D upodobitev v spletnih brskalnikih v realnem času in abstrahira nizkonivojske funkcionalnosti WebGL-a. Shema delovanja sistema je prikazana na sliki 1, v nadaljevanju pa je predstavljen posamezni korak.



Slika 1: Na sliki je prikazana shema razvitega sistema.

2.1 Priprava podatkov

Podatke o geometriji detektorjev in podatke o posameznih dogodkih v eksperimentih, smo pridobili s strani izziva TrackML[1] objavljenega na spletišču Kaggle. Podatki so na voljo v pomoč pri reševanju omenjenega izziva, namenjenega rekonstrukciji poti delcev razpisanega v letu 2018 v CERN-u.

Podatki o geometriji detektorjev so na voljo kot atributi v datoteki tipa CSV, ki jih je potrebno pred uporabo pretvoriti v primerno obliko za izris. Posamezni vnos v datoteki prestavlja centroid modula posameznega detektorja, ki jo moramo s primerno transformacijo, glede na tip detektorja, razširiti v pravokotni ali trapezoidni modul primernih dimenzij in postaviti na ustrezen položaj v prostoru. Končne detektorske module s primerno orientacijo in indeksacijo vozlišč shranimo v medpomnjeno geometrijo in za potrebe osvetlitve poračunamo normale v ogliščih.

Podatki o dogodkih eksperimenta znotraj detektorjev so prav tako zapisani datotekah tipa CSV. Posamezni dogodek zajema informacije o delcih ob trku znotraj detektorja, informacije o vseh interakcijah delcev z detektorji in informacije, ki povezujejo interakcije s posameznimi delci. Informacije povežemo in shranimo v obliki primerni za izris.

2.2 Optimizacija izrisa

Z optimizacijo izrisa zagotovimo potrebi po visoki stopnji interaktivnosti vizualizacije in s tem možnosti uporabe na vsakdanjih sistemih.

2.2.1 Medpomnjena geometrija

Z namenom učinkovite predstavitve geometrije (poligonov, črt ali točk) smo uporabili medpomnjeno geometrijo, ki zahteva nižjenivojsko razumevanje, a omogoča hitrejši izris. Vsi atributi so shranjeni v medpomnilnikih kar skrajša čas prenosa podatkov na grafično kartico (GPE).

2.2.2 Indeksiranje geometrije

Za dodatno pohitritev smo za predstavitev geometrije detektorjev in trajektorij uporabili indeksirano obliko zapisa. S tem zmanjšamo porabo pomnilnika kot tudi skrajšamo čas prenosa na GPE. Z uporabo indeksiranja na GPE omogočimo možnost uporabe predpomnjenja (Post Transform Cache). S tem omogočimo optimizacijo nad prekrivajočimi podatki na GPE.

2.2.3 Minimizacija klicev za izris

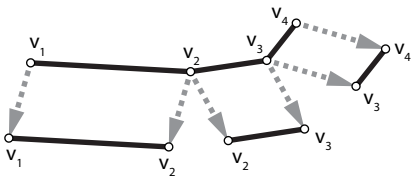
Ker v opisu posameznega dogodka trka nastopa velika količina informacij (npr. 12263 delcev, 120939 trkov in 10566 trajektorij), smo se pri optimizaciji najbolj osredotočili na minimizacijo klicev za izris (angl. draw calls), kar se je izkazalo za pomemben dejavnik in je razvidno iz rezultatov v poglavju 3.

Geometrijo posameznih detektorjev upodobimo z uporabo trikotniških primitivov, ki za skupke treh vozlišč celotne geometrije priredi trikotnike. To nam omogoča, da celotno geometrijo posameznega detektorja upodobimo z enim klicem za izris.

Podatke vseh delcev v začetnih položajih in položaje detekcij trkov upodobimo kot točke. Na ta način upodobimo celotno množico začetnih položajev delcev in vse detekcije trkov vsako z enim klicem za izris.

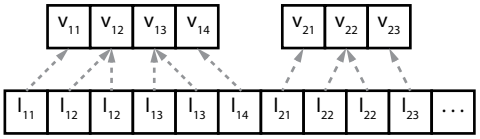
Posamezno trajektorijo razdelimo na segmente daljic (slika 2). Segmente daljic vseh trajektorij shranimo kot skupno geometrijo v medpomnilniku na način, ki nam omogoča ločevanje med trajektorijami, obenem pa povezuje daljice iste trajektorije (slika 3). Ker se vozlišča znotraj trajektorije podvajajo (glej sliko 3), izvedemo že prej omenjeno indeksiranje geometrije, pri čemer obdržimo

[1] https://www.kaggle.com/c/trackml-particle-identification



Slika 2: Na sliki je prikazana zgradba posamezne trajektorije iz posameznih daljic.

relacije daljic v posameznih trajektorijah. S tem pridobimo še dodatno pohitritev. Množico vseh trajektorij upo-



Slika 3: Slika prikazuje indeksiranje geometrije trajektorij.

dobimo kot verigo črt. Na ta način lahko vse trajektorije upodobimo zgolj z enim klicem za izris.

S predstavljenimi pristopi omogočimo izris z minimalnim številom klicev za izris in zagotovimo visoko stopnjo neodvisnega delovanja GPE (analiza v poglavju 3).

2.3 Scena

Za namene intuitivne vizualizacije smo za prikaz celotnega eksperimenta upodobili sceno, ki zajema:

- geometrijo eksperimenta z detektorji,
- začetne položaje delcev,
- zaznane trke delcev z detektorji in
- rekonstruirane trajektorije.

Po sceni navigiramo z uporabo miške in tipkovnice. Objekti so osvetljeni s šestimi lučmi, postavljenimi na ploskve navidezne omejevalne kocke.

2.3.1 Detektorji

Osvetljevanje geometrije detektorja izračunamo z uporabo Lambertovega modela osvetljevanja[2], senčenje pa izvedemo z Gouraudovo tehniko senčenja[3].

Ker se posamezni detektorji v prostoru prekrivajo, smo prosojnost simulirali tako, da jih najprej izrisujemo od zunanjih proti notranjim brez testa globine. S tem omogočimo prosojen pogled skozi detektorje iz središča navzven. Ker smo test globine pri tem onemogočili, so detektorji izrisani v obratnem vrstnem redu. Zato v drugem koraku detektorje izrišemo še obratno (od notranjega proti zunanjemu) s testom globine. S tem omogočimo še prosojen pogled v notranjost, zagotovimo pa pravilen izris geometrije glede na globino.

[2] `https://en.wikipedia.org/wiki/Lambertian_reflectance`

[3] `https://en.wikipedia.org/wiki/Gouraud_shading`

2.3.2 Delci

Senčenje delcev smo izvedli z nadzorom nad velikostjo, barvo in prosojnostjo. Pri tem lahko barvo določimo glede na atribut gibalne količine delca ali pa priredimo enotno barvo vsem delcem. Velikost upodobljenih točk, s katerimi ponazorimo delce izračunamo glede na oddaljenost od kamere. Tako delce, ki so bližje kameri izrišemo večje kot bolj oddaljene. V primeru, da smo v sceni izbrali določen element dogodka za opazovanje, to dodatno poudarimo s spremembo velikosti točke delca, barve in prosojnosti.

2.3.3 Trki

Tudi senčenje zaznanih trkov delcev smo izvedli z nadzorom nad velikostjo, barvo in prosojnostjo. Barvo določimo glede na tip detektorja, kjer je bil delec zaznan, ali pa priredimo enotno barvo. Za boljšo predstavo razporeditve položajev detekcij izrisujemo tiste, ki so bližje kameri z večjimi točkami. Izbiro elementa dogodka v sceni poudarimo s spremembo velikosti točke ki ponazarja trk, barve in prosojnosti.

2.3.4 Trajektorije

Senčenje rekonstruiranih trajektorij smo realizirali z nadzorom nad barvo in prosojnostjo. Barvo določimo glede na atribut gibalne količine v delu trajektorije ali pa priredimo enotno barvo. Zaradi velikega števila trajektorij smo pri senčenju trajektorij barvo in prosojnost dodatno utežili glede na število vseh trajektorij in oddaljenostjo od kamere. Pri izbiri trajektorije le-to poudarimo s spremembo debeline, barve in prosojnosti črt.

Pri spremembi debeline trajektorije smo testirali naslednje rešitve: *metodo WebGL API-ja* `glLineWidth` - kjer z `lineWidth()` določimo širino rasteriziranih linij. Ta pristop ni podprt na vseh platformah in ne bi zadostili cilju platformne neodvisnosti.

enakomerno porazdeljene točke - kjer poskušamo linijo simulirati z enakomerno porazdeljenimi točkami. Pri tem pristopu, glede na velikost točke in dolžino posameznega segmenta trajektorije, izračunamo položaje točk za zapolnitev prostora med dvema vozliščema in s tem ponazorimo linijo. To se je izkazalo za izvedljiv pristop, vendar pa take linije niso zvezne v vseh primerih.

izris s trakom trikotnikov - kjer krajišči vsake linije razdelimo na pare točk in jih pravokotno odmaknemo v smeri normale za želeno debelino (slika 4). Pri tem upoštevamo položaj kamere in dimenzije platna. Dobljene trakove nato izrišemo. To je tudi končna uporabljena rešitev.



Slika 4: Ponazoritev izračuna debeline za posamezni segment trajektorije.

2.4 Uporabniški vmesnik

Za lažjo interakcijo smo implementirali uporabniški vmesnik z uporabo knjižnice dat.GUI[4].

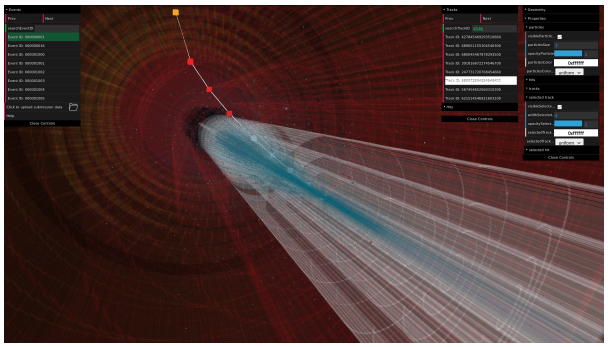
Ob zagonu sistema se v sceni prikaže geometrija detektorja in uporabniški vmesnik. Premikamo se lahko z orbitalno kamero. Uporabniški vmesnik nam ponuja nastavitve posameznih atributov (rezalne ravnine, vidljivost, barva, prosojnost, velikost, tip obarvanja), prikaza detektorjev, delcev, njihovih trkov z detektorji in rekonstruiranih trajektorij. Posamezni dogodek za opazovanje lahko izberemo iz prikazanega seznama. Ob izbiri dogodka se v sceno naloži delce, trke in trajektorije.

Ker je elementov veliko, smo implementirali možnost iskanja preko identifikacijske številke. Za opazovanje lahko izberemo določen trk ali trajektorijo delca, pri čemer se kamera postavi ob izbrani element. V primeru izbire trajektorije se kamera postopoma pomakne med začetno in končno vozlišče trajektorije (odmaknjeno v smeri normale) ter se usmeri v njen centroid. Ob izbiri trka se kamera postopoma pomakne med središče sistema in trk ter se vanj usmeri.

Z namenom pomoči tekmovalcem CERN-ovega izziva TrackML smo dodali tudi možnost nalaganja lastnega dogodka eksperimenta, ki ga lahko primerjamo s katerim izmed izbranih. Na tem mestu smo implementirali funkcionalnost, ki ob izbiri trajektorije priredi seznam najbolj podobnih trajektorij glede na centroid in število segmentov.

3 Analiza in rezultati

Končni izgled razvitega sistema je prikazan na sliki 5, ki prikazuje primer upodobitve enega dogodka trka.



Slika 5: Primer upodobitve fizikalnega dogodka trka delcev z našim sistemom.

Sistem smo testirali na napravi z 2-jedrnim procesorjem *Intel Core i7 4510U* pri taktu 2.00 GHz, 8 GB glavnega pomnilnika in grafično kartico *AMD Radeon R7 M260*. Pri testiranju smo v sceno naložili isti dogodek trka delcev, ob tem pa smo kamero pri vseh testih postavili v enak začetni položaj. Merjenje smo izvedli na vzorcu 10.000 časov upodabljanja, katere smo povprečnici za končni rezultat.

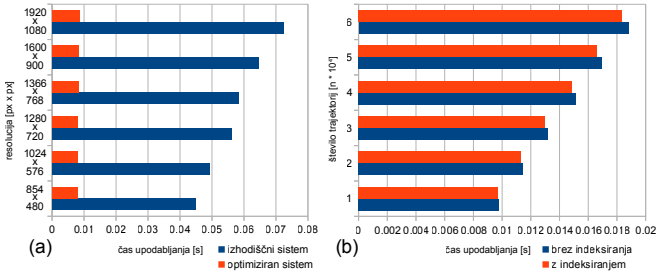
[4] https://github.com/dataarts/dat.gui

3.1 Pohitritev sistema

Testirali smo celotno pohitritev našega sistema v primerjavi z izhodiščnim sistemom (slika 6 (a)). Primerjali smo čas upodabljanja obeh sistemov v odvisnosti od dimenzije platna. Iz testa je razvidno, da skozi spremembo dimenzije naš sistem upodablja v krajšem času kot izhodiščni. To nam omogoča visoko interaktivnost in možnost, da lahko vizualizacijo dodatno nadgradimo (npr. implementiramo obojestransko prosojnost detektorjev za boljšo preglednost).

3.2 Indeksiranje trajektorij

Testirali smo doprinos indeksiranja trajektorij, rezultati so prikazani na sliki 6 (b). Merili smo doprinos uporabe indeksiranja geometrije trajektorij našega sistema v odvisnosti od njihovega števila. Razvidno je, da se doprinos veča s številom prisotnih trajektorij, kar je tudi pričakovano.



Slika 6: (a) pohitritev sistema, (b) čas upodabljanja trajektorij.

4 Zaključek in nadaljnje delo

Poleg omenjenega sistema želimo analizirati in uporabiti še hibridni način upodabljanja z uporabo platforme Med3D[5] [5], ki omogoča odloženo upodabljanje. S tem načinom upodabljanja pričakujemo, da se bo izkazal za še bolj učinkovito in hitrejše predvsem zaradi pospešenega izračuna osvetlitve, kot tudi možnosti združevanja različnih načinov izrisa.

5 Zahvala

Za sodelovanje se zahvaljujemo članom skupine za vizualizacijo pri organizaciji CERN in posameznih fizikalnih eksperimentih (ATLAS, CMS in Alice).

Literatura

[1] Theo Beffart, Maximilian Früh, Christoph Haas, Sachin Rajgopal, Jonas Schwabe, Christoph Wolff, and Marek Szuba. RootJS: Node.js Bindings for ROOT 6. *CoRR*, abs/1704.07887, 2017.

[2] Matthew Bellis, Riccardo Maria Bianchi, Sebastien Binet, Ciril Bohak, Benjamin Couturier, Hadrien Grasland, Oliver Gutsche, Sergey Linev, Alex Martyniuk, Thomas McCauley, Edward Moyse, Alja Mrak Tadel, Mark Neubauer, Jeremi Niedziela, Leo Piilonen, Jim Pivarski, Martin Ritter, Tai Sakuma, Matevz Tadel, Barthélémy von Haller, Ilija Vukotic, and Ben Waugh. Hep software foundation community white paper working group – visualization, 2018.

[3] Brian Danchilla. *Three.js Framework*, pages 173–203. Apress, Berkeley, CA, 2012.

[5] https://github.com/UL-FRI-LGM/Med3D

[4] Dean Jackson and Jeff Gilbert. Webgl specification. Technical report, The Khronos Group Inc., 2015.

[5] Primož Lavrič, Ciril Bohak, and Matija Marolt. Spletno vizualizacijsko ogrodje z možnostjo oddaljenega sodelovanja. In *Zbornik petindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2016, 19. - 21. september 2016, Portorož, Slovenija*, pages 43–46, 2016.

[6] Mark Segal and Kurt Akeley. The OpenGL Graphics System: A Specification (Version 3.1). Technical report, The Khronos Group Inc., March 2009.

Psevdokanonična predstavitev grafov na podlagi sosednosti

Luka Fürst

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, SI-1000 Ljubljana
E-pošta: luka.fuerst@fri.uni-lj.si

Neighborhood-based Pseudo-canonical Representation of Graphs

A canonical representation of a graph is a string that is invariant to the numbering of the graph vertices. Consequently, two graphs are isomorphic if and only if they have equal canonical representations. The problem of computing a canonical representation is thus at least as hard as the graph isomorphism problem. In this paper, we present a polynomial-time algorithm for finding a pseudo-canonical graph representation with the property that a pair of equal representations implies isomorphism but the opposite is not necessarily true. Nevertheless, for a given graph the algorithm iteratively builds representations that are increasingly more likely to be numbering-invariant. To validate our approach, we applied the proposed algorithm to the complete set of connected simple undirected graphs with up to 9 vertices and to a set of real-world graphs.

1 Uvod

Takšna in drugačna omrežja igrajo v današnjem svetu čedalje pomembnejšo vlogo. Za učinkovito preučevanje omrežij pa potrebujemo učinkovite grafne algoritme. V številnih postopkih za delo z grafi oz. omrežji nam pride prav *kanonični zapis* [1]. Gre za zapis, ki enolično določa strukturo grafa, od oštevilčenja vozlišč pa je povsem neodvisen; če številke vozlišč med seboj premešamo, se kanonični zapis ne sme spremeniti. Grafa imata enak kanonični zapis natanko tedaj, ko sta *izomorfna*, torej ko imata enako strukturo.

Eden od temeljnih problemov na področju odkrivanja vzorcev v grafih (angl. graph data mining) [2] je problem naštevanja neizomorfnih podgrafov v danem gostiteljskem grafu [3]. Pri reševanju tega problema nam kanonični zapisi koristijo takrat, ko preverjamo, ali smo izomorfno kopijo pravkar odkritega podgrafa že odkrili: namesto da bi iskali izomorfizme, med seboj preprosto primerjamo kanonične zapise. Kanonični zapisi se uporabljajo tudi na področju kemije, saj omogočajo učinkovito iskanje spojin in funkcionalnih skupin glede na njihovo strukturo [4].

Ker se preverjanje izomorfizma prevede na preverjanje enakosti kanoničnih zapisov, je problem določitve kanoničnega zapisa vsaj tako težak kot problem grafnega izomorfizma. Ta problem pa je GI-poln, zato ne vemo, ali zanj obstaja algoritem s polinomsko časovno zahtevnostjo.[1] Najboljši znani algoritem teče v kvazipolinomskem času [5].

V nekaterih aplikacijah lahko zahteve glede kanoničnega zapisa nekoliko omilimo in lahko, denimo, izomorfna grafa »občasno« proglasimo za neizomorfna [6]. V takih primerih zadošča *psevdokanonični* zapis, ki ga bomo opredelili s sledečo zahtevo:

> Če sta psevdokanonična zapisa grafov med seboj enaka, sta grafa izomorfna.

Seveda je zaželeno, da lastnost čim večkrat drži tudi v obratni smeri, vendar pa tega formalno ne zahtevamo.

V članku predstavljamo algoritem za tvorbo psevdokanoničnega zapisa grafov. Algoritem se izvaja v polinomskem času in postopoma gradi čedalje boljše predstavitve; čim dlje ga pustimo teči, tem večja je verjetnost, da bo zapis, ki ga bo zgradil za podani graf, dejansko kanoničen (*popolnoma* neobčutljiv na oštevilčenje vozlišč), čeprav, kot bomo videli, obstajajo grafi, pri katerih algoritem ne more zgraditi takšnega zapisa.[2] Algoritem smo preizkusili na popolnem naboru povezanih neusmerjenih enostavnih grafov z največ 9 vozlišči in na nekaterih grafih iz realnega sveta. Po nam dosegljivih podatkih lahko algoritem in njegovo teoretično in praktično analizo obravnavamo kot prispevek k znanosti.

Psevdokanonično predstavitev gradimo na podlagi sosednosti vozlišč, načeloma pa bi lahko algoritem zasnovali na poljubni kombinaciji *grafnih invariant* — lastnosti, ki so odvisne samo od strukture grafa, ne pa od konkretnega oštevilčenja vozlišč.

V razdelku 2 bomo opredelili pojme, ki jih bomo potrebovali skozi celotni prispevek, v razdelku 3 pa bomo predstavili algoritem za izdelavo psevdokanoničnega zapisa podanega grafa. V razdelku 4 bomo delovanje metode eksperimentalno potrdili, z razdelkom 5 pa bomo prispevek zaključili.

2 Terminologija in notacija

Graf G je dvojica (V_G, E_G), kjer je V_G množica *vozlišč*, $E_G \subseteq V_G \times V_G$ pa množica *povezav*. Namesto

[1] Kratica GI se nanaša na problem grafnega izomorfizma (angl. graph isomorphism problem).

[2] Če razreda problemov GI in P nista istovetna, potem taki grafi obstajajo pri *poljubnem* polinomskem algoritmu.

V_G in E_G bomo pisali V in E, kadar bo jasno, kateremu grafu pripadata množici, ali pa ko to ne bo pomembno. Brez pretirane izgube splošnosti se bomo omejili na neusmerjene grafe brez zank: $(v, v) \notin E$, $(u, v) \in E \iff (v, u) \in E$. Prav tako bomo predpostavili, da velja $|V| = n$ in $V = \{1, 2, \ldots, n\}$. Rekli bomo, da je vozlišče u *manjše* (oz. *večje*) od vozlišča v, če velja $u < v$ (oz. $u > v$). Množico vseh sosedov vozlišča u bomo označili z $\mathcal{N}(u) = \{v \in V \mid (u, v) \in E\}$.

Grafa G in H sta *izomorfna* (oznaka: $G \simeq H$), če obstaja bijektivna preslikava $h\colon V_G \to V_H$, ki ohranja sosednosti in nesosednosti, torej če za vsak par vozlišč $u, v \in V_G$ velja $(u, v) \in E_G \iff (h(u), h(v)) \in E_H$. Takšni preslikavi pravimo *izomorfizem*.

Niz $z^*(G)$ je *kanonični zapis* grafa G, če za vsak par grafov G in H velja $z^*(G) = z^*(H) \iff G \simeq H$. Niz $z(G)$ je *psevdokanonični zapis* grafa G, če za vsak par grafov G in H velja $z(G) = z(H) \implies G \simeq H$.

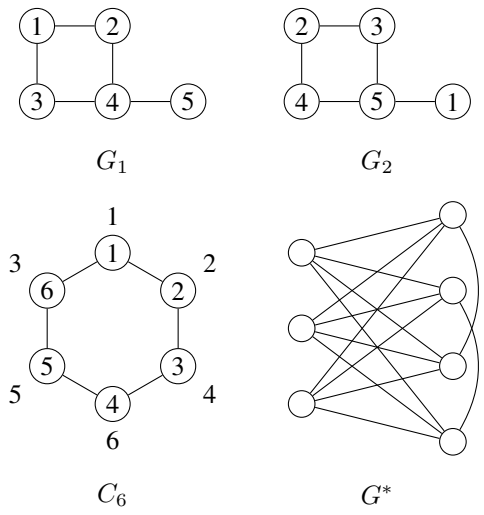
3 Algoritem

Za začetek pokažimo sledečo trditev:

Trditev 1. Niz $z(G) = S_1/S_2/\ldots/S_{n-1}$, kjer je $S_u = s_{u,u+1}s_{u,u+2}\ldots s_{u,n}$ za vsak $u \in \{1, \ldots, n-1\}$ in $s_{u,v} = 1$ (v primeru $(u, v) \in E_G$) oziroma 0 (v nasprotnem primeru) za vsak $v > u$, je psevdokanonični zapis grafa G.

Na primer, za grafa G_1 in G_2 na sliki 1 velja $z(G_1) = 1100/010/10/1$ in $z(G_2) = 0001/110/01/1$.

Dokaz. Če velja $z(G) = z(H)$, potem imata grafa G in H gotovo enako število vozlišč. Poleg tega za vsak par (u, v) z lastnostjo $u < v$ velja $(u, v) \in E_G \iff (u, v) \in E_H$. Ker sta grafa neusmerjena, velja ta lastnost za vse pare (u, v), od tod pa že sledi, da sta grafa izomorfna. □



Slika 1: Primeri grafov.

Žal pa opisana predstavitev sama po sebi ni nujno odporna na spremembe v oštevilčenju vozlišč. Grafa G_1 in G_2 na sliki 1 sta izomorfna, vendar pa, kot smo videli, nimata enakih psevdokanoničnih zapisov. Zato se nam ponuja sledeča ideja: vozliščem grafa dodelimo takšne številke, da bo zapis, izdelan po metodi iz trditve 1, enak pri čimveč različnih začetnih oštevilčenjih vozlišč. Problem (psevdo)kanonične predstavitve grafa tako prevedemo na problem (psevdo)kanoničnega oštevilčenja vozlišč grafa.

Pri iskanju (psevdo)kanoničnega oštevilčenja se moramo opreti na lastnosti vozlišč, ki so neodvisne od začetnega oštevilčenja. Algoritem, ki ga bomo predstavili, obravnava soseščine vozlišč. Na primer, pri grafih G_1 in G_2 na sliki 1 ima eno vozlišče enega soseda, eno vozlišče tri sosede, tri vozlišča po dva soseda. Vozlišču z enim sosedom se splača dodeliti (denimo) številko 1, vozlišču s tremi sosedi (denimo) številko 5, ostalim vozliščem pa številke 2, 3 in 4. Številki 1 in 5 bosta tako vedno, ne glede na začetno oštevilčenje grafa, dodeljeni vozliščema z enim oziroma tremi sosedi. Če poleg neposrednih sosedov upoštevamo tudi sosede sosedov, sosede sosedov sosedov itd., pa dobimo še robustnejše oštevilčenje.

Postopek za določanje psevdokanoničnega oštevilčenja vozlišč je prikazan kot algoritem 1. Algoritem se prične s klicem funkcije IZVRSI na podanem grafu. Funkcija na začetku vsem vozliščem dodeli številko $id = 1$, nato pa v zanki tako dolgo kliče funkcijo NAPREJ in po potrebi še funkcijo RAZPLETI, dokler nima vsako vozlišče svoje unikatne številke.

Funkcija NAPREJ v k-ti iteraciji obravnava množico W, ki jo tvorijo vozlišča s številko k, a le v primeru, če velja $|W| \geq 2$. (Ko namreč neko vozlišče dobi unikatno številko, ga ne obravnavamo več; njegova številka postane dokončna.) Za vsako vozlišče w iz množice W se zgradi zaporedje številk vozlišč njegovih sosedov, nato se to zaporedje leksikografsko uredi (funkcija UREDI), da dobimo zaporedje $T(w)$, nazadnje pa se vozliščem v množici W dodelijo številke glede na položaje njihovih zaporedij $T(w)$ v leksikografsko urejenem zaporedju zaporedij $T(w_1), T(w_2), \ldots, T(w_{|W|})$. (Operator $\prec$ v vrstici 24 preberemo kot »je leksikografsko manjši od«.) Vozlišča z enakim zaporedjem dobijo isto številko, zato se bo z njimi še treba ukvarjati.

Če klic funkcije NAPREJ v ničemer ne spremeni oštevilčenja, moramo spremeniti vsaj eno neunikatno številko, saj bi se sicer ujeli v neskončno zanko. V funkciji RAZPLETI poiščemo množico vozlišč z najmanjšo neunikatno številko in vsem vozliščem iz te množice (razen najmanjšemu) povečamo številko za 1.

Tabela 1 prikazuje delovanje algoritma na primeru grafa C_6 s slike 1. (Na sliki podajajo številke v vozliščih začetno, zunanje številke pa dobljeno končno oštevilčenje.) Na začetku vsem vozliščem dodelimo številko 1. Ker imajo vsa vozlišča enako zaporedje T ($\langle 1, 1\rangle$), se številke vozlišč ne spremenijo, zato pokličemo funkcijo RAZPLETI. Vozlišče 1 dobi unikatno številko 1 (z njim se ne bomo več ukvarjali), ostala pa številko 2. V naslednjem klicu funkcije NAPREJ velja $T(2) = T(6) = \langle 1, 2\rangle$ in $T(3) = T(4) = T(5) = \langle 2, 2\rangle$, zato vozlišči 2 in 6 ohranita številko 2, vozlišča 3, 4 in 5 pa dobijo številko 4. Postopek nadaljujemo, dokler nima vsako vo-

Algoritem 1 Psevdokanonično oštevilčenje vozlišč.

```
1: function IZVRSI(G = (V, E))
2:     for all v ∈ V do
3:         id(v) := 1
4:     end for
5:     repeat
6:         sprememba := NAPREJ(G, id)
7:         konec := (∀u, v ∈ V : id(u) ≠ id(v))
8:         if ¬konec ∧ ¬sprememba then
9:             RAZPLETI(G, id)
10:        end if
11:    until konec
12:    return id
13: end function
14: function NAPREJ(G = (V, E), id)
15:    id' := id
16:    for all k ∈ {1, 2, ..., |V|} do
17:        W := VRSTNIKI(V, id', k)
18:        if |W| > 1 then
19:            for all w ∈ W do
20:                T_0 := ⟨id'(w') | w' ∈ N(w)⟩
21:                T(w) := UREDI(T_0)
22:            end for
23:            for all w ∈ W do
24:                P := {w' ∈ W | T(w') ≺ T(w)}
25:                id(w) := |P| + 1
26:            end for
27:        end if
28:    end for
29:    return (id ≠ id')
30: end function
31: function RAZPLETI(G = (V, E), id)
32:    k* := min{k | |VRSTNIKI(V, id, k)| > 1}
33:    W := VRSTNIKI(V, id, k*)
34:    for all w ∈ W \ {min(W)} do
35:        id(w) := id(w) + 1
36:    end for
37: end function
38: function VRSTNIKI(V, id, k)
39:    return {v ∈ V | id(v) = k}
40: end function
```

Tabela 1: Delovanje algoritma na grafu C_6.

	1	2	3	4	5	6
id	1	1	1	1	1	1
T	$\langle 1,1\rangle$	$\langle 1,1\rangle$	$\langle 1,1\rangle$	$\langle 1,1\rangle$	$\langle 1,1\rangle$	$\langle 1,1\rangle$
id	1	1	1	1	1	1
id	**1**	2	2	2	2	2
T		$\langle 1,2\rangle$	$\langle 2,2\rangle$	$\langle 2,2\rangle$	$\langle 2,2\rangle$	$\langle 1,2\rangle$
id		2	4	4	4	2
T		$\langle 1,4\rangle$	$\langle 2,4\rangle$	$\langle 4,4\rangle$	$\langle 2,4\rangle$	$\langle 1,4\rangle$
id		2	4	**6**	4	2
id		**2**	4		4	**3**
T			$\langle 2,6\rangle$		$\langle 3,6\rangle$	
id			**4**		**5**	

zlišče svoje unikatne številke. Vidimo, kako algoritem na začetku upošteva samo neposredne sosede, nato pa spremembe številk vplivajo na vse bolj oddaljena vozlišča.

Časovno zahtevnost algoritma lahko določimo na podlagi opažanja, da v vsakem obhodu zanke vsaj eno vozlišče grafa pridobi svojo končno številko. Če funkcija NAPREJ vrne **false**, potem funkcija RAZPLETI nekemu vozlišču dodeli unikatno številko, unikatna številka pa je samodejno tudi dokončna. V nasprotnem primeru pa funkcija NAPREJ bodisi pridela vsaj eno unikatno številko ali pa poveča številke vsaj eni skupini vozlišč. Številke vozlišč v tej skupini sicer niso unikatne, toda vsaj eno vozlišče iz skupine bo ohranilo svojo številko do konca izvajanja algoritma. Od tod sledi, da imamo največ $(n-1)$ iteracij postopka, vsaka iteracija pa zahteva čas $O(n^2 \log n)$: za vsako od n vozlišč uredimo zaporedje številk vseh njegovih sosedov v času $O(n \log n)$, nato pa za celoten graf uredimo zaporedje zaporedij številk v času $O(n^2 \log n^2) = O(n^2 \log n)$. Skupna časovna zahtevnost algoritma potemtakem znaša $O(n^3 \log n)$.

4 Eksperimentalni rezultati

Psevdokanonično predstavitev lahko obravnavamo kot kanonično, če je povsem neobčutljiva na začetno oštevilčenje vozlišč. To lahko preverimo tako, da zgradimo psevdokanonični zapis za vsako od $n!$ možnih oštevilčenj. Če so vsi zapisi med seboj enaki, gre za pravo kanonično predstavitev.

Algoritem 1 smo izvedli na celotni množici povezanih enostavnih neusmerjenih grafov z največ 9 vozlišči. Tabela 2 za vsak $n \in \{2, \ldots, 9\}$ in $i \in \{1, \ldots, n-1\}$ prikazuje, pri koliko grafih z n vozlišči dobimo po i iteracijah zanke v funkciji IZVRSI zapis, ki ni povsem neobčutljiv na začetno oštevilčenje vozlišč. Graf G^* na sliki 1 je najmanjši graf (in edini tak graf s sedmimi vozlišči), čigar končni psevdokanonični zapis ni kanoničen. Kot vidimo, imajo vsa vozlišča grafa enako število sosedov. Takšni grafi povzročajo našemu algoritmu največ preglavic.

V drugi skupini poskusov smo algoritem preizkusili na nekaterih realnih grafih iz zbirk KONECT [7] in SNAP [8]. V fazi predobdelave smo v vseh grafih odstranili morebitne oznake vozlišč in povezav, usmerjene povezave spremenili v neusmerjene in odstranili morebitne večkratne povezave in zanke. Pri poskusih smo algoritem 1 primerjali z algoritmom za računanje pravih kanoničnih zapisov, ki je vgrajen v sistem Sage.[3] Algoritem 1 smo implementirali v programskem jeziku C, vse poskuse pa smo izvršili na računalniku z 8-jedrnim procesorjem Intel Core i7-3770 s taktom 3,40 GHz.

Rezultati izvajanja poskusov so zbrani v tabeli 3. Poleg porabe časa v milisekundah (t_{PK} za algoritem 1 in t_{K} za algoritem v sistemu Sage) podajamo tudi število iteracij (N) zanke v funkciji IZVRSI. V skladu s pričakovanji se psevdokanonični zapisi v skoraj vseh primerih izračunajo hitreje od kanoničnih. (Odgovora na vprašanje, zakaj je graf Advogato izjema, zaenkrat še nimamo.) Ker je število permutacij vozlišč preveliko, ne moremo zanesljivo trditi, ali so dobljeni psevdokanonični zapisi

[3] http://www.sagemath.org/

Tabela 2: Število nekanoničnih psevdokanoničnih zapisov za grafe v množici vseh povezanih enostavnih neusmerjenih grafov z n vozlišči ($\mathcal{G}_n$).

		i							
n	$\lvert\mathcal{G}_n\rvert$	1	2	3	4	5	6	7	8
2	1	0							
3	2	0	0						
4	6	2	1	0					
5	21	12	7	3	0				
6	112	88	52	27	7	0			
7	853	781	427	235	89	20	1		
8	11 117	10 873	6032	2780	1164	319	52	19	
9	262 180	260 105	145 392	44 952	16 295	4691	899	201	113

Tabela 3: Rezultati na realnih grafih.

Graf	$\lvert V\rvert$	$\lvert E\rvert$	t_{PK}	t_{K}	N
Les Misérables	77	254	1,8	8,8	28
David Copperfield	112	425	0,7	5,6	5
Jazz	198	2742	4,5	16,9	10
US Power Grid	4941	6594	32 700	58 700	481
Advogato	6541	39 285	230 000	124 000	2013
ca-HepTh	9877	25 973	615 000	1 450 000	2266
ca-CondMat	23 133	93 439	9 220 000	23 700 000	5999

odporni na spremembe v oštevilčenju vozlišč, vendar pa se to po preizkušanju z naključnimi permutacijami začetnega oštevilčenja zdi dokaj verjetno.

5 Zaključek

Predstavili smo polinomski algoritem za izdelavo psevdokanoničnega zapisa grafa in ga preizkusili na umetnih in realnih grafih. Vsaj za grafe z največ devet vozlišči lahko trdimo, da je zapis, ki ga proizvede algoritem, v veliki večini primerov dejansko kanoničen, torej neobčutljiv na oštevilčenje vozlišč.

Pri gradnji psevdokanoničnih zapisov bi si lahko pomagali s simetrijami. Na primer, pri grafu K_6 so pravzaprav vsa oštevilčenja kanonična: ne glede na to, kako vozliščem dodelimo številke, bo dobljeni zapis vedno enak. Pri grafu C_6 imamo manj svobode; izkaže se, da lahko med seboj permutiramo le številke vozlišč 1, 3 in 5. Množice vozlišč z opisano lastnostjo tvorijo *preiskovalno ekvivalentno particijo* grafa [9, 10]. V prihodnosti bi bilo potemtakem smiselno raziskati, kako preiskovalna ekvivalenca vpliva na psevdokanonični zapis in ali jo lahko izkoristimo za dosego večje učinkovitosti algoritma ali robustnejših psevdokanoničnih predstavitev.

Literatura

[1] L. Babai, E. M. Luks, "Canonical labeling of graphs," *Symposium on Theory of Computing (STOC) 1983, Boston, Massachusetts, ZDA*, str. 171–183, 1983.

[2] D. Cook, L. Holder, *Mining Graph Data*. John Wiley & Sons, 2006.

[3] M. Kuramochi, G. Karypis, "Finding frequent patterns in a large sparse graph," *Data Mining and Knowledge Discovery*, vol. 11, št. 3, str. 243–271, 2005.

[4] D. Weininger, "SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, št. 1, str. 31–36, 1988.

[5] L. Babai, "Graph isomorphism in quasipolynomial time [extended abstract]," *Symposium on Theory of Computing (STOC) 2016, Cambridge, Massachussetts, ZDA*, str. 684–697, 2016.

[6] L. Fürst, M. Mernik, V. Mahnič, "Graph grammar induction as a parser-controlled heuristic search process," *Applications of Graph Transformations with Industrial Relevance (AGTIVE) 2011, Budimpešta, Madžarska*, LNCS vol. 7233, str. 121–136, 2012.

[7] J. Kunegis, "KONECT: the Koblenz network collection," *International Web Observatory Workshop (WWW) 2013, Rio de Janeiro, Brazilija*, str. 1343–1350, 2013. `http://konect.uni-koblenz.de/`

[8] J. Leskovec, A. Krevl, "SNAP Datasets: Stanford large network dataset collection." `http://snap.stanford.edu/data`, 2014.

[9] J. Mihelič, L. Fürst, U. Čibej, "Exploratory equivalence in graphs: definition and algorithms," in *Federated Conference on Computer Science and Information Systems (FedCSIS) 2014, Varšava, Poljska*, str. 447–456, 2014.

[10] L. Fürst, U. Čibej, J. Mihelič, "Maximum exploratory equivalence in trees," in *Federated Conference on Computer Science and Information Systems (FedCSIS) 2015, Łódź, Poljska*, str. 507–518, 2015.

Adaptive Diversity-Control Mechanism Used in Popular Nature-Inspired Algorithms to Tackle the Real-World Problems

Petr Bujok[1], **Radka Poláková**[2]

University of Ostrava,
[1]*Faculty of Science, Department of Informatics and Computers,*
[2]*Centre of Excellence IT4Innovations, Institute for Research and Applications of Fuzzy Modeling,*
Czech Republic
{*petr.bujok, radka.polakova*}*@osu.cz*

Abstract

A recently proposed adaptive mechanism for population size is used to control a level of population diversity. The mechanism is employed in seven popular population-based optimisation algorithms. The experiments are carried out on a set of real-world problems. Results show that the mechanism increases the efficiency of some algorithms and even decreases their time complexity.

1 Introduction

Many researchers used simple popular methods to solve various optimisation problems. Recent results showed that some popular optimisation methods perform poorly in the case of real-world problems [1, 2]. The goal of this study is to show how new adaptive technique of population' diversity-control influences performance of several popular population-based optimisers. A set of real-world problems of CEC 2011 is used as a benchmark [3].

2 Diversity-based Adaptation of Population Size

An adaptive mechanism which adapts population size was proposed in 2017 [4]. Although the mechanism was originally developed for differential evolution (DE) algorithm, a population diversity control is possible to be employed by any population-based optimisation method. In a case of a large population diversity, the evolutionary process has bigger power to create new point/s in different areas of search space than in a case of lower diversity. By adding a randomly located point in the search space into the population, the diversity can be increased and the requirement on the amount of function evaluations increases slightly. In the mechanism, the population diversity is measured by *DI* characteristic defined by the expression:

$$DI = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D}(x_{ij} - \bar{x}_j)^2}\,, \qquad (1)$$

where $\bar{x}_j$ is the mean of jth coordinate of the points in the current population, N is current size of population,

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}\,, \qquad (2)$$

DI is square root of average square of distance between a population point and centroid $\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_D)$.

When the initial population is created, the diversity is measured and labelled as DI_{init}. Then, it is used as reference value in the definition of relative measure *RD* of the diversity in the current generation of population (iteration of cycle in an optimisation algorithm) (3),

$$RD = \frac{DI}{DI_{\text{init}}}\,. \qquad (3)$$

The size of population changes dependently on current relative diversity. It is suggested to keep the relative diversity *RD* near its required value *rRD* linearly decreasing from the value 1 at the beginning of the search to value near 0 at the end of the search. It means to change the population size only when the current relative diversity is less than $0.9 \times rRD$ or larger than $1.1 \times rRD$. It is suggested for first nine tenths of search, the zero value is strictly required in the last tenth.

The *RD* characteristic is computed after each generation. The size of population is increased by 1 and a random point from search space is added when the *RD* is less than $0.9 \times rRD$. The population size is decreased by 1 and the worst point is excluded when the *RD* is larger than $1.1 \times rRD$. The minimal and maximal values of population size for the approach are $N_{min} = max[5, D+1]$, $N_{max} = 5 \times D$. The search process starts with $N_{init} = 50$ as was suggested and also used in our experiments in this paper (as in [4]).

3 Algorithms in Experiments

The ABC (Artificial Bee Colony) algorithm proposed by Karaboga in 2005 [5] models the behaviour of a bee swarm, which consists of three types of bees - employed bee, onlookers bee, and scout. An employed bee's $\boldsymbol{x}_i$ jth position is updated by $\boldsymbol{y}_{i,j} = \boldsymbol{x}_{i,j} + U(-1,1) \times (\boldsymbol{x}_{i,j} - \boldsymbol{x}_{r,j})$, where j is randomly selected index from $\{1, 2, \ldots, D\}$ of the position to be updated (D is the dimension of the problem), r is randomly selected bee different from current $\boldsymbol{x}_i$ and $U(-1,1)$ is a random number from the uniform distribution with parameters given in parentheses. Then greedy selection between $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ proceeds. If the food source of an employed bee is depleted (the objective function is not improved for more than *limit* trials), the bee becomes the scout bee.

African Buffalo optimisation (ABO) was introduced in 2015 [6]. It was inspired by the behavior of African buffaloes tracking the rainy seasons in different parts of Africa. Parameters of ABO are set to the recommended values: learning parameters $lp_1 = 0.7, lp_2 = 0.5$. These parameters enable the movement of the buffaloes towards greater exploitation or exploration.

The Bat algorithm simulates an echolocation behavior of real bats controlled by emission rate and loudness. The artificial representation of bats uses parameter setting, which follows the publication of Yang [7]. Maximal and minimal frequencies are set up $fr_{max} = 2$ and $fr_{min} = 0$, respectively. A local-search *loudness* parameter is initialised $A_i = 1.2$ for each bat-individual and reduced if a new bat position is better than the old one using coefficient $\alpha = 0.9$. The *emission rate* parameter is initialised for each bat-individual $r_i = 0.1$ and increased by parameter $\gamma = 0.9$ in the case of a successful offspring.

The Cuckoo (Cuckoo search) algorithm was inspired by cuckoo birds 'nest-parasitism' and it was introduced by Yang and Deb in 2009 [8]. There are N nests in population, in each generation, a new solution is generated by Lévy flight random walk. A p_a-fraction of nests is abandoned and new nests are generated, the best nests survive into next generation. Probability of the cuckoo eggs' laid in a bird-host nest is $p_a = 0.25$ and the control parameter of Lévy flight random walk is $\lambda = 1.5$ in our study.

The Firefly algorithm models the life of a swarm of fireflies in a simplified form. The most important thing of the swarm's life for the inner operation is the situation how a firefly attracts another firefly. The artificial model of fireflies' attraction has several control parameters, randomisation parameter $\alpha = 0.5$, *light absorption* coefficient is set as $\gamma = 1$, and *attractiveness* is updated using its initial $\beta_0 = 1$ and minimal $\beta_{min} = 0.2$ values. The algorithm was proposed by Yang in 2007 [9].

From the family of the algorithms modelling the life of plants, the Flower algorithm (Flower Pollination Algorithm for Global Optimisation, 2012) [10] was selected. Initial positions of all points are updated by either global or local pollination. The global pollination is done by using step vector $\boldsymbol{L}$, which obeys a Lévy distribution with probability $p = 0.8$. If the new point is better than the original one, the point replaces the original into the population. A parameter controlling Lévy distribution is set up $\lambda = 1.5$ as in the Cuckoo algorithm.

Particle swarm optimisation (PSO) proposed in 1995 belongs to very popular and studied nature-inspired algorithms [11]. The basic variant of PSO with slightly enhanced particles' velocities updating by the variation coefficient w and coefficient c is used in this experimental comparison. The control parameter of variation w is set for each generation as a linear interpolation from maximal value $w_{\max} = 1$ to $w_{\min} = 0.3$. Parameter controlling a local and a global part of the velocity updating is set $c = 1.05$. Velocity is updated by $\boldsymbol{v}_{i,G+1} = w_{G+1}\, \boldsymbol{v}_{i,G} + c\, U(0,1)\, (\boldsymbol{p}_{\rm best} - \boldsymbol{x}_i) + c\, U(0,1)\, (\boldsymbol{g}_{\rm best} - \boldsymbol{x}_i)$, where G denotes generation, $U(0,1)$ is random number generated from uniform distribution with parameters given in parentheses, $\boldsymbol{x}_i$ is current particle, $\boldsymbol{p}_{\rm best}$ is up-to-now best historical position of the current particle, and $\boldsymbol{g}_{\rm best}$ is the best particle in swarm history.

4 Experiments and Results

The seven popular optimisation algorithms described in section 3 were tested on test set CEC 2011 [3], which consists of 22 real-world problems in Special Session on Real-Parameter Numerical optimisation. Different computational complexity and the dimensionality of the search space which varies from $D = 1$ to $D = 240$ are for the objective functions in the benchmark. The functions are described in [3] in detail, including the experimental settings required for the competition.

The diversity-based adaptive mechanism for population size described in section 2 was implemented into the seven popular optimisation algorithms described in section 3. Obtained algorithms, namely ABCdiv, ABOdiv, BatDiv, CuckooDiv, FireflyDiv, FlowerDiv, and PSOdiv, were tested under the same conditions as the original ones.

For each algorithm and problem, 25 independent runs were carried out. The run stops if the prescribed number of function evaluation *MaxFES* $= 150,000$ is reached. The partial results of the algorithms after reaching one third and two thirds of *MaxFES* were also recorded. The point in the terminal population with the smallest function value is the solution of the problem found in the run.

Table 1: Counts of wins, loses and similar results of proposed diversity-control mechanism from Wilcoxon rank-sum test.

Alg.	wins	equal	loses
ABC	3	15	4
ABO	2	16	4
Bat	5	13	4
Cuckoo	9	6	7
Firefly	7	12	3
Flower	13	7	2
PSO	5	16	1
%	28.6	55.2	16.2

The population size $N = 90$ was used in all the original algorithms. The other control parameters are described in section 3 and are set as recommended by authors in the original papers.

In this experiment, there are 14 algorithms and therefore only selected results are presented. At first, Wilcoxon rank-sum test was applied to compare the pairs of the original and the diversity-control variant. The counts of significant wins of proposed variants, the count of problems, where both variants perform similarly, and the counts of loses of the algorithms with proposed mechanism are presented in Table 1. The proposed mechanism increases efficiency in more than 28 % of problems and performs worse in 16 % of problems. The best results of the mechanism are achieved in Flower (FlwDiv) and the worst in ABO algorithm.

Table 2: Counts of best and shared best, worst and shared worst positions from Kruskal-Wallis and Dunn's tests.

position	FlwDiv	CkoDiv	Flower	ABC
best	11	10	10	8
worst	0	0	1	1
PSOdiv	ABCdiv	Cuckoo	PSO	ABO
7	7	6	4	0
0	1	0	0	4
ABOdiv	BatDiv	Bat	Firefly	FflDiv
0	0	0	0	0
4	12	12	14	16

If we compare these results with the results of previous study [12] where a set of CEC 2014 benchmark problems was used, the efficiency of mechanism in ABC algorithm achieved in earlier study was better. A ratio of number of wins and loses shows ability of newly proposed diversity control mechanism to cope with real-world problems. A natural (original) diversity level of the algorithms with lower ratio (especially ABO and ABC) is sufficient. A further control of diversity level does not increase capability of these algorithms to solve real-world problems.

For a better insight into all 14 algorithms efficiency solving each problem separately, Kruskal-Wallis test with Dunn's multiple comparison methods are applied. The zero hypothesis about algorithms' performance was rejected in each problem at significance level 0.01. In Table 3, the best and the worst algorithms are depicted, and in Table 2, the counts of first and the last position of all algorithms are presented. Algorithms in Table 2 are ordered from better to worse performing. Two best positions are for the diversity-control mechanism (Flower and Cuckoo algorithms). The worst positions are for Firefly and Bat algorithms and its diversity-control versions (FflDiv and BatDiv).

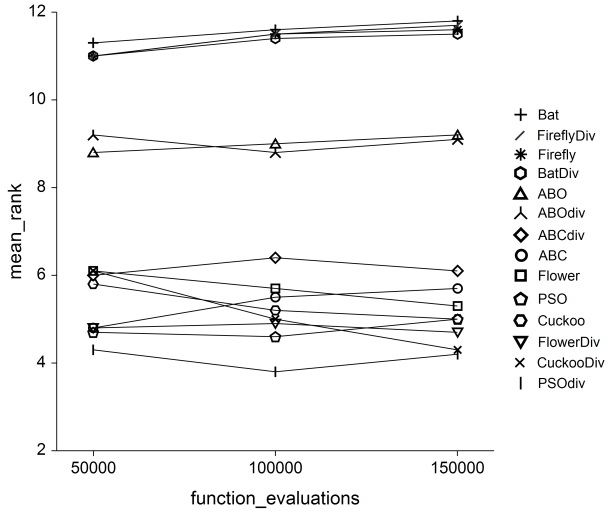


Figure 1: Mean ranks of all algorithms from Friedman test.

Global insight to overall algorithms performance provides Friedman statistical test. The zero hypothesis about equality of algorithms' results was rejected in each stage of the run, significance level was set as 0.01. For better illustration, the mean ranks are illustrated in the Figure 1. It is clear that the version with the adaptive population size mechanism is better than its original version for all algorithms except ABC and Firefly. The best performing algorithms employ diversity-control mechanism: PSOdiv, CuckooDiv, and FlowerDiv. There are visible three 'groups' of algorithms: the worst performing Bat, Firefly and its diversity-control versions; the middle ABO and its new version; and remaining-better performing methods. Although PSOdiv seems the best performing algorithm, the CuckooDiv variant seems to be better optimiser when increasing time.

Generally, from the results of Friedman and Kruskal-Wallis tests we can decided that for real-world problems, PSO, Cuckoo, Flower, and ABC algorithms provide better results.

One expected that adding of the new mechanism into the existing method increases the time complexity. We estimate the time complexity of original and new versions of algorithms by using three functions with various dimensionality $F1$ $(D = 6), F2$ $(D = 30)$, and $F12$ $(D = 240)$. Results show that new mechanism substantially decreases time complexity in all selected problems of ABC, Bat, Flower, and PSO, in ABO the decreasing appears for $D = 30, 240$. Proposed mechanism decreases the complexity of Cuckoo for $D = 240$, and in the case of Firefly for $D = 6, 30$. Minimal percentage proportion of time necessary to achieve the results is 19% (Firefly, $F1$), maximal is 235% (Firefly, $F12$), and average time reduction by mechanism is 67%.

5 Conclusions

An adaptive diversity-based mechanism for population size setting was implemented into seven popular optimisation algorithms. Experiments were done on a benchmark set of real-world problems CEC 2011. Overall, the new modifications of the algorithms behaved more effectively than original versions in more than 28 % out of tested cases, whilst the original versions of the algorithms outperformed the new versions in about 16 % out of tested cases. Only the Firefly and ABC algorithms' results were deteriorated by the implementation of the new adaptive mechanism. The largest improvement appeared for the Flower and PSO algorithms in our experiments. Regarding the estimated time complexity of using of the proposed mechanism, new approach decreases time demands of algorithms substantially in most of the selected problems. Finally, in a case of Flower and PSO algorithms, diversity-control mechanism increase efficiency and decrease the time demands of the optimisation process. Further research will be focused on study and tuning of mechanism' parameters.

Table 3: Shared best and shared worst positions from Kruskal-Wallis and Dunn's tests.

F	D	best	worst
T01	6	Cuckoo,CuckooDiv,Flower	BatDiv,Bat,FireflyDiv
T02	30	ABCdiv,ABC,PSOdiv	Bat,BatDiv,Flower
T03	1	all except→	Firefly,FireflyDiv,ABO,ABOdiv
T04	1	many algorithms	FireflyDiv
T05	30	ABC,ABCdiv,CuckooDiv	BatDiv,Bat,FireflyDiv
T06	30	CuckooDiv,FlowerDiv,PSOdiv,ABC	Bat,Firefly,FireflyDiv,BatDiv
T07	20	CuckooDiv,Cuckoo,FlowerDiv,Flower	BatDiv,Bat,Firefly,FireflyDiv
T08	7	many algorithms	BatDiv,FireflyDiv
T09	126	ABCdiv,ABC	Firefly,FireflyDiv
T10	12	CuckooDiv,Cuckoo	BatDiv,FireflyDiv,Firefly Bat,ABOdiv,ABO
T11.1	120	ABCdiv,ABC,Cuckoo	Firefly,FireflyDiv
T11.2	240	ABCdiv,ABC,Cuckoo	Bat,BatDiv,Firefly
T11.3	6	FlowerDiv,Flower,CuckooDiv	Bat,Firefly,FireflyDiv
T11.4	13	FlowerDiv,Flower,CuckooDiv	ABCdiv,ABO,ABC,ABOdiv
T11.5	15	FlowerDiv,Flower,CuckooDiv	BatDiv,Bat,FireflyDiv,Firefly
T11.6	40	FlowerDiv,Flower	Firefly,FireflyDiv,ABO
T11.7	140	FlowerDiv,Flower,PSOdiv	Firefly,ABOdiv
T11.8	96	PSOdiv,FlowerDiv,PSO,Flower	Firefly,FireflyDiv,Bat,BatDiv
T11.9	96	FlowerDiv,PSO,PSOdiv,Flower	Firefly,FireflyDiv
T11.10	96	FlowerDiv,PSOdiv,PSO,Flower	Firefly,FireflyDiv
T12	26	PSOdiv,ABC,PSO,CuckooDiv ABCdiv,FlowerDiv,Cuckoo	BatDiv,Bat
T13	22	ABCdiv,ABC,CuckooDiv	BatDiv,Bat

References

[1] P. Bujok, J. Tvrdík, and R. Poláková, "Nature-inspired algorithms in real-world optimization problems," *MENDEL Soft Computing Journal*, vol. 23, pp. 7–14, 2017.

[2] ——, "Adaptive differential evolution vs. nature-inspired algorithms: An experimental comparison," in *2017 IEEE Symposium Series on Computational Intelligence (IEEE SSCI)*, 2017, pp. 2604–2611.

[3] S. Das and P. N. Suganthan, "Problem definitions and evaluation criteria for CEC 2011 competition on testing evolutionary algorithms on real world optimization problems," Jadavpur University, India and Nanyang Technological University, Singapore, Tech. Rep., 2010.

[4] R. Poláková, J. Tvrdík, and P. Bujok, "Adaptation of population size according to current population diversity in differential evolution," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI) Proceedings*. IEEE, 2017, pp. 2627–2634.

[5] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Erciyes University, Kayseri, Turkey, Technical Report-TR06, 2005.

[6] J. B. Odili, M. N. M. Kahar, and S. Anwar, "African buffalo optimization: A swarm-intelligence technique," *Procedia Computer Science*, vol. 76, pp. 443–448, 2015.

[7] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nicso 2010: Nature Inspired Cooperative Strategies for Optimization*, ser. Studies in Computational Intelligence, J. Gonzalez, D. Pelta, C. Cruz, G. Terrazas, and N. Krasnogor, Eds., vol. 284, 2010, pp. 65–74.

[8] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *2009 World Congress on Nature Biologically Inspired Computing NaBIC*, 2009, pp. 210–214.

[9] X.-S. Yang, *Nature-Inspired Optimization Algorithms*. Elsevier, 2014.

[10] ——, "Flower pollination algorithm for global optimization," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7445 LNCS, pp. 240–249, 2012.

[11] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *1995 IEEE International Conference on Neural Networks Proceedings, Vols 1-6*. IEEE, Neural Networks Council, 1995, pp. 1942–1948.

[12] R. Poláková and P. Bujok, "Popular optimisation algorithms with diversity-based adaptive mechanism for population size," in *MENDEL 2018: Recent Advances in Soft Computing*, R. Matousek, Ed. Springer-Verlag, 2018, *In press*.

Spletno upodabljanje LiDAR podatkov Slovenije z dodano barvno informacijo in senčenjem

Jaka Kordež, Ciril Bohak

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
E-pošta: jk5456@student.uni-lj.si, ciril.bohak@fri.uni-lj.si

Web-Based visualization of Slovenian LiDAR data with added color information and shading

In this paper we present an adaptation of web-based point-cloud data visualization framework Potree for use with Slovenian LiDAR data. We enriched LiDAR data with color information obtained from Ortho-Photo data. We supplemented the dataset with artificial points, where data was missing (water surfaces). For better spatial perception we also calculate the normals for individual points in the data and use them for shading. We evaluate the presented sistem step by step and present the final results.

Povzetek

V članku predstavimo prilagoditev spletnega ogrodja Potree za prikaz podatkov v obliki oblakov točk za uporabo s slovenskimi podatki LiDAR. V podatke smo dodali barvno informacijo iz zbirke orto-foto posnetkov. Podatke smo dopolnili z umetno ustvarjenimi točkami, kjer so bila manjkajoča območja (vodne površine). Za boljšo prostorsko predstavo pri pregledovanju podatkov smo za posamezne točke v zbirki izračunali normale, ki jih uporabimo pri izračunu osvetlitve. Končni sistem smo korak za korakom performančno ovrednotili in predstavili končne rezultate.

1 Uvod

V zadnjem času se je zaradi napredka tehnologije zelo povečala količina zajetih okoljskih podatkov na različnih področjih. To velja tudi za področje geodezije, geografije in geologije, kjer si raziskovalci vse bolj pogosto pri snemanju zemeljskega površja pomagajo tudi s tehnologijo LiDAR (angl. Light Imaging, Detection and Ranging) [4]. Z uporabo senzorjev LiDAR, nameščenih na letala, lahko pridobimo izredno natančno informacijo o obliki površja, kot tudi o vrsti podlage od katere se laserski žarek odbije. Zbirko takšnih podatkov so pripravili tudi pri Agenciji Republike Slovenije za okolje [1].

Uporabnost takšne zbirke pa se močno poveča kadar lahko uporabniku njeno vsebino predstavimo na enostaven in prijazen način. Eden takšnih pristopov je vizualizacije takšne zbirke podatkov. Za široko dostopnost je dandanes smiselno uporabiti spletne tehnologije, saj so podprte na širokem naboru naprav (od namiznih računalnikov pa do mobilnih telefonov). Vizualizacija takšne zbirke podatkov v obliki kock z robom 1 m z uporabo upodobljevalne metode sledenja žarkom je predstavljena v delu [6], kjer pa sistem ne omogoča interaktivnega vpogleda zaradi zahtevnega izračuna upodobitve. Za vizualizacijo takšnih podatkov je zaradi interaktivnosti smiselno uporabiti strojno pospešen izris. To nam omogoča uporaba vmesnika WebGL [5], ki ga uporablja tudi ogrodje Potree [1] [7], namenjeno prav vizualizaciji oblakov točk na spletu.

Ogrodje Potree je namenjeno prikazu podatkov, ki imajo točke razporejene po širšem 3D prostoru in za optimizacijo prikaza podatkov uporablja strukturo osmiškega drevesa (angl. octree). Za primer podatkov v že omenjeni zbirki velja, da so točke večinoma razporejene v tankem pasu v bližini površja in je posledično uporaba takšne podatkovne strukture nepotrebna in celo potratna, saj delitev prostora v tretji dimenziji ni potrebna. Prav to pa naslavlja v svojem delu Jan Gašperlin [3], ki je ogrodje Potree za namene prikaza površja prilagodil z implementacijo štiriškega drevesa za delitev prostora, kar se izkaže za dobro odločitev, tako v performančnem smislu kot pri prihranku prostora.

Zbirka podatkov LiDAR za posamezno točko vsebuje podatke o položaju, intenziteti odboja, zaporedni številki odboja in druge. Za dobro vizualizacijo površja pa bi bilo smiselno za posamezno točko hraniti tudi barvno informacijo. Ta problem je v svojem delu naslovil Matej Slemenik [8], ki je točkam pripisal barvo na podlagi poravnanih barvnih orto-foto posnetkov. Še boljšo vizualizacijo pa lahko dosežemo z uporabo primerne metode za senčenje. Pri tem nam močno pomaga informacija o normalah, ki definira odboj svetlobe. To problematiko je v svojem delu naslovil Peter Fajdiga [2], ki za izračun normal v posamezni točki zbirke uporabi metodo analize glavnih komponent (angl. Principal Component Analysis - PCA) sosednjih točk. V našem sistemu to izkoristimo za možnost izračuna osvetlitve glede na položaj sonca, izbranega datuma in ure.

Za še boljši končni izgled vizualiziranih podatkov pa je smiselno nasloviti tudi napake oz. pomanjkljivosti v podatkih. Tako v pričujočem delu predstavimo tudi metodo za dopolnitev podatkov na območjih vodnih površin, kjer zaradi lastnosti površine površja senzorji podatkov niso zajeli. Za določanje vodnih površin uporabimo po-

[1] http://potree.org/

datke iz Atlasa voda Ministrstva za okolja in prostor RS [2].

V članku v naslednjem poglavju predstavimo razviti sistem in njegove posamezne korake, v poglavju 3 razvit sistem performančno ovrednotimo ter predstavimo rezultate, na koncu pa v poglavju 4 podamo zaključke in možne korake za nadaljnje delo.

2 Spletni sistem za vizualizacijo terena

Celoten sistem je sestavljen iz dveh delov: (a) dela za pripravo podatkov in (b) uporabniškega dela za vizualizacijo podatkov. Prvi del teče na strežniku, kjer se podatki iterativno pripravljajo v obliko primerno za vizualizacijo. Tako pripravljeni podatki so nato na voljo uporabniškemu delu sistema namenjenega vizualizaciji. Uporabniški del je razvit kot spletna aplikacija.

2.1 Priprava podatkov

Območje celotne Slovenije razdelimo v pakete, ki jih sestavljajo kvadrati s stranico dolžine 1 km tako, da so koordinate poravnane z Gauß-Krüger-jevim koordinatnim sistemom. Da bi zagotovili lažji nadzor nad procesiranjem podatkov, je sistem sestavljen iz posameznih stopenj, obdelava podatkov pa poteka cevovodno. Vsak paket mora skozi naslednje stopnje:

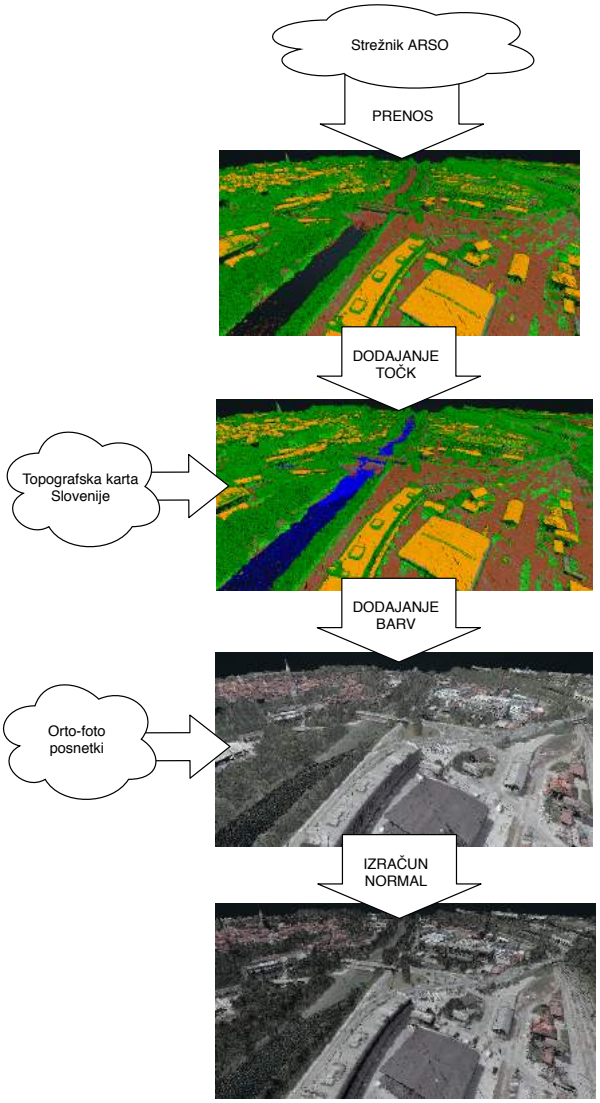
- *Prenos* - oblak točk se prenese iz strežnika.
- *Dodajanje točk* - dodajo se manjkajoče točke na vodnih površinah.
- *Dodajanje barve* - vsaka točka dobi barvno informacijo iz ustreznege slikovne pike na zračne posnetku površja.
- *Izračun normal* - izračun normal z metodo PCA.

Vsaka od stopenj sistema (prikazanega na sliki 1) se izvaja kot ločen proces in ima določeno vhodno in izhodno datoteko. Z vsem skupaj upravlja glavni program, ki zagotavlja tudi ponavljanje procesiranja v primeru napak. Za čim večjo učinkovitost se več paketov obdeluje sočasno. Vsaka od stopenj je podrobneje opisana v naslednjih podpoglavjih.

2.1.1 *Dodajanje manjkajočih točk*

S senzorjem LiDAR zaznavamo okolico na podlagi odboja laserskih žarkov. To pomeni, da se mora vsaj del žarka ob stiku s površino tudi odbiti nazaj proti senzorju. Tehnika deluje zelo dobro na trdnih površinah, na vodi pa zelo redko pride do ustreznega odboja. Rezultat so manjkajoče točke na jezerih in rekah. Da bi popravili oblake točk na teh območjih moramo tam točke torej naknadno dodati.

Da bi točke dodali na pravo mesto potrebujemo podatke o vodnih površinah. V pomoč nam je topografska karta Slovenije, ki je v vektorski digitalni obliki dostopna na strežniku Agencije Republike Slovenije za okolje[3]. Karta vsebuje tudi sloj na katerem so vodne površine



Slika 1: Prikaz zaporednih korakov cevovodnega delovanja sistema za pripravo podatkov.

opisane v obliki dvodimenzionalnih poligonov. Vsak ima pripisano geografsko ime, širino, tip in še druge podatke.

Algoritem za vsak paket najprej prenese vse vodne površine za to območje v obliki JSON. Vsaka površina je predstavljena z množico točk, ki določajo poligon. Iz oblaka točk se vsakemu poligonu dodajo točke, ki glede na tloris ležijo znotraj poligona in so označene kot tla. Točke, ki določajo poligon so predstavljene z dvema koordinatama in nimajo nadmorske višine. V naslednjem koraku se zato vsaki priredi višina izračunana kot povprečje najbližjih nekaj točk iz izbranega dela oblaka točk. Da bi zapolnil manjkajoče vrzeli, program izriše navidezno mrežo točk preko celotnega poligona z razmakom 1 m. V oblak se doda le tiste točke, ki so vsaj 1 m oddaljene od že obstoječih. Pred dodajanjem se višina določi kot povprečje višin najbližjih robnih točk poligona, koordinati X in Y pa se naključno zamakneta da dobimo bolj naravno porazdelitev.

[2] http://www.evode.gov.si/atlas-voda/

[3] http://gis.arso.gov.si/arcgis/rest/services/Topografske_karte_ARSO_nova/MapServer

2.1.2 *Dodajanje barv*

Oblaki točk v zbirki LiDAR Ministrstva za okolje in prostor nimajo dodane informacije o barvi. To informacijo v podatke dodamo s pomočjo letalskih orto-foto posnetkov površja. Algoritem prenese ustrezen posnetek iz strežnika Agencije RS za okolje[4] in vsaki točki v oblaku doda barvo najbližjega slikovnega elementa iz orto-foto posnetka.

2.1.3 *Senčenje*

Oblak točk sam po sebi ne definira nobenih poligonov. Vsaka površina je zato predstavljena z nizom točk, ki ležijo na njeni ravnini. Da bi lahko model pravilno osenčili je potrebno določiti normale za vsako točko. To najlažje storimo z uporabo metode PCA. Za vsako točko iz oblaka izberemo vse sosednje točke v določenem radiju in izračunamo skupno kovariančno matriko po enačbi (1), kjer je k število sosednjih točk, p_i je i-ta sosednja točka in P pa točka normale.

$$C = \sum_{i=1}^{k} (p_i - P)(p_i - P)^T \tag{1}$$

Nad to matriko se izvede razcep na lastne vrednosti in lastne vektorje. Ker so vektorji in matrika tridimenzionalni dobimo 3 lastne vrednosti in 3 lastne vektorje, ki so med seboj ortogonalni. Normala je tisti lastni vektor, katerega lastna vrednost je najmanjša, saj najmanj sovpada z sosednjimi točkami.

2.2 Vizualizacija podatkov

Uporabniški del aplikacije je razvit kot spletna aplikacija in za prikaz uporablja prilagojeno različico že omenjenega ogrodja Potree. Ogrodje Potree smo prilagodili tako, da namesto osmiške delitve prostora zaradi učinkovitosti uporablja štiriško delitev.

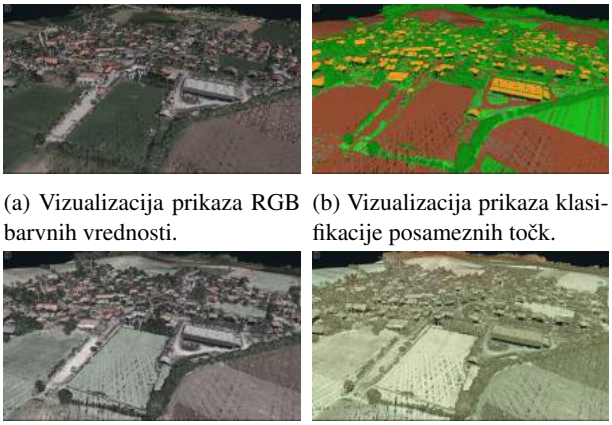
Na strežniku pripravljeni podatki izbranega območja se pretočijo k uporabniku, kjer se uporabniku prikažejo. Izris je strojno pospešen in izrablja vmesnik WebGL. Spletna aplikacija omogoča prikaz različnih lastnosti terena shranjenih v podatkih, podpira pa tudi njihovo združevanje. Nekaj primerov vizualizacije istih podatkov je prikazanih na sliki 2.

3 Evalvacija in rezultati

Da bi ocenili hitrost delovanja celotnega sistema smo poskusno pognali celoten sistem in merili čas izračuna. V tabeli 1 so prikazani povprečni izmerjeni časi posameznih stopenj cevovoda. Izmerjeni so bili na vzorcu 12 različnih datotek. Povprečno število točk na datoteko je bilo 14.829.274. Celotna obdelava paketov je v povprečju trajala 10 minut in 29 sekund. To pomeni, da je bilo na sekundo obdelanih 22.998 točk. Testiranje je potekalo na procesorju AMD Ryzen 1700 pri frekvenci 3,1 GHz.

Ker vsaka datoteka pokriva območje 1 km^2, površina Slovenije pa je nekaj več kot 20.000 km^2, pričakujemo, da bo celotno območje obdelano v približno 3605 urah.

[4] http://gis.arso.gov.si/arcgis/rest/services/DOF_2016/MapServer



(a) Vizualizacija prikaza RGB barvnih vrednosti.

(b) Vizualizacija prikaza klasifikacije posameznih točk.

(c) Vizualizacija prikaza kombinacije RGB barvnih vrednosti in intenzitete odboja laserskega žarka.

(d) Vizualizacija prikaza kombinacije intenzitete odboja laserskega žarka in višine posamezne točke.

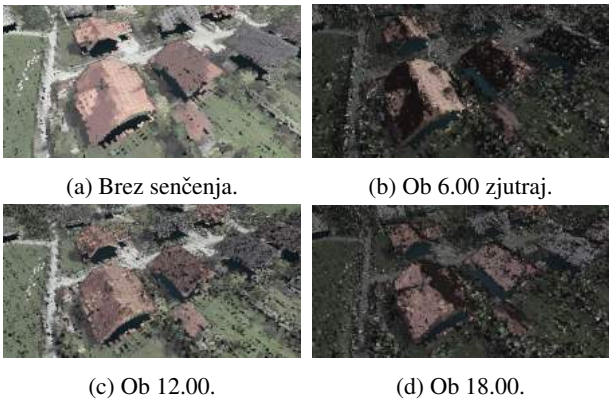
Slika 2: Prikaz različnih lastnosti shranjenih za posamezno točko in njihove kombinacije.

Tabela 1: Povprečni časi stopenj cevovoda

KORAK	ČAS
Prenos	35 s
Dodajanje točk	46 s
Dodajanje barv	47 s
Izračun normal	8 min 42 s
Skupaj	**10 min 49 s**

Program lahko izkoristi več jeder sistema s tem, da obdeluje več datotek hkrati. Na ta način lahko z uporabo npr. 8 jeder zaključimo v približno 450 urah oz. v 18 dnevih in 18 urah.

Z izračunanimi normalami lahko v koraku senčenja izračunamo osvetlitev v različnih delih dneva. Na sliki 3 je tako prikazan del podatkov brez izračuna senčenja in nato še z izračunom senčenja ob treh delih dneva: zjutraj, opoldne in zvečer.



(a) Brez senčenja.

(b) Ob 6.00 zjutraj.

(c) Ob 12.00.

(d) Ob 18.00.

Slika 3: Senčenje terena ob različnih urah v dnevu.

Na slikah 4a - 4d je razviden končni rezultat vizualizacije podatkov obdelanih s predstavljenim sistemom za izbrana področja Slovenije.



(a) Pogled na Ljubljano.



(b) Pogled na Ljubljano.



(c) Pogled na Piran.



(d) Pogled na Triglav.

Slika 4: Prikaz končnih upodobitev podatkov pripravljenih in vizualiziranih s predstavljenim sistemom.

Da bi ugotovili kolikšno dodatno breme predstavlja senčenje za grafični pospeševalnik, smo izvedli še performančni test spletnega pregledovalnika. V našem testnem računalniku je Nvidia GT 430, testirali pa smo v brskalniku Firefox 61 na operacijskem sistemu Windows 10. Pregledovalnik je bil nastavljen na upodabljanje enega milijona točk. En oblak točk smo prikazali iz desetih različnih kotov in vsakič zabeležili dobljeno število sličic na sekundo (angl. frames per second - FPS) za prikaz osenčenega in neosenčenega modela. V povprečju je izris neosenčenega modela potekal z 28,66 FPS osenčenega modela pa z 27,5 FPS. Dodatno senčenje tako zateva $4,2\%$ več časa.

4 Zaključki in nadaljnje delo

Predstavljen sistem je bil razvit z namenom izdelave dostopnega orodja za prikaz izbranih geodetskih podatkov za območje Slovenije. Uporabniški del razvitega sistema omogoča vpogled v podatke tako na namiznem računalniku kot na mobilnih telefonih in tablicah. Uporabniku ponuja interaktiven, odziven in intuitiven vmesnik, kjer lahko izbrane parametre prilagaja med samo uporabo.

V okviru nadaljnjega dela želimo s predstavljenim sistemom pripraviti podatke za celotno Slovenijo in jih ponuditi na vpogled širši javnosti. Prav tako želimo sistem nadgraditi z dodatnimi funkcijami in podatke dodatno avtomatsko obdelati kjer prihaja do artefaktov (npr. pri mostovih in jezovih). Sistem je enostavno prilagodljiv in nadgradljiv in upamo, da pritegne širši nabor uporabnikov.

Literatura

[1] Geodetski inštitut slovenije. izvedba laserskega skeniranja slovenije. blok 35 – tehnično poročilo o izdelavi izdelkov. geodetski inštitut slovenije. Technical report, 2015.

[2] Peter Fajdiga. Združevanje ortofora in podatkov LiDAR v oblak točk z informacijo o barvi in normali - Poročilo seminarja. Technical report, University of Ljubljana, Faculty of Computer and Information Science, 2018.

[3] Jan Gašperlin. Web-based visualization of large terrain point cloud dataset - Seminar report. Technical report, University of Ljubljana, Faculty of Computer and Information Science, 2018.

[4] Norbert Haala, Michael Peter, Jens Kremer, and Graham Hunter. Mobile LiDAR Mapping for 3D Point Cloud Collection in Urban Areas — A Performance Test. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37:1119–1127, 2008.

[5] Dean Jackson and Jeff Gilbert. Webgl specification. Technical report, The Khronos Group Inc., 2015.

[6] Miha Lunar, Ciril Bohak, and Matija Marolt. Distributed ray tracing for rendering voxelized lidar geospatial data. In *Zbornik petindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2016, 19. - 21. september 2016, Portorož, Slovenija*, pages 55–58, 2016.

[7] Markus Schütz. Potree: Rendering large point clouds in web browsers. Master's thesis, Institute of Computer Graphics and Algorithms, Vienna University of Technology, Favoritenstrasse 9-11/186, A-1040 Vienna, Austria, 2016.

[8] Matej Slemenik. Združevanje ortofotov in lidar podatkov za izgradnjo oblaka točk z informacijami o barvi in normali - poročilo seminarja. Technical report, University of Ljubljana, Faculty of Computer and Information Science, 2018.

Application of non-linear regression for modeling the kinetics of essential oils extraction

Dragana Radosavljevic
Faculty of Technical Sciences of University in Pristina
Kosovska Mitrovica, Serbia
dragana.radosavljevic@pr.ac.rs

Sinisa Ilic
Faculty of Technical Sciences of University in Pristina
Kosovska Mitrovica, Serbia
sinisa.ilic@pr.ac.rs

Petar Spalevic
Faculty of Technical Sciences of University in Pristina
Kosovska Mitrovica, Serbia
petar.spalevic@pr.ac.rs

Nadica Milenkovic
Faculty of Technical Sciences of University in Pristina
Kosovska Mitrovica, Serbia
nadica.milenkovic@pr.ac.rs

Jelena Todorovic
Faculty of Technical Sciences of University in Pristina
Kosovska Mitrovica, Serbia
jelena.todorovic@pr.ac.rs

***Abstract*— In this paper it has been shown that data processing techniques can be successfully applied to the modelling of physical processes. The kinetics of essential oil hydrodistillation process is modelled with a polynomial-logarithmic model, and nonlinear regression is then applied. Modelling results for various plant materials are presented. Using several different deviation measures, it has been shown that the new polynomial-logarithmic model gives better matching with experimental values compared to the existing, generally accepted kinetic models of this process.**

Keywords — non-linear regression, modelling, essential oil, hydrodistillation.

I. INTRODUCTION

Essential oils (as complex mixtures of compounds) are the main ingredients of aromatic drugs. They have fragrances and other characteristic features of the drugs (plants) from which they are extracted.

Some essential oils in their composition may have a relatively small number of main components, but there are also essential oils that contain more than 300 components [1]. It is almost impossible to artificially make equally complex, but still identical, combination of components that naturally occur in some oil. All synthetically obtained oils can be only good copies.

The greatest importance of the original, natural structure of essential oils is in medicine (for therapeutic purposes) and pharmacy. Furthermore, nowadays essential oils are most commonly used in cosmetics, perfumery, food industry, agri-culture, alcohol and non-alcoholic beverage industry, paints and varnishes industry, etc. [2]. For the production of essential oils from aromatic drugs, four methods are common: hydro-distillation, cold pressing, organic solvent extraction and supercritical extraction. The most commonly used technique for obtaining essential oils is hydro-distillation [3].

Most papers dealing with the separation of essential oils from various plant materials refer to the composition, yield and quality of essential oils. A smaller number of authors study the kinetics of the essential oils extraction process [4].

To describe the kinetics of the essential oils extraction, researchers mainly use models that are based on physical laws. Empirical models can also be used to model the kinetics of the essential oils extraction from various plant materials. They are, as a rules, simpler than models based on physical laws, with relatively high precision and can be used in engineering. Empirical models are based on functional mapping of inputs values into output values without explicit description the process itself. In these models, using mathematical modelling and artificial intelligence algorithms over the available experimental data, optimal models are formed and the errors related to the known data are minimal.

Regression analysis is one of the techniques used to express the relationship between variables and to evaluate the impact of one or more variables on other variables. This is achieved by choosing the appropriate regression equation (can be linear or nonlinear) [5]. If the data being processed are linearly dependent, obtaining of the linear equation parameters is very simple (e.g. by using the smallest squares method). However, for data that are not linearly dependent, the additional most important step is selecting a suitable function (model) with a set of parameters that best fits to experimental data.

In this paper, one such nonlinear model is presented. It describes the extraction of essential oils from various plant materials. The quality of this model is compared with the relevant models most commonly used for modelling the kinetics of essential oils hydrodistillation from various plant materials.

II. MATERIALS AND METHODS

A. Experimental data

The database of experimental data consists of data that in 60 different samples follow the kinetics of essential oils hydro-distillation from the following plant material: celery, cinnamon, lavender, citronella, melissa, thyme, mentha, anis, rosemary, wild marigold, sage, bitter orange, eucalyptus, california fir, spruce, juniper, white pine, douglas fir, basil, fennel, agar wood, oregano, patchouli, candleberry, coriander, mastic, caraway, pepper, ginger, parsley [6-32] (available from the author on request).

B. Hydrodistillation kinetics models

Two most commonly used models that describe the kinetics of the essential oils extraction from plant material are the model Sovova and Aleksovski [33] and the model Milojević et al. [12]. Beside these two models that are based on physical laws, in this work the validity of using the new (empirical) polynomial-logarithmic model is confirmed [4].

Model 1 - model Sovova - Aleksovski

$$\frac{q}{q_\infty} = 1 - f \cdot e^{-k_1 \cdot t} - (1-f) \cdot e^{-k_2 \cdot t} \quad (1)$$

Model 2 - model Milojević

$$\frac{q}{q_\infty} = 1 - (1-b) \cdot e^{-kt} \quad (2)$$

Model 3 – polynomial-logarithmic model

$$\frac{q}{q_\infty} = a \cdot (\log t)^2 + b \cdot \log t + c \quad (3)$$

Modified Model 3

$$q = a_1 \cdot (\log t)^2 + b_1 \cdot \log t + c_1 \quad (4)$$

where:

q - yield (mass fraction of extracted essential oil) at time t

q_∞ - yield of essential oil at time $t \to \infty$,

$f, T_1, T_2, b, k, a, b, c, a_1, b_1, c_1$ - the parameters of the corresponding models that are to be calculated for each sample (take a note that $a_1 = a \cdot q_\infty$, $b_1 = b \cdot q_\infty$, $c_1 = c \cdot q_\infty$).

The idea to make a polynomial-logarithmic model is obtained by observing the experimental data in which the kinetics of the essential oil hydro-distillation process is monitored. It is clear that the amount of extracted essential oil over time can be described by exponential growth (Fig. 1a). Namely, the essential oil extraction process is intensive at the beginning, while at the end of the process runs very slowly (extensively). If, instead of the linear, the logarithmic scale of a time is used (Fig. 1b), it becomes obvious that the dependence of the amount of the extracted essential oil and time in logarithmic axis is polynomial. The small values of a time in a logarithmic axis are "stretched", while large values are condensed without loss of data [4]. By observing most of experimental data, the polynomial of a second order was promising the best results.

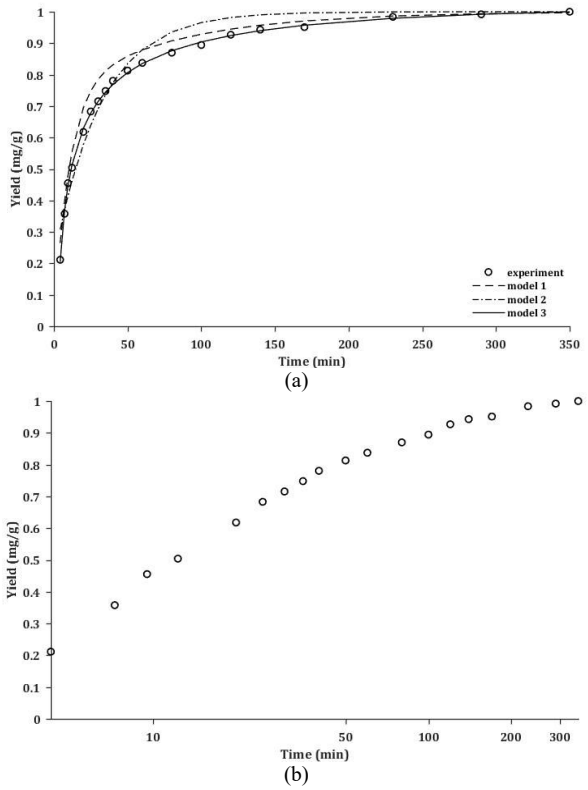


Fig. 1. The change of essential oil yield during the extraction process in (a) linear scale and (b) logarithmic scale of time (sample 26 [6])

C. Methods of modeling

For data processing the R software is used - integrated software environment for data management and graphical and numerical analysis [34].

A nonlinear method of the least squares based on the Levenberg-Marquardt algorithm is used here. In order to implement aforementioned software and to determine the unknown model parameters using the nonlinear regression method, it is necessary first to define the estimated form (the equation) of the kinetics model and for each sample collect the time (t) and yield of essential oil data (q).

Two experiments are performed: 1) based on the given input data (t and q), the estimated values of the parameters (equation coefficients) in eqs. 1, 2 and 3 of all models are obtained through the finite number of iterations for each sample, and 2) based on first 2/3 of input data values (t and q) the estimated values of parameters in eq. 4 of the new model are calculated and using this model the last 1/3 of yield values (q) are estimated for known t for each sample. The obtained parameters are calculated along with different statistical indicators that assess the quality of the model.

From the equations 1 and 2 it can be seen that the value q_∞ (in practice the value of q in the last moment of t) is needed to calculate parameters of the models. By modifying eq. 3 into eq. 4, it becomes obvious that the value of q_∞ is not needed for calculation of $q(t)$, hence the parameters of the new model can be calculated from the first several (q, t) experimental values and remained values can be then estimated by performing regression.

III. Assessment of Nonlinear Regression Model Parameters and Model Validation

Beside the visual assessment, there are also different quantitative indicators for determining the conformity of the evaluated model with the original data [35].

The estimated parameters of the nonlinear regression model are obtained by calculating the minimum squares of the distance of the experimental points from the points obtained by the corresponding model. For this purpose, the following are counted:

Residual Sum of Squares (RSS) (the lower the better):

$$RSS = \sum_{i=1}^{n} (y_{\exp} - y_{\mathrm{mod}})^2 \quad (5)$$

where: $y_{\exp}$ – the experimental values, a y_{mod} – values obtained by the corresponding mathematical model.

Akaike's Information Criterion (AIC) (again the lower the better) is also good indicator for determining which models with a different number of parameters is better. It represents a relative measure of the model quality, applied to the data set. In many scientific fields this criterion is often used as a key in the choice between competing models. If two bad models are compared, the AIC will only provide information that the model is less bad, but not the information that both models are bad. The statistics of this criterion are based on the function of maximum credibility and the number of parameters of the tested distribution [36]:

$$AIC = 2 \log(L) + 2p \quad (6)$$

where: L - the maximum value of the credibility function and p - the number of parameters.

IV. RESULTS AND DISCUSSION

In order to analyse the efficiency of the model comparison, for each sample from the plant database is drawn a diagram of the essential oil extraction in time. As an illustration of the performed model comparison procedure, in Fig. 1a. estimations by using aforementioned models of a randomly selected sample are shown.

In order to determine the validity of the model, the values of the quantitative indicators RSS and AIC were calculated for all samples. By comparing these quantitative indicators for all three models from the first experiment, it can be seen that model 3 is more efficient than the other two models.

In Table 1 are shown the average values of the quantitative indicators (for all 60 samples). It can be seen that Model 3 gives the best results with the lowest average values of all quantitative indicators of this model.

This conclusion is noticeable also in Figure 2 where graphical representation of the comparison of quantitative indicators of a complete set of analysed data using all three models are shown. (the labels on X axis are serial number of samples, and labels on Y axis represent values of corresponding statistical indicators).

TABLE I. COMPARISON OF QUANTITATIVE MODEL INDICATORS

Model	Comparison	
	Indicator	Average
Model 1	*RSS*	0.016
	AIC	-32.37
Model 2	*RSS*	0.017
	AIC	-31.58
Model 3	*RSS*	0.007
	AIC	-41.693

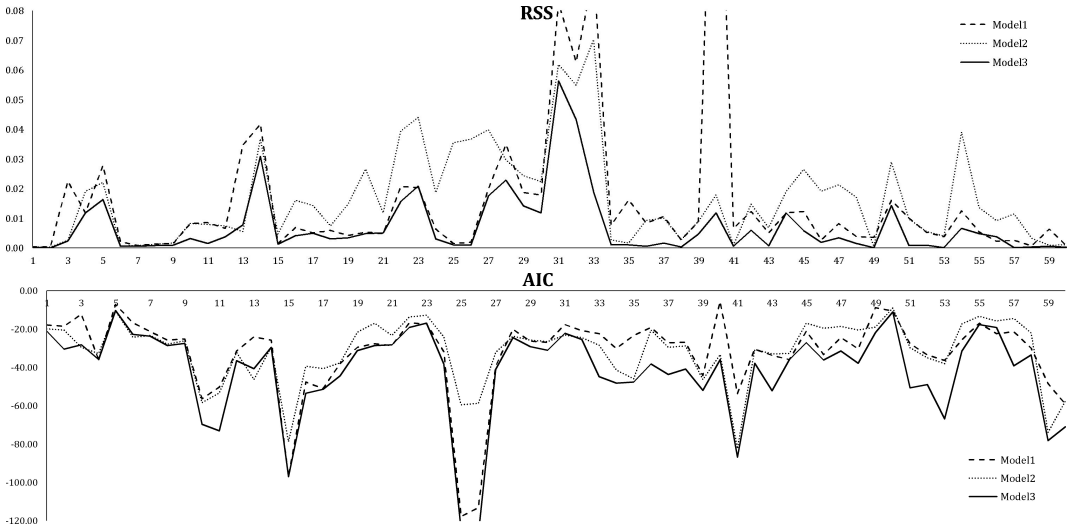


Fig. 2. Graphical representation of Residual Sum of Squares (RSS)and Akaikes's Information Criterion (AIC)

The graphical representation of RSS indicates that in most cases Model 3 gives either the best results, or results closest to the values of the best model in the appropriate sample. Also, from the graphic representation of AIC, it is noted that Model 3 achieves the lowest values. By observing the obtained results (Table I, Fig. 2), it can be seen that all quantitative indicators in the majority of considered samples are the smallest for Model 3 (the lower values of the quantitative indicators indicate a better model).

In the second experiment only samples with more than 10 values are considered (36 of 60 samples). The models based on eq. 4 are obtained (trained) based on first 2/3 values and tested on last 1/3 values of experimental data in time. The residual sums of squares are calculated for training and test data and their averaged and individual values shown in Table 2 and Figure 3.

TABLE II COMPARISON RSS FOR TRAINING AND TEST SET

RSS for modified Model 3 (eq. 4)	
Training set	Test set
0.006	0.013

From the table 2 it can be seen that even estimated values obtained by regression have better RSS (0.013) compared to calculated values from other two models (0.016 and 0.017).

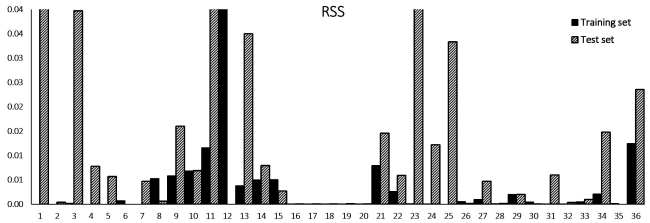


Fig. 3. Graphical representation of Residual Sum of Squares (RSS) for training and test set

CONCLUSION

This paper presents the application of data research techniques for modelling the kinetics of essential oils hydro-distillation process. A nonlinear regression analysis procedure is used to determine a mathematical model based on a well-presumed form of the equation, which faithfully describes the dynamics of the oil hydro-distillation process

from various plant materials. The polynomial-logarithmic model is better than those in the available literature and quite accurately "simulates" the measured values. For the prediction of process results in this work were used artificial intelligence algorithms with the support of program package R. By analysing the values of statistical indicators (RSS and AIC) obtained using the new model and comparing them with the corresponding indicators of the two most widely used models, it is obvious that the proposed model provides better match with the experimentally obtained data.

Using the proposed model, the estimation of yield in time is also performed for selected samples from data in early (initial) phase of the process and again very good results are obtained. The verification and improvement of the model is planned for other similar experiments found on the Internet.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Science and Technological Development of the Republic of Serbia under Projects No.TR-35026 and III47016.

REFERENCES

[1] Marković T. (2011) Etarska ulja i njihova bezbedna primena. Beograd: Institut za proučavanje lekovitog bilja "dr Josif Pančić'

[2] Sovilj, M., Spasojević, M. (2001) Proizvodnja i primena etarskih ulja iz domaćeg lekovitog bilja. Časopis za procesnu tehniku i energetiku u poljoprivredi / PTEP, vol. 5, br. 1-2, str. 34-38

[3] Skala D., Žižović I., Petrović S., Etarska ulja – destilacija, ekstrakcija, izbor tehnologije i kvalitet, Hem. Ind. 53 (1999) 123-139

[4] Radosavljević D., Ilić S., Milojević S., Bojović Ž., Marković M, (2017). Modeling the kinetics of essential oil hydrodistillation from juniper berries (*Juniperus communis* L.) using non-linear regression, Hem. Ind. pp 48-48; doi:10.2298/HEMIND160715048R

[5] Graybill F., Iyer H., (1994), Regression Analysis: Concepts and Applications, Duxbury Press, Belmont

[6] Pavićević V. P., (2016). Bezotpadna tehnologija prerade ploda kleke (Juniperus communis L.), doktorska disertacija, Univerzitet u Beogradu, Tehnološko-metalurški fakultet.

[7] Cassel E., Vargas R. M. F., (2006). Experiments and Modeling of the Cymbopogon winterianus Essential oil Extraction by Steam Distillation, J. Mex. Chem. Soc. 50, 126-129.

[8] Hancı S. S., Sahin S., Yılmaz L., (2003). Isolation of volatile oil from thyme (Thymbra spicata) by steam distillation, Nahrung/Food 47, 252–255.

[9] Sowbhagya H. B., Sampathu S. R., Krishnamurthy N.,(2007). Evaluation of size reduction on the yield and quality of celery seed oil, Journal of Food Engineering 80, 1255–1260.

[10] Morin P., Gunther C., Peyron L., Richard H., (1985). Etude des phénomènes physico-chimiques intervenant lors du procédé d'hydro-distillation, Bulletin de la Société chimique de France 5, 921-930.

[11] Romdhane M., Tizaoui C., (2005). The kinetic modelling of a steam distillation unit for the extraction of aniseed (Pimpinella anisum) essential oil, J. Chem. Technol. Biotechnol. 80, 759–766.

[12] Milojević S. Ž., Stojanović T. D., Palić R., Lazić M. L., Veljković V. B., (2008). Kinetics of distillation of essential oil from comminuted ripe juniper (Juniperus communis L.) berries, Biochem. Eng. J. 39, 547–553.

[13] Benyoussef E. H., Hasni S., Belabbes R., Bessiere J.M., (2002). Modélisation du transfert de matiére lors de l'extraction de l'huile essentielle des fruits de coriandre, Chem. Eng. J. 85, 1-5.

[14] Veličković D., Ristić M., Stojiljković D., Šmelcerović A., (2001). Kinetics of obtaining the essential oil by different technological procedures from flowers, leaves and stems of sage (Salvia officinalis L.), Lek. sirov. 21, 67–72.

[15] Babu G. D. K., Singh B., (2009). Simulation of Eucalyptus cinerea oil distillation: A study on optimization of 1,8-cineole production, Biochem. Eng. J. 44, 226–231.

[16] Stanojević Lj., Stanković M., Cakić M., Nikolić V., Nikolić Lj., Ilić D., Radulović N., (2011). The effect of hydrodistillation techniques on yield, kinetics, composition and antimicrobial activity of essential oils from flowers of Lavandula officinalis L., Hem. Ind. 65, 455–463.

[17] Bousbia N., Vian M. A., Ferhat M. A., Petitcolas E., Meklati B. Y., Chemat F., (2009). Comparison of two isolation methods for essential oil from rosemary leaves: Hydrodistillation and microwave hydrodiffusion and gravity, Food Chem. 114, 355–362.

[18] Miletić P., Grujić R., Marjanović-Balaban Ž., (2009). The application of microwaves in essential oil hydrodistillation processes, Chem. Ind. Chem. Eng. Q. 15, 37–39.

[19] Pornpunyapat J., Chetpattananondh P.and Tongurai C., (2011). Mathematical modeling for extraction of essential oil from Aquilaria crassna by hydrodistillation and quality of agarwood oil, A Journal of the Bangladesh Pharmacological Society (BDPS) Bangladesh J Pharmacol, 6: 18-24.

[20] Masango P., (2005). Cleaner production of essential oils by steam distillation, J Clean Prod. 13, 833–839.

[21] Stanković M. Z., Nikolić N. C., Stanojević Lj. P., Cakić M. D., (2004). The effect of hydrodistillation technique on the yield and composition of essential oil from the seed of Petroselinum crispum (Mill.) Nym.ex.A.W.Hill., Hem. Ind. 58, 409–412.

[22] Hasmita I., Adisalamun, Alam P. N., Mahlinda, SupardanD. M., (2015). Effect of Drying and Hydrodistillation Time on the Amount of Ginger Essential Oil, International Journal on Advanced Science, Engineering and Information Technology, Vol.5, No. 5

[23] Megawati M., Handayani P. A., Sediawan W. B., (2013). Extraction of P. cubeba Essential Oil by Microwave Assisted Hydrodistillation: Modeling and Process Optimization, American Journal of Oil and Chemical Technologies,; Vol. 1, Issue 10, December 2326-6570.

[24] Ellouze I., (2014). Manef Abderrabba, Kinetics of extraction of Citrus aurantium essential oil by hydrodistillation: influence on the yield and the chemical composition, J. Mater. Environ. Sci. 5 (3), 841-848, ISSN : 2028-2508 CODEN: JMESCN.

[25] Viana M. A., Fernandezb X., Visinonic F., Chemata F., (2008). Microwave hydrodiffusion and gravity, a new technique for extra-ction of essential oils, Journal of Chromatography A, 1190, 14–17.

[26] Kusuma H. S., Mahfud M., (2016). Microwave-assisted Hydrodistillation for Extraction of Essential Oil from Patchouli (Pogostemon cablin) Leaves, Periodica Polytechnica Chemical Engineering, OnlineFirst, paper 8676.

[27] Chenni M., El Abed D., Rakotomanomana Nj., Fernandez X. and Chemat F., (2016). Comparative Study of Essential Oils Extracted from Egyptian Basil Leaves (Ocimum basilicum L.) Using Hydro-Distillation and Solvent-Free Microwave Extraction, Molecules, 21, 113.

[28] Bertouche S., Sahraoui N., Boutekedjiret C., (2016). Extraction of Thyme (Thymus pallecens de Noé) Essential Oil by Steam-Distillation, Steam-Diffusion and Hydro-Distillation Processes: Optimization of Operating Conditions and Antioxidant Activity.

[29] Ammar A. Haj, Meniai A. H., Zagrouba F., (2014). Experimental Study and Modeling of Essential Oil Extraction from Plants by Hydrodistillation, Chem. Eng. Technol., 37, No. 00, 1–9.

[30] Zeković Z., at all. (2015). Coriander seeds processing: Sequential extractionof non-polar and polar fractions using supercriticalcarbon dioxide extraction and ultrasound-assistedextraction, Food and bioproducts processing 95, 218–227.

[31] Mebrouk K., Saibi S., Nacer-Bey N., Benyoussef E.H.,(2012). Modelling of Water Distillation Kinetic of Pistacia lentiscus L. Essential Oil, Journal of essential oil-bearing plants JEOP.

[32] Bensebia O., Allia K., (2015). Drying and Extraction Kinetics of Rosemary Leaves: Experiments and Modeling, Journal of Essential Oil Bearing Plants, 18:1, 99-111.

[33] Sovova H., Aleksovski S., (2006). Mathematical model for hidrodistillation of essential oils, Flavour Fragr. J. 21, 881-889.

[34] Ritz C., Streibig J. C.,(2008) Nonlinear Regression with R, Springer Science & Business Media, New York.

[35] Cichosz, P., (2015). Data Mining Algorithms: Explained Using R., John Wiley & Sons.

[36] Akaike H., (1974). A new look at the statistical model identification, IEEE Trans. on Autom. Control 19(6), 716 - 723.

[37] Bates D. M., Watts D. G., (1988). Nonlinear Regression Analysis and Its Applications, Wiley, New York

Programiranje sistemov industrijske avtomatizacije z domensko specifičnimi jeziki

Tomaž Kos

Dewesoft d.o.o.
Gabrsko 11a, 1420 Trbovlje
E-pošta: tomaz.kos@dewesoft.com

Programming industrial automation systems with domain-specific languages

Abstract. *We can say that today there is almost no industry that does not have at least one process automatic. It is possible to automate almost anything. There are various programming languages that can automate processes. For that reason, the correct chose of programming language is important. In this article two domain-specific visual programming languages are presented that can be used for this purpose. The first is Dewesoft Sequencer and the second is Ladder Diagram. Each has its own advantages and disadvantages, but which programming language to choose is always a question.*

1 Uvod

Avtomatizacija v industriji pomeni uporabo nadzornih sistemov, kot so računalniki in roboti, ter uporabo informacijskih tehnologij za upravljanje različnih procesov in strojev v industriji, ki nadomestijo človeka. Dandanes je avtomatizacija del vsakega proizvodnega ali tehnološkega procesa, tako pri izdelavi avtomobila, letala ali pa v kontinuiranem procesu kot je na primer nadzor celotnega sistema elektrarne. Tekoči trakovi in roboti so le del avtomatizacije. Poleg omenjenega, med avtomatizacijo štejemo tudi obdelavo procesnih podatkov, komunikacijo med porazdeljenimi sistemi, vizualizacijo, sledljivost, končno kontrolo, vzdrževanje kakor tudi vodenje podjetja. Prvotni namen industrijske avtomatizacije je bil povečati produktivnost in zmanjšati stroške povezane z zaposlenimi operaterji, saj lahko avtomatizirani sistemi delujejo 24 ur na dan. Danes se avtomatizacija osredotoča tudi na povečanje kakovosti in prožnosti v proizvodnem procesu.

V zadnjem desetletju se je na področju krmilne tehnike precej spremenilo. V šestdesetih letih prejšnjega stoletja je avtomatizacija večine industrijskih procesov vsebovala analogne ožičene krmilnike, kar so jih kasneje zamenjali mikrokrmilniki in klasični programirljivi logični krmilniki (angl. Programmable Logic Controllers - PLC) [1]. Ti krmilniki se uporabljajo še danes, vendar se vse več industrije odloča za krmilnike, ki temeljijo na splošno namenskih računalnikih (npr. Acontis, Beckoff, Dewesoft, LabVIEW). Takšni krmilniki prinašajo večjo fleksibilnost in razširljivost, vendar po drugi strani zaradi kompleksnosti lahko privedejo tudi do manjše stabilnosti sistema. Z njim lahko zajamemo, vizualiziramo in analiziramo procesne podatke, izvajamo krmiljenja in regulacije, varujemo procese v konfliktnih situacijah in opozarjamo na kritične situacije.

Za reševanje problemov se uporabljajo različni programski jeziki in pristopi. Poznamo tako imenovane splošno namenske jezike (angl. General-Purpose Languages - GPL). Med te jezike prištevamo C, C++, Visual Basic, C# itd. S temi jeziki lahko razvijamo aplikacije iz različnih problemskih področji. Nasprotje pa predstavljajo t.i. domensko specifični jezi (angl. Domain-Specific Lanuages – DSL) [2]. Ti jeziki so se razvili predvsem zaradi velikih potreb po hitrem razvoju programov za specializirane problemske domene. V ospredje postavljajo domeno abstrakcije in semantiko, ki omogoča domenskim strokovnjakom programiranje direktno z domenskimi koncepti.

Zaradi visokega nivoja abstrakcije so domenski specifični jeziki zelo primerni za načrtovanje kompleksnih sistemov industrijske merilne avtomatizacije. Programski paket Dewesoft [3], ki se uporablja za zajem, obdelavo in kasneje tudi analizo izmerjenih meritev, uporablja domensko specifičen jezik Sekvencer za definiranje merilnih postopkov [4]. S tem domensko specifičnim jezikom lahko uporabnik na zelo enostaven način, brez strokovnega računalniškega znanja, izdela poljubno avtomatizacijo merilnega procesa, ki se vrši na industrijskem sistemu.

Članek je zgrajen iz 6 poglavij. V drugem poglavju predstavimo osnovno zgradbo industrijskega krmilnika. Sledi tretje poglavje kjer sta opisana dva DSL jezika, ki se uporabljata za programiranje avtomatizacije procesov. V četrtem poglavju je narejena primerjava med jezikoma Sekvencer in lestvičnim diagramom. Za boljše razumevanje je v 5 poglavju predstavljen praktični primer uporabe. Zadnje poglavje povzame vsebino članka.

2 Zgradba industrijskega krmilnika

Splošno namenski krmilniki za industrijske procese so običajno industrijski računalniški nadzorni sistemi, ki nenehno spremljajo stanje vhodnih modulov. Ti moduli so lahko digitalni vhodi, analogni vhodi, in komunikacijski ali posebni funkcijski moduli (npr. časovni števci). Na osnovi vhodnih stanj in programa napisanega s strani uporabnika, sprejema odločitve in upravlja z delovanjem izhodnih naprav, ki direktno komunicirajo in krmilijo izhodne periferije. Tovrstni krmilniki so tipičen primer realno-časovnih sistemov,

saj je zakasnitev v odzivnem času glede na spremembo vhodov manjša od določenega majhnega časa [5].

Krmilni program se v krmilniku izvaja znotraj krmilnega cikla (angl. scan cycle) in se odvija v štirih ponavljajočih se fazah: branje vhodov, izvajanje programske kode, pisanje na izhode in komunikacija ter diagnostika sistema. Tovrstno izvajanje predstavlja tudi nekatere nevarnosti. Lahko se zgodi, da pride do sprememb na vhodu signala v času, ki je manjši od odzivnega časa krmilnika. V tem primeru krmilnik teh sprememb ne zazna. V ta namen imajo različni krmilniki vgrajene dodatne mehanizme, ki zaznavajo spremembe tudi v tem času. Nekateri podpirajo tudi možnost delovanja s pomočjo prekinitev.

Vsak program je običajno zapisan v posebni aplikaciji namenjeni za programiranje in simulacijo na osebnem računalniku. Ko je program zaključen, se prenese na krmilnik kjer se izvaja. Program je shranjen v PLC-ju, v pomnilniku (RAM-u), ki se napaja iz baterije ali v drugem trajnem pomnilniku.

Nekateri krmilniki med delovanjem potrebujejo interakcijo z ljudmi kar opravijo s pomočjo vmesnika človek-stroj (angl. Human-Machine Interface - HMI). Z grafičnim vmesnikom lahko spreminjamo različne parametre krmilnika, spremljamo opozorilo o alarmih ali nadzorujemo sistem. Za nadzor na višjem nivoju in povezovanje različnih sistemov v celoto, se uporabljajo računalniški nadzorni sistemi (angl. Supervisory Control and Data Acquisition - SCADA). Nadzorni sistemi komunicirajo s krmilniki preko standardnih protokolov kot je na primer protokol OPC-UA [6].

3 Programiranje logičnih krmilnikov

Naloga avtomatizacije merilnih procesov je nadomestiti človeško odločanje z uporabo mehanske opreme in ukazov za logično programiranje. Za opisovanje logike delovanja krmilnika so se razvili številni DSL jeziki.

V nadaljevanju poglavja si bomo podrobneje pogledali dva vizualna programska jezika. Prvi je Sekvencer podjetja Dewesoft, ki ga primerjamo z enim izmed bolj uporabljenih programskih jezikov, lestvičnim diagramom (angl. Ladder Diagram - LD).

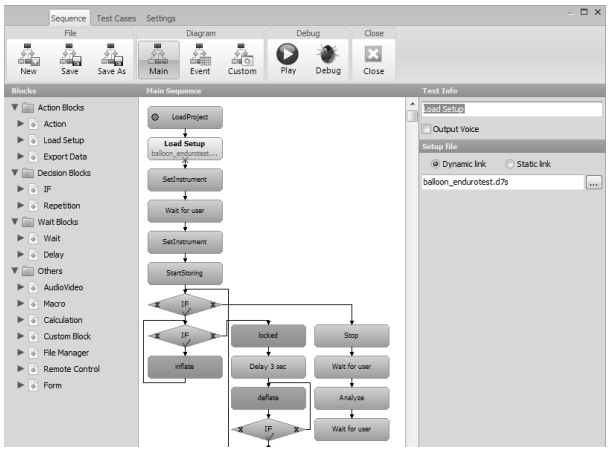
3.1 Dewesoft Sekvencer

Jezik Sekvencer je vključen v programski paket Dewesoft X3 (trenutna verzija X3 SP4). V osnovi je bil načrtovan v merilni industriji za kreiranje testnih postopkov med samim testiranjem. Danes pa se uporablja tudi v industrijski avtomatizaciji na najrazličnejših področjih (avtomobilska, letalska, elektro industrija itd.).

Krmiljenje najrazličnejših postopkov lahko kreiramo v vizualnem ali tekstovnem modelirnem okolju. Okolji sta prilagojeni za izgradnjo logike procesov v merilni domeni. Vizualne gradnike premikamo po metodi »primi in spusti« in jim določamo pripadajoče lastnosti.

Programski gradniki predstavljajo akcije, ki se izvajajo. Akcije se izvajajo od začetnega gradnika (označen s krogcem) in se nadaljuje na naslednjem gradniku kamor kaže puščica. Gradniki lahko vsebujejo tudi lokalne in globalne spremenljivke, ki so namenjeni za hranjenje trenutne vrednosti. Ker lahko ob velikem številu gradnikov vizualno programiranje postane neobvladljivo, modelirni jezik omogoča tudi vgnezdene modele (angl. custom blocks) s pomočjo katerih lahko združimo gradnike v celoto in se nanje sklicujemo v preostalem delu programa.

Slika 1 prikazuje razvojno orodje. Na levi strani najdemo gradnike domenskega jezika. Na sredini je delovna površina, kjer končni uporabniki programirajo proces. Na desni strani zaslona pa najdemo lastnosti, ki jih lahko uporabnik izbira za izbrani gradnik iz generiranega modela.



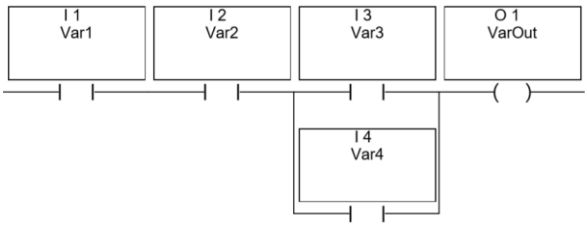
Slika 1: Primer programa zapisanega v jeziku Sekvencer

3.2 Standard IEC 61131-3: lestvični diagram

Standard IEC 61131-3 [7] govori o smernicah PLC programskih jezikov. Definira strukturo programov, podatkovne tipe in spremenljivke, standardne funkcije in funkcijske bloke, ter definira štiri osnovne programske jezike: seznam ukazov (angl. Instruction List - IL), strukturiran tekst (angl. Structured Text - ST), lestvični diagram (angl. Ladder Diagram - LD) ter funkcijski blokovni diagram (angl. Function Block Diagram - FBD). Vsak program je lahko zapisan v enem od omenjenih jezikov, možno pa jih je med seboj tudi kombinirati. Dodaten programski jezik, ki je namenjen strukturiranju opravil in programskih modulov je tako imenovani sekvenčni funkcijski diagram (angl. Sequential Function Chart - SFC).

Najstarejši in najbolj uporaben od osnovnih programskih jezikov je lestvični diagram. Jezik izhaja iz relejske tehnike, kjer so bili koncepti jezika razviti iz električnih shem in so namenjeni za procesiranje logičnih digitalnih signalov. Osnovni simboli, ki se uporabljajo so: normalno odprt kontakt, normalno zaprt kontakt in tuljava oz. izhodni kontakt. Izvajanje lestvičnega diagrama je iz leve proti desni in od zgornje do spodnje vrstice.

Slika 2 prikazuje enostaven program zapisan v lestvičnem diagramu. Normalno odprta kontakta Var1 in Var2 sta povezana zaporedno in predstavljata logično operacijo »IN«. Nato sledita elementa Var3 in Var4, ki sta povezana vzporedno in predstavljata logično operacijo »ALI«. Elementi od Var1 do Var4 so vhodni signali, kjer se izhodna vrednost zapiše v VarOut element ali izhodni kontakt. Spodnje grafične elemente bi lahko zapisali tudi v obliki OutVar = Var1 IN Var2 IN (Var3 ALI Var4).

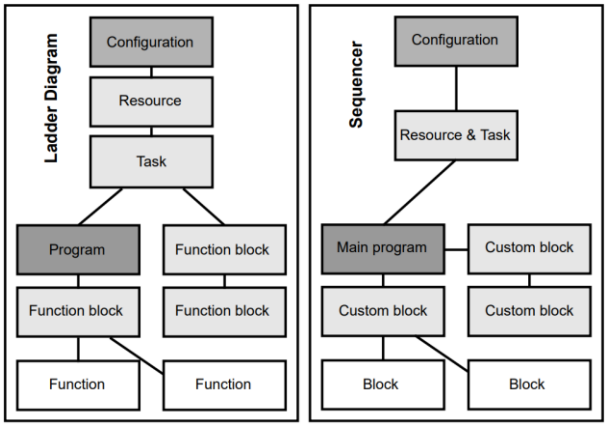


Slika 2: Primer programa zapisanega z lestvičnim diagramom

4 Primerjava Sekvencer jezika z lestvičnim diagramom

Jezika Sekvencer in lestvični diagram spadata v skupino domensko specifičnih vizualnih jezikov. Vendar sta si tako po zgradbi, kakor tudi po tipu programskih jezikov različna.

Struktura lestvičnega diagrama, ki je definirana po standardu, je zasnovana v več nivojih (Slika 3). Na najvišjem nivoju je konfiguracija (angl. configuration), ki določa kako in kje se koda izvaja. Nato sledi vir (angl. resource), ki je dejansko krmilnik oz. procesna enota. Znotraj vira so definirana opravila (angl. tasks). Opravilo je enota, ki vsebuje informacije o tem kako in kdaj naj se nek program izvede. Nato hierarhično sledi izvajalni program, ki je sestavljen iz programa (angl. program), funkcijskih blokov (angl. function block) in funkcij (angl. function). Tudi programski jezik Sekvencer ima podobno hierarhično strukturo. Na vrhu tako kot pri lestvičnem diagramu je konfiguracija sistema. Ker programski jezik Sekvencer ne podpira večopravilnosti in je vir vezan na trenutni sistem, sta naslednja nivoja enakovredna. Nato sledi glavni program (angl. main program), ki je kasneje razdeljen na vgnezdene (angl. custom blocks) in osnovne konstrukcijske bloke (angl. blocks).



Slika 3: Struktura programskih orodij

Programski jezik Sekvencer se izvaja koračno. Akcije se izvajajo sosledno in se prekinjajo glede na krmilni cikel. Programski jezik omogoča mehanizem dogodkov (angl. events), ki prinaša nekatere prednosti [8]. Lestvični diagram ne podpira dogodkov, izvaja se v krmilnem ciklu, ki v splošnem namenskem jeziku predstavlja neskončno zanko, ki se jo programerji skušamo izogniti zaradi slabega performančnega izkoristka.

Lestvični diagram se razlikuje od običajnega postopkovnega programiranja kot je Sekvencer na več načinov. Jezik je precej primitiven, vendar zahteva dobro znanje Boole-ove algebre. Večina uporabnikov ne uporablja podprogramov, zato se lestvični diagram uporablja skupaj s sekvenčnim funkcijskim diagramom, kar predstavlja dvojno poznavanje in učenje jezikov.

Prenosljivost programske kode obeh jezikov je med sistemi istega proizvajalca dobra. Problem se pojavi pri prenosu med različnimi proizvajalci, saj ni definiranega standardnega formata za izmenjavo programske kode.

Čeprav je lestvični diagram zasnovan na bitnih vrednostih, omogoča in podpira tudi druge podatkovne tipe kot so celoštevilčna števila in števila s plavajočo vejico. Podobno velja za programski jezik Sekvencer, ki poleg omenjenih tipov podpira tudi kompleksna števila in tekstovne nize. Oba jezika omogočata nekatere kompleksne tipe kot so digitalni in časovni števci. Spremenljivke omenjenih tipov pri obeh jezikih so lahko zunanje ali notranje, globalne ali lokalne, omogočajo pa tudi uporabo konstant. Posebnost jezika Sekvencer je v tem, da spremenljivke v osnovi ne hranijo samo trenutne vrednosti, ampak tudi zgodovino za določeno obdobje, ki se lahko kasneje uporabi pri analizi obnašanja sistema.

Ker je lestvični diagram standardiziran obstaja na trgu ogromno različnih razvojnih orodij. Prav tako obstajajo različni odprtokodni projekti kot je na primer OpenPLC, Beremiz. Razvojna orodja omogočajo različne funkcionalnosti in delujejo v večini samo za specifičen tip naprav določenega proizvajalca. Ker programski jezik Sekvencer ni standariziran, je v omenjenem jeziku mogoče programirati samo znotraj produkta Dewesoft in njihovih sistemov. Orodje omogoča tudi nekatere dodatne funkcije kot so na primer preoblikovanje programa (angl. refactoring) [9].

Splošno je znano, da napake, ki se pojavijo znotraj kontrolnega sistema, težko najti in odpraviti. Še posebej je problem lokalizirati napako v realno časovnih sistemih. V teh primerih uporabniki za odpravo problema uporabljajo dodaten sistem s pomočjo katerega zajemajo in shranijo vhodne ter izhodne podatke (signale) iz realnega sistema in jih kasneje analizirajo. V ta namen je bil razvit razhroščevalnik, ki pomaga odkriti napake, ki se pojavijo bodisi na strojni ali programski opremi (sekvenčnem programu). Pri tem orodju se lahko uporabijo različne tehnike, kot so koračno izvajanje, prekinitvene točke, animacija, pregled spremenljivk in pregled sledenja modela, s katerimi hitreje lokaliziramo in odpravimo napake.

Pri samem razvoju programov za avtomatizacijo je prav tako zelo pomembno testiranje zgrajenih programov. V ta namen se največkrat uporabljajo

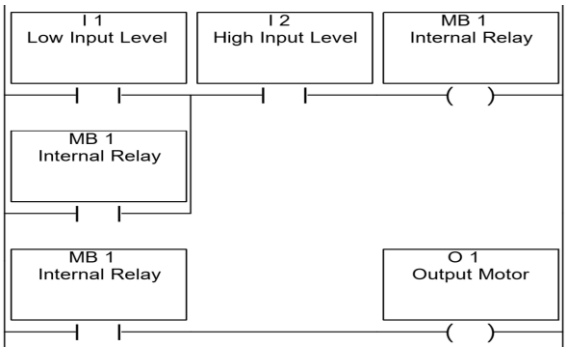
različne simulacije, ki pripomorejo k pravilnem delovanju. Pri orodju Sekvencer je bilo razvito orodje, ki omogoča kreiranje testnih primerkov tako za strojno kakor tudi programsko opremo.

Takšni vizualni jeziki imajo tudi nekatere slabosti. Jezike je mogoče uporabljati le na grafičnih operacijskih sistemih. Pri kreiranju programa zahtevajo več pomnilnika, večji prostor za shranjevanje in hitrejši procesor v primerjavi s tekstovnimi jeziki. Ena od slabosti je tudi razvoj takšnega jezika, ki je težak, saj je potrebno dobro poznavanje domene kot tudi znanje iz gradnje programskih jezikov. Zaradi omenjenih slabosti je cena načrtovanja, implementacije in vzdrževanja višja v primerjavi s splošno namenskimi jeziki.

5 Praktičen primer uporabe

Uporabo omenjenih jezikov bomo prikazali na primeru regulacije nivoja olja v rezervoarju. Za detekcijo nivoja olja bomo uporabili dva senzorja. Prvi senzor (High Input Value) bo nameščen na vrhu in bo določal visoki nivo, drugi senzor (Low Input Value) pa na dnu rezervoarja in bo določal nizki nivo olja. Senzorja sta aktivna (logična »1«), ko je tekočina pod nivojem. Dokler senzor ne doseže določenega praga, črpalka (Output Motor) polni olje v posodo. Ko se bo vklopil senzor visokega nivoja, bomo izklopili motor in počakali dokler nivo olja ne pade pod senzor nizkega nivoja. Ko pade pod ta nivo, ponovno vklopimo motor in ponavljamo postopek.

Slika 4 prikazuje program zapisan z lestvičnim diagramom. Na začetku je tank prazen (I1 in I2 sta aktivna), zato je interna spremenljivka (MB1) aktivna in motor prižgan (O1). Ko doseže olje nizek nivo, postane neaktivna L1 in pri visokem nivoju še L2. Po tem dogodku postane neaktivna MB1 in posledično se motor ustavi.

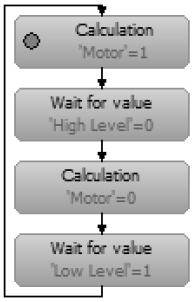


Slika 4: Program zapisan z lestvičnem diagramom

Na Sliki 5 je predstavljen algoritem zapisan z jezikom Sekvencer. Kontrola tanka je zgrajena iz štirih blokov. Najprej se vklopi motor za polnjene tanka (začeten blok je označen s krogcem). Nato se počaka, da doseže visoki nivo ('High Level'=0). Po tem dogodku ustavimo motor in ponovno počakamo, da se tank sprazni do nizkega nivoja ('Low Level'=0).

6 Zaključek

Industrijska avtomatizacija prihaja vse bolj in bolj v ospredje in postaja vse bolj pomembna zaradi velikih prednosti kot so povečanje produktivnosti, kvalitete in zmanjšanje cene. Zato je kratek čas izdelave produkta zelo pomemben. Pri tem moramo izbrati pravi jezik, ki je visokonivojski, enostaven in razumljiv. V članku sta predstavljena domensko specifična jezika lestvični diagram in Sekvencer. Prvi je zaradi zgodovinskih lastnosti močno uporabljen, standariziran in primeren za programiranje krmilnih sistemov v realnem času (angl. hard real-time systems). Drugi jezik Sekvencer je zaradi njegove enostavnosti namenjen za avtomatizacijo, vendar zaradi arhitekturne zasnove uporabljen za mehke sisteme v realnem času (angl. soft real-time systems), kjer majhni odziven čas ni nujno potreben.



Slika 5: Program zapisan v jeziku Sekvencer

Literatura

[1] G. Dunning. Introduction to Programmable Logic Controllers, 3rd Edition. Delmar Thomson Learning, 2006.

[2] M. Mernik, J. Heering, A. Sloane. When and how to develop domain-specific language. ACM computer surveys, 2005, štev. 37, zv. 4, str. 316-344.

[3] Dewesoft, https://www.dewesoft.com

[4] T. Kos, T. Kosar, M. Mernik, Development of data acquisition systems by using a domain-specific modeling language, Computers in Industry, 2012, štev. 63, str. 181-192.

[5] H. Kopetz, Real-Time Systems Design Principles for Distributed Embedded Applications, 2nd Edition, U.K., London:Kluwer, 2011.

[6] W. Mahnke, S. Leitner, M. Damm, OPC Unified Architecture, 2nd Edition, Springer, 2011.

[7] K.-H. John, M. Tiegelkamp, IEC 61131-3: Programming Industrial Automation Systems, 2nd Edition, Springer, 2011.

[8] F. Dabek, N. Zeldovich, F. Kaashoek, D. Mazières, and R. Morris., Event-driven programming for robust software. In Proceedings of the 10th workshop on ACM SIGOPS European workshop (EW 10). ACM, NY, 2002, USA, str. 186-189.

[9] T. Kos, T. Kosar, M. Mernik, Princip preoblikovanja domensko specifičnih modelov s programskim paketom DEWESoft, Zbornik dvajsete mednarodne konference Elektrotehniške in računalniške konference, 2011, štev. 20, str. 119-122.

Uporaba knjižnice Blockly za razvoj blokovnega jezika

Matej Moravec[1], **Tomaž Kosar**[1], **Marjan Mernik**[1]

[1] *Univerza v Mariboru,*
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46, 2000 Maribor, Slovenia
E-pošta: matej.moravec@student.um.si, tomaz.kosar@um.si, marjan.mernik@um.si

Developing a visual language using Blockly library

Blockly is a client-side JavaScript library for creating visual block programming languages and editors. Visual block programming languages are becoming more and more popular for children and older people who want to learn the basic concepts of programming. They are also used to build more complex programs in different domains. So in this paper we show how to develop a visual block programming language using Blockly library. We show practical example of application in which we create a custom blocks and code generators for each block. From created custom blocks we assemble one example of program, generate code and show the results. In this paper we also mention some pros and cons about use of visual block programming languages.

1 Uvod

Programski jeziki so programerjeva najbolj osnovna orodja. Stroške izdelave nove aplikacije lahko drastično zmanjšamo z izbiro ustreznega programskega jezika. Prav tako so lahko stroški vzdrževanja z izbiro ustreznega programskega jezika občutno manjši. V zadnjih desetletjih je tehnologija programskih jezikov veliko napredovala v smislu tradicionalnih programskih paradigem, kot so funkcijsko, logično in objektno orientirano programiranje. Prav tako je napredovalo vizualno orientirano programiranje [1]. Glavna gonilna sila razvoja je bila in bo boljše izražanje programerjevih idej. Ravno zato so raziskave na področju programskih jezikov neskončna dejavnost in jedro računalništva. V prihodnosti lahko predvidimo nove lastnosti jezikov, nove programske paradigme in mehanizme z boljšim časom prevajanja in zagona.

Lažje izražanje v specifični domeni pa prinašajo tudi domensko specifični jeziki (ang. Domain-Specific Language (DSL)) [2]. DSL je jezik, ki je zasnovan z namenom, da zagotovi notacijo, ki je blizu aplikacijski domeni, in temelji samo na konceptih in lastnostih te domene. Kot tak je DSL sredstvo za opis in generiranje "članov" programske družine v določeni domeni. DSL z zagotavljanjem notacije, ki je blizu aplikacijski domeni, ponuja veliko prednosti pred t. i. splošno namenskimi jeziki (angl. General-Purpose Languages (GPL)) [3]. Ena izmed prednosti je, da pri uporabi DSL ni potrebno znanje splošnega programiranja in ga tako lahko uporabljajo tudi končni uporabniki [4].

Domensko specifični jeziki se v marsičem prekrivajo z domensko specifičnimi modelirnimi jeziki (angl. Domain-Specific Modeling Languages (DSML)) [5]. Ti specificirajo programe z modeli in uporabljajo vizualne domenske koncepte ter tako v nekaterih domenah še dodatno dvignejo stopnjo abstrakcije izven nivoja kodiranja. Končni produkt je lahko na podlagi teh specifikacij generiran na zelo visoki ravni.

Programi, ki jih pišemo z DSML-ji, so v veliki meri podobni programom, ki temeljijo na blokih (angl. block-based programming languages) [6]. Ti so se dobro uveljavili za uvajanje otrok v programiranje [7]. Blokovni jeziki, ki jih poznamo, pokrivajo različne domene, ki učence zabavajo in motivirajo za programiranje preprostih iger (Scratch), interaktivno multimedijo (Pixly), preprosto robotiko (RoboScratch) in mobilne telefone (MIT AppInventor). V tem delu nas je zanimalo, kako se blokovni jeziki razlikujejo od DSML-jev. Naš cilj je bil raziskati to področje programiranja, ki se v zadnjem času zelo razvija – programski jeziki, osnovani na blokih.

Članek je strukturiran v pet poglavij. V drugem poglavju prikažemo zapis vizualne notacije posameznega bloka. V tretjem poglavju predstavimo, kako se na podlagi vizualnega programa v ozadju generira koda, zapisana v poljubnem programskem jeziku. V četrtem poglavju v praksi predstavimo primer uporabe vizualnega jezika. V zadnjem poglavju povzamemo naše ugotovitve o uporabnosti knjižnice Blockly [8] za izdelavo vizualnega jezika in učinkovitost avtomatskega generiranja kode.

2 Zapis vizualne notacije

Blokovni jeziki omogočajo uporabniku specifikacijo programa z različnimi vizualnimi elementi, namesto s tekstom. Z uporabo knjižnice Blockly [9] lahko uporabniku pripravimo poljubno število različnih vizualnih elementov, s katerimi lahko sestavi kompleksen program.

V aplikaciji, izdelani s knjižnico Blockly, posamezen konstrukt v kodi predstavimo z blokom. Določene dele procesov pri konfiguraciji Blockly lahko avtomatiziramo z uporabo spletnega orodja Blockly Developer Tools [9]. Pri tem nam orodje z zlaganjem različnih gra-

dnikov omogoča samodejno generiranje notacije posameznega bloka. Notacija je lahko zapisana s programskim jezikom JavaScript ali z notacijo JSON.

Orodje Blockly Developer Tools vključuje ustvarjanje blokov po meri, gradnjo seznama, kjer se nahaja nabor razpoložljivih blokov, in konfiguriranje spletne delovne površine, ki je namenjena združevanju blokov v končni program. Razvojni proces Blockly z uporabo orodja sestoji iz treh delov:

- Ustvaritev oz. definiranje bloka in možnost izvoza bloka.
- Gradnja seznama, kjer se nahaja nabor blokov, in določitev izgleda privzete delovne površine.
- Konfiguracija oz. določitev funkcionalnosti delovne površine.

Z uporabo orodja Blockly Developer Tools nam je omogočeno poenostavljeno ustvarjanje blokovnih definicij in generatorjev kode za posamezen blok. Bloke lahko enostavno ustvarimo, spremenimo in jih shranimo. Ustvarjene bloke in generatorje kode lahko iz orodja izvozimo in jih uporabimo v aplikaciji. Z uporabo orodja nam je omogočena enostavna konfiguracija seznama, kjer se nahaja nabor razpoložljivih blokov za gradnjo programa. Določimo lahko tudi, kateri bloki se bodo ob zagonu aplikacije že nahajali na delovni površini.

2.1 Primer zapisa vizualne notacije za posamezen konstrukt v kodi

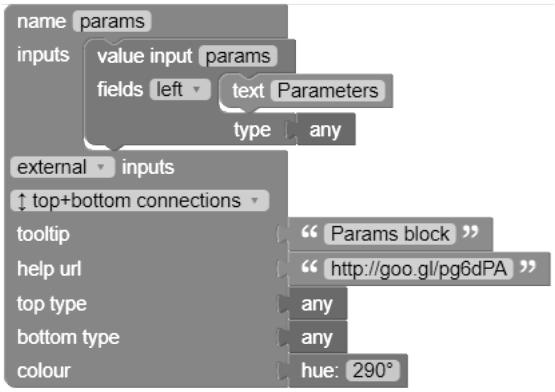
Na vsak blok se navezujemo z njegovim imenom, zato mora biti vsak posamezen blok predstavljen s svojim unikatnim imenom. Videz bloka in njegovo obnašanje opišemo z definicijo bloka, ki vključuje besedilo, barvo, obliko in opis, kako se drugi bloki lahko z njim povežejo [8].

V programski kodi 1 vidimo, da moramo bloku v prvi vrstici najprej definirati njegovo ime. Nato v vrsticah 2-12 definiramo celotno vizualno predstavitev bloka. Vrstice 3-5 definirajo vhod v blok oz. vrednost, ki jo blok lahko prejme. Vrstice po vrsti definirajo najprej oznako vhodne vrednosti, s katero se lahko kasneje pri generiranju kode nanjo navežemo oz. do nje dostopamo. Zatem je definiran tip, ki ga lahko blok na vhodu prejme (v našem primeru je nastavljen na *any* oz. *null*). Z določitvijo tipa vplivamo na pravilno sestavljanje programa. Pri tem Blockly podpira preverjanje več vrst tipov, kot so *Boolean*, *Number*, *String*, itd. V primeru, da želimo k bloku pripeti blok z blokom, ki ne vrača ustreznega tipa, kot je zapisano v definiciji bloka, Blockly teh dveh blokov ne bo povezal med sabo. Za določitvijo tipa je bloku dodano polje, s katerim nanj dodamo poljubno besedilo. Smiselno je, da na blok dodamo besedilo, ki uporabniku pomaga pri razumevanju pomena bloka. Vrstice 6-11 definirajo lastnosti bloka, ki povejo, kako blok spada v kontekst oz. kako se blok lahko v končnem programu poveže oz. združi z drugimi bloki programa. S temi lastnostmi definiramo, kako naj se vhodni bloki k bloku pripnejo, določimo, če se blok lahko pripne k predhodnemu bloku in če se neki drug blok lahko pripne nanj. Z metodo *setColour* definiramo odtenek barve, ki pri definiciji bloka pove, kako bo blok izgledal. S tem uporabniku intuitivno pomagamo razumeti, kateri bloki spadajo skupaj. S funkcijama *setTooltip* in *setHelpUri* uporabniku nudimo dodaten opis bloka oz. koristne informacije o uporabi bloka.

```
1  Blockly.Blocks['params'] = {
2    init: function () {
3      this.appendValueInput("params")
4          .setCheck(null)
5          .appendField("Parameters");
6      this.setInputsInline(false);
7      this.setPreviousStatement(true, null);
8      this.setNextStatement(true, null);
9      this.setColour(290);
10     this.setTooltip("Params block");
11     this.setHelpUrl("http://goo.gl/pg6dPA");
12   }
13 };
```

Programska koda 1: Definicija bloka “params”

Ekvivalent programski kodi JavaScript 1 predstavlja slika 1. Na sliki je prikazana sestava bloka v orodju Blockly Developer Tools. Na podlagi te sestave se nam programska koda 1 samodejno generira.



Slika 1: Vizualna definicija bloka “params”

Na sliki 2 je prikazana vizualna predstavitev bloka “params”, ki je zgrajena na podlagi programske kode 1. Vidimo, da se blok lahko pripne v kontekst drugega bloka in k njemu se lahko pripne naslednji blok v programu. Na desni strani lahko blok na vhodu prejme drug blok, katerega koda se pri generiranju kode združi s kodo bloka “params”. Na bloku podano besedilo “Parameters” pripomore k lažjemu razumevanju, čemu pri gradnji programa je blok namenjen.



Slika 2: Vizualna predstavitev bloka “params”

3 Generiranje kode

Večina aplikacij Blockly mora v programu uporabljene bloke, pretvoriti v izvajalno kodo. V ta namen za vsak

blok zapišemo generator kode. Izvajalna koda je lahko zapisana v različnih jezikih (JavaScript, Python, PHP, Lua, Dart, C#, itd.).

Prva naloga kateregakoli bloka je, da si pri generiranju kode pridobi vse argumente in podatke drugih blokov, ki so z njim združeni na vhodnem polju. Knjižnica Blockly vsebuje več funkcij, ki se lahko uporabijo za to nalogo:

- *getFieldValue*.
- *valueToCode*.
- *statementToCode*.

V programski kodi 2 je prikazan primer generatorja kode za blok "params". V kodi se najprej navežemo na ime bloka, za katerega želimo zapisati generator kode. Ko smo se navezali na ustrezen blok, zapišemo funkcijo, ki predstavlja generator kode. V vrsticah 3-5 s funkcijo *valueToCode* dostopamo do bloka "params", preko katerega pridobimo generirano kodo bloka, ki se nahaja na vhodu bloka "params". Kodo zapišemo v spremenljivko *value_params* in za tem v spremenljivko *code* ter slednjo, ki predstavlja generirano kodo, vrnemo. Spremenljivka *code* v tem primeru prejme samo vrednost spremenljivke *value_params*. V primeru, ko je blok bolj kompleksen in ima več možnih argumentov in parametrov, se v spremenljivki *code* vsi ti argumenti in vrednosti iz vhodov združijo v kodo, ki se vrne kot rezultat funkcije.

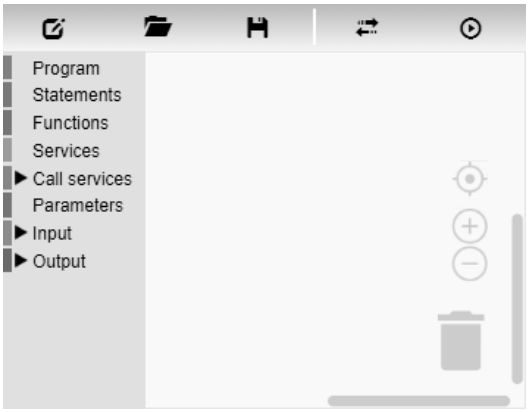
```
1  Blockly.JavaScript['params']=function(block)
2  {
3     var value_params = Blockly.JavaScript
4        .valueToCode(block, 'params',
5           Blockly.JavaScript.ORDER_ATOMIC);
6
7     var code = value_params;
8
9     return code;
10 };
```

Programska koda 2: Generiranje kode za blok "params"

4 Primer uporabe

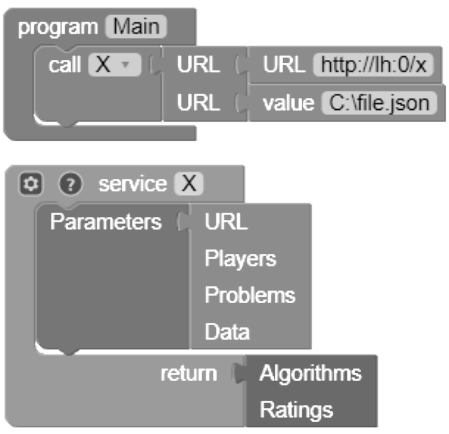
V tem poglavju prikažemo praktični primer izdelane spletne aplikacije. Praktični primer obsega že izdelane bloke, zapisano generiranje kode za vsak blok in funkcionalnosti, kot so priprava nove delovne površine, odpiranje že sestavljenih programov, shranjenih v XML-zapisu, itd.

Na sliki 3 je prikazana spletna aplikacija, ki jo lahko uporabljamo v brskalniku. Strukturirana je iz treh delov. Na levi strani vsebuje seznam vseh blokov, ki jih lahko uporabimo pri gradnji programa. Na vrhu aplikacije se nahaja menijska vrstica, v kateri se nahajajo gumbi, ki predstavljajo funkcionalnosti, namenjene za lažjo pouporabo že sestavljenih programov, in gumbi za generiranje izvajalne kode v programski jezik C#. Na desni strani aplikacije največji del predstavlja površina, na kateri lahko uporabnik sestavi želeni program. Na skrajni desni strani delovne površine se nahajajo gumbi, katerih namen je lažji pregled nad sestavljenim programom. Te funkcionalnosti omogoča knjižnica Blockly in z njimi lahko bloke v sestavljenem programu povečujemo oz. zmanjšujemo, jih na delovni površini centriramo in s potegom posameznega bloka ali skupka blokov na ikono v obliki koša blok oz. skupek blokov izbrišemo. Povečevanje oz. zmanjševanje in centriranje blokov nam koristi predvsem v primeru, ko imamo sestavljen bolj kompleksen program in se želimo po tem programu pomikati.



Slika 3: Aplikacija Blockly

Z uporabo spletne aplikacije lahko sestavimo različne programe. Primer sestavljenega programa prikazuje slika 4. Program je sestavljen iz dveh blokov, ki lahko vsebujeta sklop drugih blokov in v generirani kodi predstavljata funkciji. Vsak izmed blokov vsebuje druge bloke, ki predstavljajo telo tega bloka oz. funkcije. Prvi blok predstavlja "glavni" program, v katerem se kliče funkcija "X", ki pa je sestavljena v drugem bloku. Ob klicu funkcije se uporabijo še bloki, ki predstavljajo vrednosti parametrov, ki jih ob klicu pošljemo v funkcijo. V drugem bloku so vsebovani bloki, ki predstavljajo strukturo funkcije "X" in na podlagi katerih se znotraj funkcije kliče storitev. Izhod iz funkcije določimo s pripetim blokom ob napisu "return". Pri gradnji programa je pomembno, da pripnemo vsebino (bloke) vsem tistim blokom, ki so strukturirani tako, da imajo definirane vhode. V primeru, da bloka na vhodu ne pripnemo, bo pri generiranju manjkal del kode in programa ne bo mogoče izvesti.



Slika 4: Program Blockly

Ko imamo želeni program sestavljen, ga lahko z uporabo aplikacije, pretvorimo v generirano kodo. Pri generiranju se koda, ki jo predstavlja posamezen blok v sestavljenem programu, združi. Koda se v primeru naše aplikacije generira v programski jezik C#. Programska koda 3 predstavlja generirano kodo sestavljenega programa na sliki 4. Odsek kode med vrsticami 1-5 predstavlja generirano kodo iz prvega "večjega" bloka, medtem ko vrstice 7-13 predstavljajo generirano kodo drugega "večjega" bloka.

```
1  static void Main(string[] args)
2  {
3     var X = CallX("http://lh:0/x",
4        "C:\file.json");
5  }
6
7  static string CallX(string url, string json)
8  {
9     WSX.MyXWebService ws =
10       new WSX.MyXWebService();
11
12    return ws.X(json);
13 }
```

Programska koda 3: Generirana koda programa Blockly

5 Zaključek

V članku smo pokazali, kako izdelamo blokovni jezik s pomočjo uporabe knjižnice Blockly. Pri izdelavi posameznih blokov smo prišli do spoznanja, da moramo najprej dobro poznati domeno, za katero razvijamo blokovni jezik, tako lahko uporabniku najbolj poenostavimo "programiranje" oz. sestavljanje programa. Poleg izdelave je tudi pomembno, kako za vsak posamezen blok zapišemo generator kode, saj se ob sestavljenem programu generatorji kode posameznih blokov združijo in predstavljajo končni program.

Ustvarili smo spletno aplikacijo, ki vsebuje seznam ustvarjenih blokov različnih oblik. Vsak blok v seznamu predstavlja določen programski konstrukt in ima zapisan generator kode za programski jezik C#. Slabost pri programiranju s pomočjo blokovnega jezika je, da je uporabnik omejen le na ponujene programske konstrukte. Blokovni jeziki pa zagotavljajo to prednost, da ne moremo narediti sintaktičnih napak in uporabljati programskih konstruktov, ki ne spadajo skupaj.

V prihodnje načrtujemo obogatitev aplikacije z dodanimi še več programskimi konstrukti in generatorji kode za sestavljanje programa. Na podoben način, kot si sedaj uporabnik lahko ustvari nov blok s svojim imenom (ki predstavlja spremenljivko), bomo dodali še opcijo, da si bo lahko uporabnik ustvaril nov blok s svojim imenom, ki bo predstavljal funkcijo. Uporabniku bomo tudi omogočili zagon sestavljenega programa znotraj spletne aplikacije.

Literatura

[1] M. Boshernitsan and M. S. Downes. *Visual programming languages: A survey*. Computer Science Division, University of California, 2004.

[2] M. Mernik, J. Heering, and A. M. Sloane. When and how to develop domain-specific languages. *ACM computing surveys (CSUR)*, 37(4):316–344, 2005.

[3] T. Kosar, S. Gaberc, Jeffrey C. Carver, and M. Mernik. Program comprehension of domain-specific and general-purpose languages: replication of a family of experiments using integrated development environments. *Empirical Software Engineering*, Feb 2018.

[4] T. Kos, T. Kosar, J. Knez, and M. Mernik. From DCOM interfaces to domain-specific modeling language: A case study on the Sequencer. *Computer Science and Information Systems*, 8(2):361–378, 2011.

[5] J. P. Tolvanen and S. Kelly. Defining domain-specific modeling languages to automate product derivation: Collected experiences. In *International Conference on Software Product Lines*, pages 198–209. Springer, 2005.

[6] D. Bau, J. Gray, C. Kelleher, J. Sheldon, and F. Turbak. Learnable programming: blocks and beyond. *Communications of the ACM*, 60(6):72–80, 2017.

[7] D. Weintrop and U. Wilensky. To block or not to block, that is the question: Students' perceptions of blocks-based programming. In *Proceedings of the 14th International Conference on Interaction Design and Children*, IDC '15, pages 199–208, New York, NY, USA, 2015. ACM.

[8] Google. Blockly. `https://developers.google.com/blockly/`. Dostopano 5. 7. 2018.

[9] Google. Blockly developer tools. `https://developers.google.com/blockly/guides/create-custom-blocks/blockly-developer-tools`. Dostopano 5. 7. 2018.

Razpoznavanje vzorcev

Pattern Recognition

Visual Psychophysics for Machine Learning

Brandon RichardWebster

Uni Notre Dame, ZDA

Abstract. By providing substantial amounts of data and standardized evaluation protocols, datasets in computer vision have helped fuel advances across all areas of visual recognition. But even in light of breakthrough results on recent benchmarks, it is still fair to ask if our recognition algorithms are doing as well as we think they are. The vision sciences at large make use of a very different evaluation regime known as Visual Psychophysics to study visual perception. Psychophysics is the quantitative examination of the relationships between controlled stimuli and the behavioral responses they elicit in experimental test subjects. Instead of using summary statistics to gauge performance, psychophysics directs us to construct item-response curves made up of individual stimulus responses to find perceptual thresholds, thus allowing one to identify the exact point at which a subject can no longer reliably recognize the stimulus class. In this talk I will highlight two applications of Visual Psychophysics for Machine Learning: evaluation of object recognition and face recognition explainability.

Learning to super-resolve faces through recognition

Klemen Grm, Simon Dobrišek, Vitomir Štruc

University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, Ljubljana, Slovenia
klemen.grm@fe.uni-lj.si

In this paper, we propose a novel high magnification factor face hallucination model that incorporates identity priors into the learning procedure. The model consists of i) a cascaded super-resolution network that upscales the low-resolution images, and ii) an ensemble of face recognition models that act as identity priors during training. Our network uses a cascade of SR models that progressively upscale the low-resolution images using steps of $2\times$. This allows us to apply appearance and recognition supervision at different resolutions. Our model is able to upscale (very) low-resolution images captured in unconstrained conditions and produce visually convincing results. We evaluate the proposed model on a large dataset of facial images and report superior performance compared to the state-of-the-art.

1 Introduction

Face hallucination represents a domain specific super-resolution (SR) problem where the goal is to recover high-resolution (HR) face images from low-resolution (LR) inputs [1]. It has important applications in image enhancement, compression and face recognition [2], but also surveillance and security [3, 4].

Like other single-image super-resolution tasks, face hallucination is inherently ill-posed. Given a fixed image-degradation model, every LR facial image can be shown to have many possible HR counterparts. Thus, the solution space for SR problems is extremely large and existing solutions commonly try to produce plausible reconstructions by "hallucinating" high-frequency information based on the provided LR evidence. While significant progress has been made in recent years in the area of super-resolution and face hallucination [5, 6, 7, 8, 9, 10, 11, 12], super-resolving arbitrary facial images, especially at high magnification factors, is still an open and challenging problem, mainly due to:

- The ill-posed nature of the face hallucination problem, where the solution space is known to grow exponentially with an increase in the desired magnification factor [3].
- The difficulty of integrating priors beyond solely the visual quality of the reconstructions. Most of the existing priors utilized for super-resolution relate to image characteristics such as gradient distribution [13], total variation [14] and smoothness [15]. If discernibility of the semantic content is the goal of the SR procedure, such priors may not be the most optimal choice.



Figure 1: Results generated with the proposed method.

The outlined limitation are most evident for challenging face hallucination problems where very low-resolution images (e.g., 24×24 pixels) of arbitrary characteristics need to be super-resolved at high magnification factors (e.g., $8\times$). In this paper, we try to address some of these limitations with a new hallucination model build using convolutional neural networks (CNNs). Our model uses a cascade of simple super-resolution models (referred to as SR modules hereafter) for image upscaling and identity priors in the form of pretrained recognition networks as constraints for the training procedure. The SR models super-resolve the LR input images in magnification increments of $2\times$ and, consequently, allow for intermediate supervision at every scale. This intermediate supervision confines the explosion of the solution space size and contributes towards more accurate hallucination results. The recognition models are trained to respond only to the hallucinated high-frequency parts of the SR images and ensure that the added facial details are not only plausible, but as close to the true details as possible. Due to availability of intermediate SR results, we incorporate the identity constraints at multiple scales in the proposed model.

Overall, we make the following main contributions:

1. We propose a new face hallucination model, that integrates identity priors at multiple scales into the training procedure of a super-resolution network. To the best of our knowledge, this is the first attempt to exploit *multi-scale identity information* to constrain the solution space of deep-learning based SR models.
2. We introduce a *cascaded SR network* architecture that super-resolves images in magnification steps of $2\times$ and offers a convenient and transparent way of incorporating supervision signals an multiple scales.

Once trained, the SR network is able to upsample 24×24 pixel LR images at magnification factors of $8\times$ and produce realistic and visually convincing hallucination results as illustrated in Fig. 1.

2 Proposed method

Our face hallucination model consists of two main components: *i) a generative SR network* for image upscaling, and *ii)* an ensemble of face recognition models that serve as identity priors. In the following sections we describe all components of the proposed model in detail and elaborate on the training procedure used to learn the model parameters.

2.1 The cascaded SR network

The generative part of our proposed model is a 53-layer deep convolutional neural network (CNN) that takes a LR facial image as input and super-resolves it at a magnification factor of $8\times$. The network progressively upscales the images using a cascade of *SR modules*. Each module upscales the image by a factor of $2\times$, which allows us to apply a loss function on the intermediate results. The cascaded architecture allows us to solve a series of better conditioned problems using repeated bottom-up inference with top-down supervision instead of one complex problem with an overwhelming amount of possible solutions.

We design our SR network around a fully-convolutional architecture that relies heavily on residual blocks [16] for all processing within one SR module and sub-pixel convolutions [17] for image upscaling. Our design choices are motivated by the success of fully-convolutional CNN models in various vision problems [16, 18, 19] and the state-of-the-art performance ensured by the sub-pixel convolutions in prior SR work [17, 10]. Similarly to [10], the residual blocks of the SR modules consist of two (convolution, batch-norm, activation) sub-blocks, followed by a post-activation element-wise sum.

The network branches off after each SR module to allow for intermediate top-down supervision during training. Each branch applies a series of large-filter convolutions to produce intermediate SR resolution results at different scales (i.e., $2\times$ and $4\times$ the initial scale) that are incorporated into the loss functions discussed in Section 2.3. However, these branches are not used at test time. The entire architecture of our network is illustrated in detail in Fig. 2.

2.2 The identity prior

Using prior information to constrain the solution space of SR models during training is a key mechanism in the area of super-resolution [14, 15, 13]. The main motivation for incorporating priors into SR models is to provide a source of additional information for the learning procedure that complements the commonly used reconstruction-oriented objectives and contributes towards sharper and more accurate SR results.

Identity is an exceptionally strong prior in this context. In fact, it seems intuitive to think about SR in terms of both *i) an image enhancement* and *ii) content preservation*. While the image enhancement perspective is covered in our model by a reconstruction-based loss, the content preservation aspect is addressed through an ensemble of face recognition models that ensure that identity information is preserved.

We associate each recognition model with one of the SR modules and use it as an identity prior for the corresponding SR output, as illustrated in Fig. 2. Since each SR module can be shown to add only high-frequency details to the input images, we train all recognition models to respond only to the hallucinated details and ignore the low-resolution content that is shared by the input and SR images. By focusing exclusively on the added details, we are able to directly link the recognition models to the desired SR outputs and penalize the results in case they alter the facial identity. This mechanism allows us to learn the parameters of the SR network by considering an identity-dependent loss in the overall learning objective.

We use SqueezeNet [20] models for this work. The main reason for our choice is the lightweight architecture of SqueezeNet, which does not impose significant runtime slowdowns due to its relatively small memory and FLOPS footprint.

2.3 Training details and SSIM loss

We train the model in two stages. In the first stage, we learn the parameters of the SqueezeNet models for all three SR outputs. In the second stage, we freeze the the weights of the recognition models and train the SR network with a combined loss.

Recognition-model training. Next to LR and HR image pairs, we also require two intermediate reference images between the lowest and the highest resolution to learn the parameters of the recognition models and SR modules. To this end, we apply a simple degradation model on the available HR images $\mathbf{x}_i^{hr}$ and generate N image quadruplets for training, i.e., $\{\mathbf{x}_i^{lr}, \mathbf{x}_i^{2\times}, \mathbf{x}_i^{4\times}, \mathbf{x}_i^{hr}\}$, where $\mathbf{x}_i^{lr}$ represents the LR input image, $\mathbf{x}_i^{2\times}$ and $\mathbf{x}_i^{4\times}$ stand for the intermediate SR results at $2\times$ and $4\times$ magnification factors, respectively, and the HR image $\mathbf{x}_i^{hr}$ corresponds to the ground truth for the magnification factor of $8\times$. Our degradation model uses Gaussian blurring followed by image decimation for down-sampling to produce training data.

To train the recognition models, we construct residual images that reflect the facial details that need to be learned by the SR modules. The residual images are computed by smoothing the ground truth images by a Gaussian kernel and subtracting the smoothed image from the original, i.e., $\Delta\mathbf{x}_i^j = \mathbf{x}_i^j - \mathbf{g} * \mathbf{x}_i^j$, for $j \in \{2\times, 4\times, hr\}$, where σ values of $\sigma_{2\times} = 1/3$, $\sigma_{2\times} = 1$ and $\sigma_{8\times} = 7/3$ are used with images at $2\times, 4\times$, and $8\times$ the LR image size, respectively. We train the SqueezeNet models based on the generated residual images using categorical cross-entropy L_{CE}:

$$L_{CE}(\theta_{SN}, \Delta\mathbf{x}) = -\sum_{k=1}^{K} p_{\Delta\mathbf{x}}(k)\, log\, \hat{p}_{\Delta\mathbf{x}}(k), \qquad (1)$$

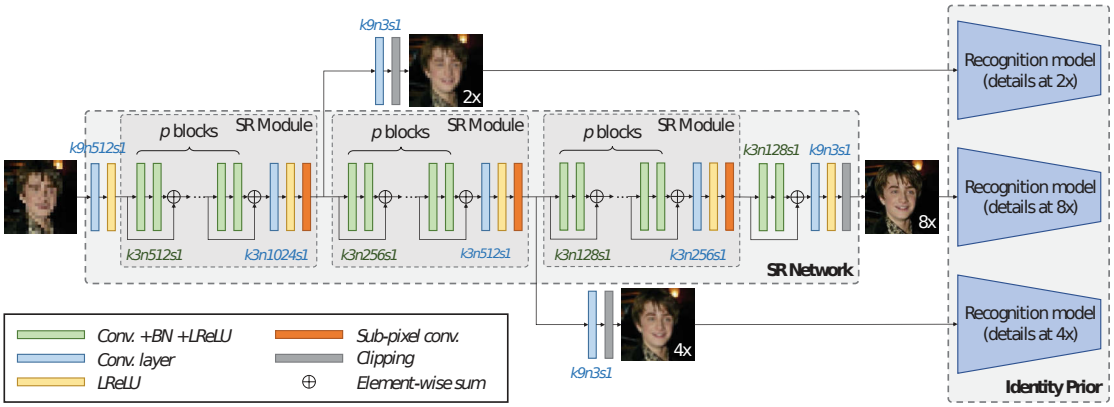


Figure 2: The proposed model consists of a generative SR network and face recognition models that serve as identity priors.

where $p_{\Delta\mathbf{x}}$ denotes the ground truth class probability distribution of the residual image $\Delta\mathbf{x}$ (i.e., $p_{\Delta\mathbf{x}} \in \{0,1\}^K$ is a class-encoded one-hot vector), $\hat{p}_{\Delta\mathbf{x}} \in \mathbb{R}^K$ stands for the output probability distribution produced by SqueezeNet's softmax layer based on $\Delta\mathbf{x}$, i.e., K stands for the number of classes in the training data and θ_{SN} represents the parameters of the network. We train the parameters of all three recognition models by minimizing the L_{CE} loss over the training dataset. The result of this training stage are three face recognition models $\hat{\theta}^{2\times}_{SN}, \hat{\theta}^{4\times}_{SN}, \hat{\theta}^{hr}_{SN}$, that serve as identity constraints for the SR network. We use the Adam [21] algorithm for training, with a batch size of 128 and an initial learning rate of 10^{-4}. The learning rate is multiplied by a factor of $\frac{1}{3}$ every 20 epochs. To avoid over-fitting, use random horizontal flipping and random crops for data augmentation.

SR network training. Standard reconstruction oriented loss functions used for training SR models, such as MSE or MAE, are known to produce overly smooth and often blurry SR results [10]. We therefore design a new loss function for our SR network around the structural similarity index (SSIM, [22]), and integrate it directly into our learning algorithm. Specifically, we use our SSIM approximation as a loss function for the hallucination model.

Given a ground truth image $\mathbf{x}$ and the corresponding SR network prediction $\hat{\mathbf{x}} = f_{\theta_{SR}}(\mathbf{x})$, we compute the SSIM-based loss as follows:

$$L_{SSIM}(\theta_{SR}, \mathbf{x}) = \frac{1}{2}\left(1 - \mathbb{E}_x\left[\hat{SSIM}(\mathbf{x}, \hat{\mathbf{x}})\right]\right), \quad (2)$$

where the SR network f is parametrized by θ_{SR}, $\mathbb{E}_x[\cdot]$ stands for the expectation operator over the spatial coordinates and $\hat{SSIM}(\mathbf{x}, \hat{\mathbf{x}})$ is a spatial similarity map between $\mathbf{x}$ and $\hat{\mathbf{x}}$ defined as:

$$\hat{SSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{(2\mu_{12} + C_1) \odot (2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1) \odot (\sigma_1^2 + \sigma_2^2 + C_2)}, \quad (3)$$

where we model the means μ_1 and μ_2 as convolutions of local patches with a Gaussian kernel, μ_{12} as their entry-wise product, and σ_1^2, σ_2^2 and σ_{12} as $\mathbf{xx}$, $\hat{\mathbf{x}}\hat{\mathbf{x}}$ and $\mathbf{x}\hat{\mathbf{x}}$ convolved with the same kernel, respectively.

The open parameters C_1 and C_2 are defined as per the SSIM reference implementation provided by the authors of [23], i.e., $C_1 \approx 6.55$, $C_2 \approx 58.98$.

Based on the pre-trained SqueezeNet models and the loss introduced above, we defined the overall loss of our face hallucination model as follows:

$$L(\theta_{SR}, \{\mathbf{x}^j\}) = \sum_{j \in \mathcal{D}} L_{SSIM}(\theta_{SR}, \mathbf{x}^j) + \alpha L_{CE}(\theta^j_{SN}, \Delta\mathbf{x}^j), \quad (4)$$

where $\mathcal{D} = \{2\times, 4\times, hr\}$, α is a weight parameter that balances the relative impact of the reconstruction- and recognition-based losses and θ_{SR} stands for the parameters of the SR network that we aim to learn.

We again use the Adam [21] algorithm for training, minimizing (4) with $\alpha = 0.001$. Due to the large memory footprint of the SR network and the face recognition models, we use a relatively small batch size of 8. We set the initial learning rate to $\frac{10}{3} \times 10^{-3}$ and multiply it by $\frac{1}{3}$ at the end of epochs 10, 25, 50 and 80.

Once the training is complete, we remove the recognition models and network branches used to generate the intermediate SR results. The final SR network takes an image $\mathbf{x}_{lr}$ of size 24×24 pixels as input and returns a 192×192 facial image $\mathbf{x}_{hr}$.

3 Experiments

We select two datasets for our experiments. To train the model we use the CASIA WebFace dataset [24] which features $494,414$ images of $10,575$ identities, (i.e., $N = 494,414$; $K = 10,575$). The CASIA WebFace images are blurred and sub-sampled to produce the necessary image quadruplets for training the recognition models and the SR network. For testing, we use the Labeled Faces in the Wild (LFW) [25] dataset with $13,233$ facial images and $5,749$ subjects. The two datasets are selected for the experiments because they feature images of variable quality captured in unconstrained conditions and thus represent a significant challenge for SR models. More importantly, they are designed to contains zero overlap in terms of identity, which is paramount to ensure a fair and unbiased evaluation of the model.

We compare our proposed model with 6 SR and face hallucination models, i.e.: the Naive Bayes Super-Resolution



Figure 3: Qualitative comparison of state-of-the-art SR models on sample images from the LFW dataset.

Table 1: PSNR and SSIM scores over the LFW dataset.

Model	Bicubic	NBSRF [8]	SRCNN [7]	VDSR [5]
PSNR	24.256	25.092	24.812	25.415
SSIM	0.7060	0.7268	0.7187	0.7411
Model	**ℓ_p [9]**	**SRGAN [10]**	**URDGN [26]**	**our model**
PSNR	26.985	25.669	25.575	**27.164**
SSIM	0.7903	0.6993	0.7516	**0.8171**

Forest (NBSRF) from [8], the Super-Resolution Convolutional Neural Network (SRCNN) from [7], the Very Deep Super Resolution Network (VDSR) from [5], the perceptual-loss based SR model (ℓ_p) from [9], the Super-Resolution Generative Adversarial Network from [10], and the Ultra Resolving Discriminative Generative Network (URDGN) from [26]. We train all models with the same data and use open-source implementations of the authors (where available) for a fair comparison. For ℓ_p we use features from the fire2, fire3 and fire4 layers of SqueezeNet for the learning criterion. We include results for bicubic interpolation as a baseline. We present the quantitative results in terms of average PSNR and SSIM scores in the table 1. A few sample SR images are presented in Fig. 3. We see that with magnification factors of $8\times$, interpolation methods are insufficient and result in the loss of facial details. Furthermore, general SR models, such as NBSRF, SRCNN and VDSR, fail to provide substantial improvements and are seen to amplify noise present in the LR images. These models fail to make use of the available facial context due to their relatively low receptive fields. The SRGAN, URDGN and ℓ_p models improve on this by including secondary networks as constraints during SR training. ℓ_p is consistently the best-performing model included in our comparison, only slightly behind our model. However, we notice it often adds high-frequency noise when trying to minimize the perceptual loss of the convolutional maps of the secondary network. We speculate the reason our model is not susceptible to these errors is the global cross-entropy loss of the secondary networks as opposed to the local conv features exploited by ℓ_p.

References

[1] S. Baker and T. Kanade, "Hallucinating faces," in *FG*, 2000.

[2] C. Liu, H. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *IJCV*, 2007.

[3] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *TPAMI*, 2002.

[4] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *TIP*, 2003.

[5] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.

[6] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *ICCV*, 2013.

[7] C. Dong, C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.

[8] J. Salvador and E. Perez-Pellitero, "Naive Bayes super-resolution forest," in *ICCV*, 2015.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.

[11] M. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.

[12] Y. Tai, J Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017.

[13] T. Cho, C. Zitnick, N. Joshi, SB Kang, R. Szeliski, and WT Freeman, "Image restoration by matching gradient distributions," *TPAMI*, 2012.

[14] Y. Wang, W. Yin, and Y. Zhang, "A fast algorithm for image deblurring with total variation regularization," 2007.

[15] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *CVPR*, 2007.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[17] S. Shi, J. Caballero, F. Huszár, J. Totz, A. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[19] O. Parkhi, A. Vedaldi, A. Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015.

[20] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5 MB model size," *arXiv preprint arXiv:1602.07360*, 2016.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[22] Z. Wang, E. Simoncelli, and AC Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, 2003.

[23] Z. Wang, AC Bovik, HR Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.

[24] D. Yi, Z. Lei, S. Liao, and SZ Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[25] GB Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[26] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *ECCV*, 2016.

Fast Spatially Regularized Correlation Filter Tracker

Alan Lukežič, Luka Čehovin Zajc, Matej Kristan

Faculty of Computer and Information Science, University of Ljubljana
{*alan.lukezic, luka.cehovin, matej.kristan*}*@fri.uni-lj.si*

Abstract

Discriminative correlation filters (DCF) have attracted significant attention of the tracking community. Standard formulation of the DCF affords a closed form solution, but is not robust and constrained to learning and detection using a relatively small search region. Spatial regularization was proposed to address learning from larger regions. But this prohibits a closed form solution and leads to an iterative optimization with significant computational load, resulting in slow model learning and tracking. We propose to reformulate the spatially regularized filter cost function such that it offers an efficient optimization. This significantly speeds up the tracker (approximately 14 times) and results in real-time tracking at the same or better accuracy.

1 Introduction

Visual object tracking is a task of continuous target localization given an initial position in first frame of the video sequence. Significant advancements have been made in the field over the recent years, largely due to a number publicly available benchmarks and steps towards standardized performance evaluation [1, 2, 3, 4].

Most recent developments have focused on discriminative correlation filters (DCF), which were primarily proposed for object detection [5]. The idea of learning discriminative correlation filters for tracking has been promoted by the seminal paper of Bolme et al. [6]. Further improvements by other authors included correlation filter with kernels [7], multiple-channel formulation [8] and scale estimation with DCF [9, 10].

In DCF tracking the target is localized using a filter which is learned on a pre-defined (Gaussian) response on the training image. The standard formulation of DCF uses circular correlation which allows to implement filter learning efficiently by Fast Fourier transform (FFT) but requires the filter and the patch size to be equal which limits the detection range. Due to the circularity, the filter is trained on many examples that contain unrealistic, wrapped-around circularly shifted versions of the target. These windowing problems were recently addressed by [11, 12] using the boundary constraints and by Danneljan et al. [13] who introduce spatial regularization to penalize filter values outside the object boundaries. All these approaches train from image patches larger than the object and thus increase the detection range.

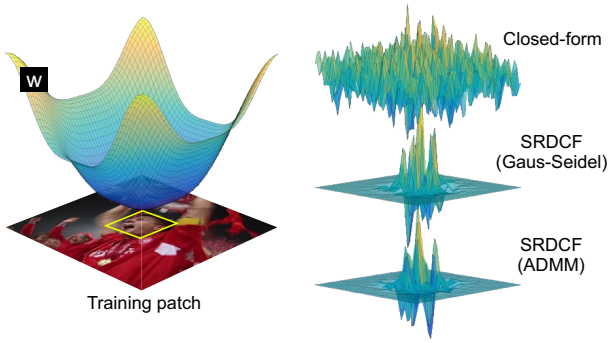


Figure 1: Training patch and spatial weight function are shown on the left. Yellow rectangle represents target region. Three filters are shown on the right (top to bottom): filter calculated with a closed form solution (Equation 1), and two filters obtained with iterative optimization methods Gauss-Seidel and ADMM optimization.

Spatially regularized discriminative correlation filter (SRDCF) [13] shows a great potential and trackers using this method [14, 15] achieve excellent performance on recent benchmarks. But the spatial regularization prevents using a closed form solution as in standard formulation [6] and iterative optimization has to be used. This optimization is based on Gauss-Seidel steps which simultaneously solves a large number of equations and leads to a very slow solution. Despite excellent performance, SRDCF is of limited use in practical application since it runs far below real-time.

We propose to reformulate the spatially regularized filter cost function such that it affords an efficient optimization by using alternating direction method of multipliers (ADMM) [16]. Since all steps in optimization are performed pixel-wise, this significantly speeds up the tracker (approximately 14 times) and results in real-time tracking of approximately 29fps at the same or better accuracy.

2 Discriminative Correlation Filters

In the standard DCF formulation [6], the filter $\mathbf{h}$ is trained to output a desired response $\mathbf{g}$ when correlated with a

training sample $\mathbf{f}$, i.e., $\mathbf{g} = \sum_{i=1}^{d} \mathbf{h}_i * \mathbf{f}_i$, where $*$ denotes a circular correlation and the training sample $\mathbf{f}$ consists of d-channel feature maps, therefore $\mathbf{h}$ contains d channels. The following cost is minimized in filter learning:

$$\epsilon(\mathbf{h}) = || \sum_{i=1}^{d} \mathbf{f}_i * \mathbf{h}_i - \mathbf{g}||^2 + \lambda \sum_{i=1}^{d} ||\mathbf{h}_i||^2 \quad (1)$$

An efficient closed-form solution for (1) is obtained via the Fourier domain:

$$\hat{\mathbf{h}}_i = \frac{\hat{\mathbf{f}}_i \odot \overline{\hat{\mathbf{g}}}}{\sum_{i=1}^{d} \overline{\hat{\mathbf{f}}}_i \odot \hat{\mathbf{f}}_i + \lambda}, \quad (2)$$

where $\hat{(\,)}$ represents Fourier transform, i.e., $\hat{a} = \mathcal{F}(a)$, the symbol $\overline{(\,)}$ is complex-conjugate operation and $\odot$ is a Hadamard product. Target is localized by finding position of the maximum peak in the correlation response $\mathbf{r}$:

$$\mathbf{r} = \mathcal{F}^{-1}\Big(\sum_{i=1}^{d} \hat{\mathbf{h}}_i \odot \hat{\mathbf{f}}_i \Big), \quad (3)$$

where $\mathcal{F}^{-1}$ represents an inverse Fourier transformation.

2.1 Spatially Regularized DCF

The SRDCF [13] tracker introduces a spatial weight function $\mathbf{w}$ which penalizes large filter values further away from the target center. It avoids the windowing problem due to the circular correlation in Fourier domain and it reduces impact of the background in the filter. The visualization of the penalty function $\mathbf{w}$ and the resulting filter is shown in the Figure 1. Penalty function is incorporated in a standard DCF formulation (1) by adding a spatial regularization term, i.e.,

$$\epsilon(\mathbf{h}) = \| \sum_{i=1}^{d} \mathbf{f}_i * \mathbf{h}_i - \mathbf{g}\|^2 + \lambda_1 \sum_{i=1}^{d} \|\mathbf{w}\mathbf{h}_i\|^2 + \lambda_0 \sum_{i=1}^{d} \|\mathbf{h}_i\|^2. \quad (4)$$

An iterative Gauss-Seidel optimization is used to solve a linear system of equations resulting from (4). The size of the linear problem is $dMN \times dMN$, where M, N, and d are width, height and number of feature channels, respectively, resulting in slow computation despite carrying it out in Fourier domain. We refer the reader to [13] for complete derivation of the optimization.

3 Proposed Method

We propose to reformulate the cost function (4), which can be decomposed into independent per-pixel (or per-frequency in the relevant Fourier-transformed terms) subproblems,

$$\epsilon(\hat{\mathbf{h}}) = \sum_{j=1}^{N} \Big(\|\hat{\mathbf{f}}_j^T \overline{\hat{\mathbf{h}}}_j - \hat{\mathbf{g}}_j\|^2 + \lambda_1 \|\mathbf{w}_j \mathbf{h}_j\|^2 + \lambda_0 \|\hat{\mathbf{h}}_j\|^2 \Big), \quad (5)$$

where N is number of pixels. Note that $\mathbf{h}_j$ and $\mathbf{f}_j$ are d-dimensional column vectors and $\mathbf{g}_j$ and $\mathbf{w}_j$ are scalars on j-th pixel. The subscript j is omitted in the following due to the compactness of the notation. An auxiliary variable $\boldsymbol{\omega}$ is introduced and the cost function (5) is augmented with the constraint $\boldsymbol{\omega} \equiv \mathbf{h}$, leading to an augmented Lagrangian formulation

$$\begin{aligned} \mathcal{L}(\hat{\mathbf{h}}, \mathbf{l}, \boldsymbol{\omega}) = \|\hat{\mathbf{f}}^T \overline{\hat{\mathbf{h}}} - \hat{\mathbf{g}}\|^2 + \lambda_1 \|\mathbf{w}\boldsymbol{\omega}\|^2 + \\ + \lambda_0 \|\hat{\mathbf{h}}\|^2 + \mathbf{l}^T(\mathbf{h} - \boldsymbol{\omega}) + \frac{\mu}{2}\|\mathbf{h} - \boldsymbol{\omega}\|^2, \end{aligned} \quad (6)$$

where $\mathbf{l}$ is a Lagrange multiplier and $\mu > 0$ is a parameter.

The cost function (6) is solved following the alternate direction method of multipliers (ADMM) [16] by alternatingly solving two subproblems:

$$\nabla_{\overline{\hat{\mathbf{h}}}} \mathcal{L} \equiv 0 \quad (7)$$

$$\nabla_{\boldsymbol{\omega}} \mathcal{L} \equiv 0. \quad (8)$$

The first derivative in (7) is equivalent to $\nabla_{\overline{\hat{\mathbf{h}}}} \mathcal{L} = \hat{\mathbf{f}}\hat{\mathbf{f}}^H \hat{\mathbf{h}} - \hat{\mathbf{f}}\overline{\hat{\mathbf{g}}} + \lambda_0 \hat{\mathbf{h}} + \frac{\mu}{2}\hat{\mathbf{h}} - \frac{\mu}{2}\hat{\boldsymbol{\omega}}$, where $(\cdot)^H$ is a Hermitian transpose. Setting the derivative to zero yields the following solution

$$\hat{\mathbf{h}} = (\hat{\mathbf{f}}\hat{\mathbf{f}}^H + \lambda_2)^{-1}(\hat{\mathbf{f}}\overline{\hat{\mathbf{g}}} + \frac{\mu}{2}\hat{\boldsymbol{\omega}}), \quad (9)$$

where $\lambda_2 = \lambda_0 + \frac{\mu}{2}$.

The first part of (9) represents a matrix inverse calculated for each pixel, which is computationally demanding. This can be avoided by using a Sherman-Morrison-Woodbury formula [17] leading to

$$\Big(\hat{\mathbf{f}}\hat{\mathbf{f}}^H + \lambda_2\Big)^{-1} = \frac{1}{\lambda_2}\Big(\mathbf{I} - \frac{\hat{\mathbf{f}}\hat{\mathbf{f}}^H}{\hat{\mathbf{f}}^H\hat{\mathbf{f}} + \lambda_2}\Big), \quad (10)$$

which simplifies (9) into

$$\hat{\mathbf{h}} = \frac{\hat{\mathbf{f}}\overline{\hat{\mathbf{g}}} - \frac{\mu}{2\lambda_2}\hat{\mathbf{f}}\hat{\mathbf{f}}^H\hat{\boldsymbol{\omega}}}{\hat{\mathbf{f}}^H\hat{\mathbf{f}} + \lambda_2} + \frac{\mu}{2\lambda_2}\hat{\boldsymbol{\omega}}. \quad (11)$$

The derivative in (8) yields the following solution: $\nabla_{\boldsymbol{\omega}} \mathcal{L} = 2\lambda_1 \mathbf{w}\boldsymbol{\omega} - \mathbf{l} - \mu\mathbf{h} - \frac{\mu}{2}\boldsymbol{\omega}$. Setting it to zero, leads to

$$\boldsymbol{\omega} = \frac{\mathbf{l} + \mu\mathbf{h}}{2\lambda_1 \mathbf{w} - \frac{\mu}{2}}. \quad (12)$$

Since $\hat{\mathbf{h}}$, $\boldsymbol{\omega}$ and $\mathbf{l}$ are used in (11) and (12) the final solution for filter $\mathbf{h}$ is calculated iteratively, i.e.,

$$\hat{\mathbf{h}}^{(k)} = \frac{\hat{\mathbf{f}}\overline{\hat{\mathbf{g}}} - \frac{\mu}{2\lambda_2}\hat{\mathbf{f}}\hat{\mathbf{f}}^H\hat{\boldsymbol{\omega}}^{(k-1)}}{\hat{\mathbf{f}}^H\hat{\mathbf{f}} + \lambda_2} + \frac{\mu}{2\lambda_2}\hat{\boldsymbol{\omega}}^{(k-1)}, \quad (13)$$

$$\boldsymbol{\omega}^{(k)} = \frac{\mathbf{l}^{(k-1)} + \mu\mathbf{h}^{(k)}}{2\lambda_1\mathbf{w} - \frac{\mu}{2}}, \quad (14)$$

$$\mathbf{l}^{(k)} = \mu(\mathbf{h}^{(k)} - \boldsymbol{\omega}^{(k)}), \quad (15)$$

where k represents the iteration index.

4 Tracking with augmented SRDCF

A tracking iteration in each frame consists of two main steps: target localization and update of the visual model. Both steps are described in the following.

Target localization. When a new frame arrives at time step t, an image patch four times larger than target

size is extracted from the image, centered on the target position from previous time step p_{t-1}. A 41-channel feature map $\mathbf{f}_t$ is calculated from the extracted image patch, corresponding to 31 HoG [18] and 10 colornames [8] channels. Circular correlation between the feature channels $\mathbf{f}_t$ and filter from previous time-step $\mathbf{h}_{t-1}$ is calculated in the Fourier domain (3). Position of the correlation output maximum $\mathbf{p}_t$ presents the new target position. Change of target size is estimated using the approach from [10].

Update. Target appearance is expected to change during tracking due to the numerous effects like, rotation, deformation or scale change of the target. The visual model is therefore updated after the target is localized. An image patch four times larger than target size is extracted from image centered at $\mathbf{p}_t$ and a feature map $\mathbf{f}_t$ is calculated. A new filter $\tilde{\mathbf{h}}_t$ is calculated using method described in Section 3. Filter from the previous time-step is denoted as $\mathbf{h}_{t-1}$ and it is combined with $\tilde{\mathbf{h}}_t$ using a moving average, resulting in an updated filter

$$\hat{\mathbf{h}}_t = (1-\eta)\hat{\mathbf{h}}_{t-1} + \eta\hat{\tilde{\mathbf{h}}}_t. \tag{16}$$

This filter is used in the next time-step to localize the target. Parameter $\eta = 0.01$ is a learning rate which controls the impact of the new filter in $\hat{\mathbf{h}}_t$. The update is performed in the Fourier domain for efficiency.

5 Experiments

The proposed tracker, the augmented SRDCF (SRDCF$_\mathrm{A}$) is compared to the original SRDCF [13] on VOT16 [2] (Section 5.1) and on OTB100 [4] (Section 5.2). Speed comparison is given in Section 5.3

5.1 VOT16 dataset

The VOT16 [2] dataset consists of 60 sequences which are chosen so that the dataset is moderately large, but rich in attributes relevant for tracking. A standard reset-based experiment [1] is performed by running a tracker on each sequence until overlap of the predicted target position with the ground-truth is zero. This event is called failure and after that the tracker is reinitialized. Tracking performance is measured by accuracy, robustness and expected average overlap (EAO). Robustness measures average number of failures and accuracy represents average overlap. Expected average overlap combines both measures into the single number and it can be interpreted as expected overlap at the typical sequence length in short-term tracking scenario.

In the Table 1 tracking performance of SRDCF$_\mathrm{A}$ is comparable to the original SRDCF. The SRDCF$_\mathrm{A}$ has 6% less failures, a lower average overlap (4%), but a 4% better expected average overlap. The similar performance indicates that the filter calculated with ADMM optimization is comparable to the filter calculated with Gauss-Seidel optimization for tracking tasks.

5.2 OTB100 dataset

The OTB100 [4] dataset consists of 100 sequences. A tracker is in each sequence initialized on the first frame and run to the end, which is called a no-reset experiment. Tracking performance is measured by the area-under-the-curve (AUC) on the success and precision plots, shown in Figure 2. The success plot shows overlap threshold values on x and the proportion of frames with the overlap between the predicted and ground truth bounding boxes as greater than a threshold on y. The precision plot shows a similar statistics computed from the center error.

Results are shown in the Figure 2. The difference between the original SRDCF and the SRDCF$_\mathrm{A}$ is negligible (approximately 1% better AUC on both graphs). This further supports the equivalence of the learned filters by the slow Gauss-Seidel and our fast ADMM formulation from tracking performance perspective.

Table 1: Accuracy (Acc.), robustness (Rob.) and expected average overlap (EAO) for SRDCF and SRDCF$_\mathrm{A}$ on the VOT16 dataset. The sign ↑ denotes *high is better* and ↓ denotes *low is better*.

Tracker	Acc. ↑	Rob. ↓	EAO ↑
SRDCF$_\mathrm{A}$	0.51	1.42	0.2520
SRDCF	0.53	1.52	0.2431

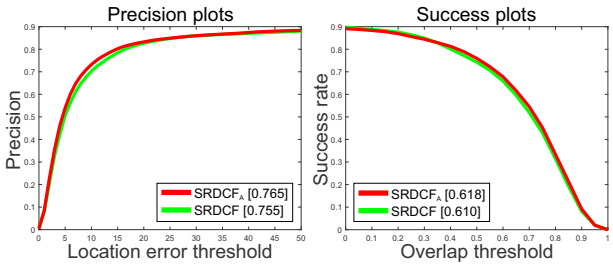


Figure 2: Tracking performance on the OTB100 [4] dataset. The area-under-the-curve (AUC) is shown in brackets in the legend.

5.3 Speed comparison

Our main motivation behind reformulation of the SRDCF cost function (6) is computational efficiency. The original tracker runs on average at 2 frames-per-second which is far from real-time and it presents a serious limitation for practical tracker application. In this section we compare speed measurements of the original SRDCF and the proposed SRDCF$_\mathrm{A}$ which uses the ADMM optimization. Both trackers are run on the 60 sequences of VOT16 [2] dataset using a no-reset experiment and tracking speed is averaged over all sequences.

In each frame time needed to process the whole frame is measured including target localization and filter calculation stage, without image loading from disk. Additionally, filter calculation step is reported separately, to demonstrate the difference between the both optimization methods. Since original SRDCF requires a significant pre-computation overhead performed only in the initialization frame, time needed to process the first frame is

reported separately. All experiments were conducted on the same desktop Intel i7 6700 CPU, 3.4 GHz and 16GB RAM computer.

Results of the speed comparison are presented in the Table 2. Time for the initialization required by the SRDCF tracker is almost 2 seconds while ADMM-based SRDCF_A takes only 47 milliseconds, which is more than 40-times faster. Considering only filter calculation step, the Gauss-Seidel optimization takes in average 329 milliseconds, while ADMM optimization needs only 9 milliseconds. Average per-frame speed of the original SRDCF is 2FPS (480ms per frame), while SRDCF_A achieves over 14-times faster performance running at 29FPS (34ms per frame).

Table 2: The average times needed for tracker initialization and filter calculation are shown in the first two columns. Average tracking speed is presented in milliseconds and in frames-per-second (third and fourth column).

Tracker	**Init.** [ms]	**Filter** [ms]	**Average** [ms]	**Average** [FPS]
SRDCF_A	47	9	34	29
SRDCF	1814	329	480	2

6 Conclusion

In this paper an optimization method for spatially regularized discriminative correlation filter based on the ADMM method [16] is derived. The optimization is used to minimize the cost function from the popular SRDCF [13] tracker. We show experimentally that tracking performance of our version of SRDCF using ADMM optimization method is slightly better than the original method. In addition, the speed of the proposed tracker is more than fourteen times faster than original version, running in real-time at 29 frames-per-second. In our future work we plan to test different spatial weight functions and to incorporate learning from multiple training samples.

Acknowledgements

This work was supported in part by the following research programs and projects funded by the Slovenian research agency ARRS: project ViAMaRo (J2-8175) and research program P2-0214.

References

[1] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[2] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, T. Vojir, G. Häger, A. Lukežič, and G. et al. Fernandez, "The visual object tracking vot2016 challenge results," in *Proc. European Conf. Computer Vision*, 2016.

[3] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, and G. Fernandez, "The visual object tracking vot2017 challenge results," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sept 2015.

[5] C. F. Hester and D. Casasent, "Multivariant technique for multiclass pattern recognition," *Applied Optics*, vol. 19, no. 11, pp. 1758–1761, 1980.

[6] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Comp. Vis. Patt. Recognition*. IEEE, 2010, pp. 2544–2550.

[7] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, 2015.

[8] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 1090–1097.

[9] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. European Conf. Computer Vision*, 2014, pp. 254–265.

[10] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, 2017.

[11] H. Kiani Galoogahi, T. Sim, and S. Lucey, "Correlation filters with limited boundaries," in *Comp. Vis. Patt. Recognition*, 2015, pp. 4630–4638.

[12] A. Lukežič, T. Vojíř, L. Čehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Comp. Vis. Patt. Recognition*, 2017, pp. 6309–6318.

[13] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Int. Conf. Computer Vision*, 2015, pp. 4310–4318.

[14] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: learning continuous convolution operators for visual tracking," in *Proc. European Conf. Computer Vision*. Springer, 2016, pp. 472–488.

[15] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Comp. Vis. Patt. Recognition*, 2017, pp. 6638–6646.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2012.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Comp. Vis. Patt. Recognition*, vol. 1, June 2005, pp. 886–893.

Quantifying uncertainty with the bootstrap: introduction and simulation study

Greta Gašparac[1], **Erik Štrumbelj**[1]

[1] *University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia*
E-mail: greta.gasparac@gmail.com, erik.strumbelj@fri.uni-lj.si

Abstract

We introduce the reader to the bootstrap, a simple and flexible resampling-based alternative for quantifying uncertainty. We describe the basic characteristics of the non-parametric bootstrap and illustrate its practical behaviour with simulations in the context of a typical task in machine learning - estimating and comparing the performance of different prediction models. We compare the standard normal, percentile and BC_a bootstrap confidence intervals. As theory suggests, the BC_a performs the best over a wide range of situations.

1 Introduction

Empirical research is an integral part of science. However, we must be aware that any result that is based on measurements has a certain degree of uncertainty. Not accounting for this uncertainty can lead to incorrect and possibly misleading conclusions. This is why it is very important to quantify uncertainty and why statistics play such an important role in modern science.

Our motivation is the common task in machine learning - estimating a model's performance (predictive accuracy, speed, ...) and/or comparing it with other models on one or more different scenarios. In many cases researchers do not even attempt to quantify the uncertainty in their estimates. When they do, they typically use either confidence intervals (CI) based on the normal distribution or null-hypothesis statistical testing (NHST) [3].

Performance data are typically not distributed normally, so CIs based on the normal distribution make sense only in cases when we reason about the mean and the sample size is not too small. The same applies to most parametric NHST which are based on the normal distribution. Using non-parametric NHST alleviates this, but is more difficult to apply and interpret correctly. Furthermore, most standard approaches apply directly only to the mean/median and cannot be readily used to quantify the uncertainty of more complicated functions (for example, the root-mean-square error or a correlation coefficient).

In this paper we advocate the use of the bootstrap for quantifying the uncertainty in our estimates. The method was introduced by Efron [4] and is a representative of resampling-based methods, which also include, among others, permutation tests and cross-validation. Most of these methods have been known for decades, but have been limited by their computationally-intensive nature. Today, however, computation has become very cheap (in particular, very cheap compared to learning how to correctly perform classical statistical analysis), and the full potential and flexibility of resampling-based methods can be realized. Cross-validation has already been widely adopted by the machine learning community as a means for estimating how the model's performance will generalize to new and unseen data. Permutation tests and the bootstrap are, however, in our opinion, still heavily under-utilized. Both due to their usefulness and their pedagogical advantage over NHST.

Bootstrap is a resampling method used for estimating standard errors, bias, computing CI and for constructing NHST. Theoretically complicated methods are replaced by computer simulation, which is much more intuitive and easier to apply. Furthermore, it is more widely applicable - the process remains the same no matter what our statistic of interest is and number of assumptions is minimized. There are many different variants of the bootstrap, but the content of this paper applies to the most common one - the non-parametric bootstrap. We aim to provide a better understanding of the method, point out the advantages it offers, and demonstrate its weaknesses.

The remainder of the paper is organized as follows. In Section 2 we explain the idea behind the bootstrap. Section 3 offers a brief overview of bootstrap CI. In Section 4 we present empirical evidence of some of the method's characteristics, which are relevant for practical application. We see how changing the number of generated bootstrap samples affects the bootstrap distribution. We also compare coverage of different CI for mean, median and the 95-th percentile when our data come from different distributions and different sample sizes. With 5 we conclude the paper and offer some direction for further work.

2 The bootstrap

Suppose $\mathbf{x} = (x_1, ..., x_n)$ is a sample of size n from population F. Let θ be the population mean, our statistic of interest, and we want to compute the 95% CI.

Assuming normality and relying on the central limit theorem, our CI is $\pm 1.96 \cdot se$, where se is the standard error of the mean. But what happens if our statistic of interest is the median or correlation coefficient? What happens if our data are not distributed normally? We could

get those values from the sampling distribution of θ, but in order to obtain such a distribution, one we would need much more than just one sample, which is in practice typically impossible.

2.1 The basic idea behind the bootstrap

If **x** is a set of identically distributed and independent observations from F, it should be a good estimate of the population. Instead of sampling from the original population F, we will sample from an estimate of the population $\hat{F}$, which we will construct from **x**. That is, we will use the so-called *plug-in principle*, where we substitute an unknown variable for an estimate. Based on what type of bootstrap we choose, $\hat{F}$ can take different forms. For non-parametric bootstrap it is the empirical distribution function (EDF)

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=0}^{n} I(x_i = x), \tag{1}$$

which assigns probability $\frac{1}{n}$ to each data point (sampling with replacement). With sampling from $\hat{F}$ we obtain bootstrap samples. In total, there are n^n different samples. Evaluating all of them is known as the theoretical bootstrap. In practice, however, this is infeasible, so we use the Monte Carlo implementation and draw B (usually $B > 10^4$) bootstrap samples $\mathbf{x}^{*1}, ..., \mathbf{x}^{*B}$. For each of them we calculate the test statistic and get the bootstrap sampling distribution $\theta^{*1}, ..., \theta^{*B}$. In order for our results to reflect the actual data, we normally draw bootstrap samples of size n. On account of the EDF representing the whole population, the bootstrap distribution is usually too narrow by approximately $\sqrt{(n-1)/n}$ [6].

3 Bootstrap confidence intervals

Efron and Tibshirani [5] provide great insight into different ways the bootstrap can be used to compute CI. Below, we describe three: a normal distribution-based approach, a first-order accurate approach, and a second-order accurate approach[1].

3.1 Standard normal interval

We assume normality and construct the CI using the bootstrap standard error:

$$\hat{se}_b = \sqrt{\frac{\sum_{i=0}^{B}(\hat{\theta}^{*i} - \overline{\hat{\theta}^{*}})}{B-1}}. \tag{2}$$

This symmetrical CI does poorly on skewed data.

3.2 The percentile interval

We construct the percentile interval by using percentiles of the bootstrap distribution. If we want to calculate a $1-2\alpha$ percentile interval, we order our bootstrap estimates $\theta^{*1}, ..., \theta^{*B}$ and take the $100 \cdot \alpha$-th and $100 \cdot (1-\alpha)$-th values as limits of our interval:

$$[\hat{\theta}_{lo}, \hat{\theta}_{up}] = [\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}]. \tag{3}$$

The percentile interval is first order accurate and transformation invariant.

3.3 The BC_a interval

BC_a stands for bias-corrected and accelerated. This method also uses percentiles of the bootstrap distribution to calculate interval endpoints, but the percentiles used are based on the shape of bootstrap distribution. Two values determine this interval: the bias-correction factor $\hat{z}_0$ and the acceleration factor $\hat{a}$:

$$[\hat{\theta}_{lo}, \hat{\theta}_{up}] = [\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}], \tag{4}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right) \tag{5}$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right). \tag{6}$$

The bias correction factor measures the median bias of $\hat{\theta}^*$ and is calculated by plugging the proportion of bootstrap replications less than our observed statistic $\hat{\theta}$ into the inverse function of a standard normal cumulative distribution function

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=0}^{B} I(\hat{\theta}^{*(b)} < \hat{\theta})\right). \tag{7}$$

The acceleration factor measures the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ:

$$\hat{a} = \frac{\sum_{i=0}^{n}\left((\sum_{i=0}^{n}\hat{\theta}_{(i)}/n) - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=0}^{n}\left((\sum_{i=0}^{n}\hat{\theta}_{(i)}/n) - \hat{\theta}_{(i)}\right)^2\right]^{3/2}}, \tag{8}$$

where $\hat{\theta}_{(i)}$ is a jackknife estimate of the statistic[2]. Note that if both factors equal zero, we get the percentile interval.

The method seeks to correct for bias and skew. It is second-order accurate, transformation invariant and range preserving. In the non-parametric setting it usually proves to be the most accurate [5].

4 Simulations

4.1 Number of bootstrap samples

Efron and Tibshirani [5] suggest that 25 bootstrap samples should be enough to get a good approximation of the standard error and 1000 samples for CI. Hesterberg [6], however, argues we should do at least 10^4 bootstrap samples to minimize sampling variability. Figure 1 illustrates how the number of bootstrap samples affect the bootstrap distribution.

[1] A CI is considered first-order accurate if the errors of the non-coverage probabilities differ from the true values by $O(n^{-\frac{1}{2}})$. It is second-order accurate if error size is of order $O(n^{-1})$.

[2] A jackknife estimate $\hat{\theta}_{(i)}$ is calculated from the original sample **x** with the i-th observation removed.

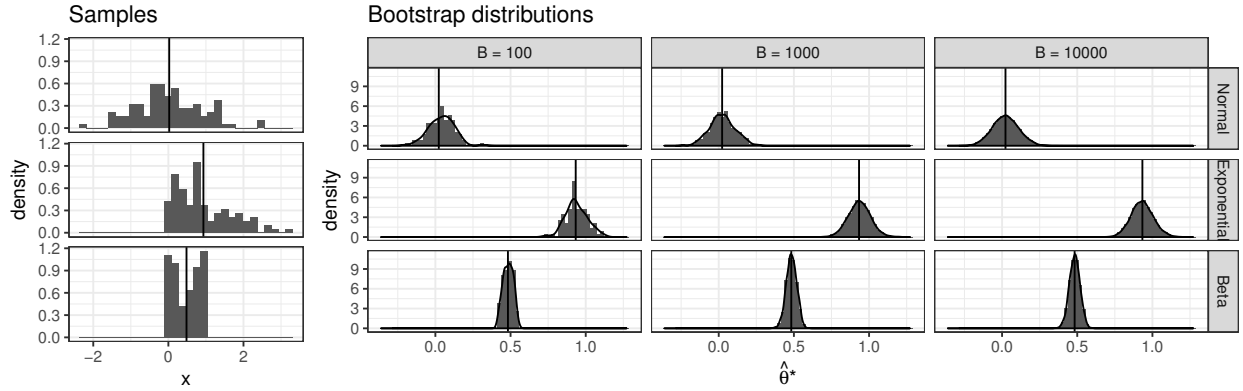


Figure 1: Changes in bootstrap distributions as the number of bootstrap samples B increases. Samples of size $n = 100$ are drawn from N(0,1) [top], Exp(1) [middle], and Beta(0.5, 0.5) [bottom]. Statistic of interest is the mean. Vertical lines are sample means.

4.2 Comparing confidence intervals

We applied 3 different variants of estimating the 95% CI to 3 distributions N(0,1), Exp(1), and Beta(0.5, 0.5) and estimated their coverage probability at different sample sizes n. The number of bootstrap samples was $B = 10^4$. The perfect CI would have coverage probability 0.95. We run the experiments with three statistics: mean, median and the 95-th percentile. To minimize variability our results are based on 10^4 repetitions for each n.

Simulation results in Figure 2 are in line with expectations. For $n \leq 25$ the bootstrap CI underestimate the CI. However, with increasing n, all three intervals converge to the specified coverage probability at their own rate. Overall, bootstrapped CI are biased (too narrow), which confirms what we previously stated and it needs to be taken into account when dealing with small n.

BC_a performs the best, but percentile intervals are almost as good and the advantage of BC_a decreases with increasing n. The normal interval performs best when the bootstrap distribution is roughly normal, so it does not perform well for the median and 95-th percentile. For the Beta distribution it even overestimates the CI, because in that case the booststrap distribution is bounded and concentrated at the upper bound.

4.3 Examples that arise in practice

To illustrate how the bootstrap can be expected to perform in practice, we prepared several examples of distributions that are the result of measuring the predictive performance of a machine learning model on a dataset:

- The distribution of the squared error obtained with leave-one-out cross-validation (LOOCV) for the *random forests* (RF) model on the *ozone* regression dataset [2]. From this distribution we bootstrap the mean and the rooted mean, effectively obtaining CI for the mean squared error (MSE) and root mean squared error (RMSE), respectively.
- The distribution of the difference in squared error obtained for the *lasso* and RF model in the same setting as above. We bootstrap the mean, effectively obtaining CI for how much the *lasso* is better/worse (in terms of mean squared error).

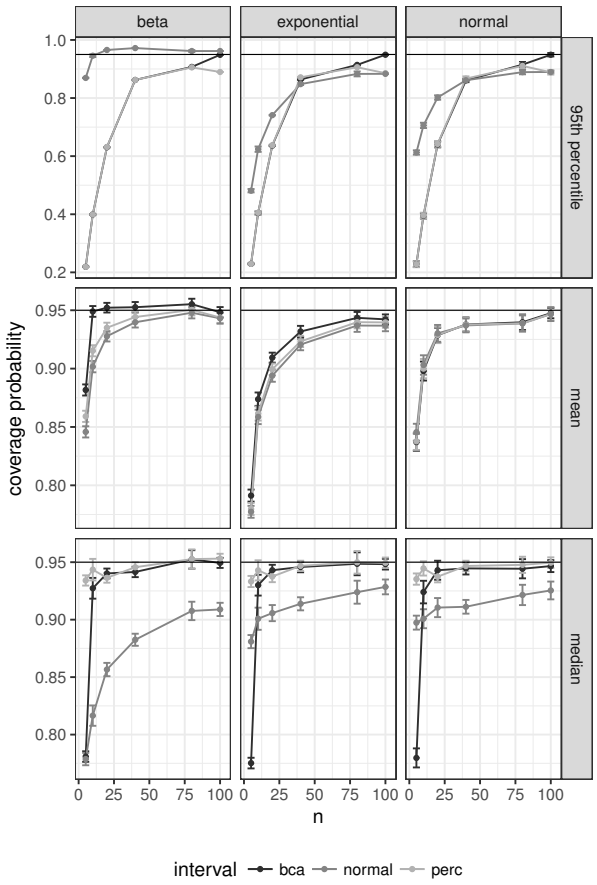


Figure 2: Observing how coverage probability changes when sample size n increases. The error bars denote 95% confidence in the interval coverages.

- The distribution of the 0-1 loss and the log-loss obtained with LOOCV for the RF model on the *Glass* dataset from the UCI Machine learning repository. We bootstrap the mean, effectively obtaining CI for the accuracy (CA) and log-score (LS).

The empirical distributions are shown in Figure 3. We pretend that these empirical distributions are the population distributions and obtain samples by resampling with

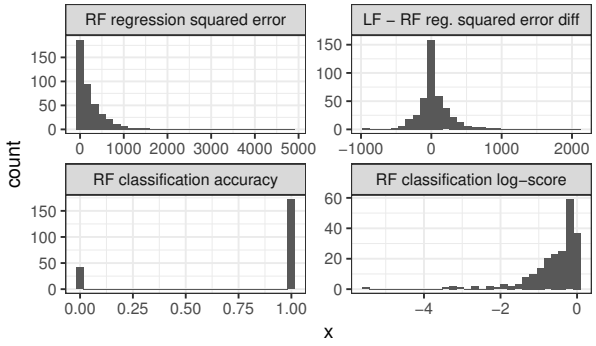


Figure 3: In practice, normally distributed performance data are not very common. 0-1 loss is distributed Bernoulli. Bounded variables, such as squared error and log-loss will often be close to exponential (or gamma) distributed. The difference of two such distributions will be close to a Laplace distribution.

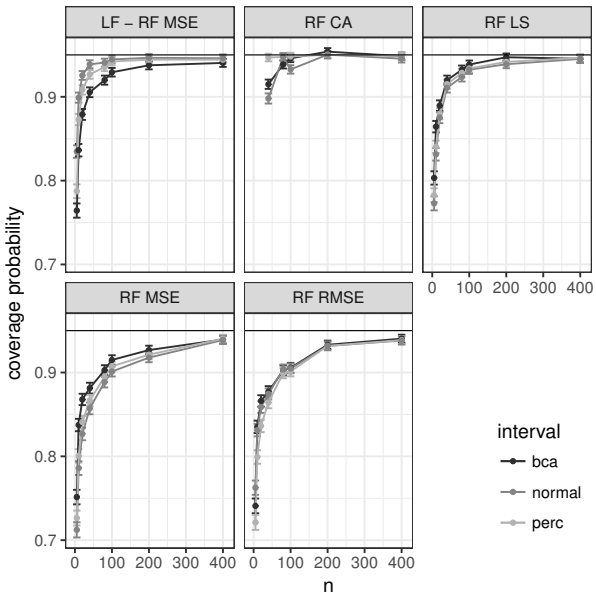


Figure 4: Coverage probabilities of CI when bootstrapping the empirical distributions.

replacement. This enables us to 'know' the true mean of the population and estimate coverage probabilities.

Again, we drew 10^4 bootstrap samples and calculated the CI based on 10^4 repetitions. Figure 4 shows that a larger sample size is needed for accurate CI for heavy-tailed distributions, such as the RF MSE. As expected, the normal interval's performance is relatively poor when the data are discrete and far from normal. The binary sample can also pose a problem when the sample size is too small (in our case $n < 40$), as sampling with replacement can result in a trivial distribution with zero variability. BCa performs the worst for a Laplace-like distribution (LF - RF). Similar results were observed in [8, 9].

5 Conclusion

The bootstrap has several advantages over classical approaches: it is easy to understand, implement, and apply and it can be applied to almost any statistic of interest.

It can fail in practice with very small sample size n and/or problematic statistics, such as extreme quantiles. However, simulations show that this is not a serious problem as long as we have at least $n \geq 50$.

BCa CI perform the best, but percentile intervals also perform well-enough. We should avoid normal intervals unless the bootstrap distribution is expected to be close to normally-distributed. Today, computation is relatively cheap, and the number of bootstrap samples B should be as large as possible, but simulations show that even 1000 is good enough for almost all practical situations.

Recently, there has been a shift towards Bayesian methods for model comparison [1]. Bayesian methods are in many ways more intuitive and therefore easier to interpret than frequentist ones. We did not touch on Bayesian approaches as the problems of frequentist approaches also apply to Bayesian approaches (distributional assumptions, problems when we are interested in more than just the mean). If one prefers the Bayesian view of probability, one can always apply the Bayesian boostrap [7].

References

[1] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

[2] Rok Češnovar and Erik Štrumbelj. Bayesian lasso and multinomial logistic regression on gpu. *PLoS ONE*, 12(6):e0180343, 2017.

[3] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[4] Brad Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.

[5] Brad Efron and Rob Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, Inc., 1993.

[6] Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015.

[7] Donald B. Rubin. The Bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

[8] Klairung Samart, Naratip Jansakul, and Mitchai Chongcheawchamnan. Exact bootstrap confidence intervals for regression coefficients in small samples. *Communications in Statistics - Simulation and Computation*, 2017.

[9] Yaqian Zhu and John Kolassa. Assessing and comparing the accuracy of various bootstrap methods. *Communications in Statistics - Simulation and Computation*, 2017.

Drone control using gestures

Jaka Cikač[1], Friedrich Fraundorfer[2], and Danijel Skočaj[1]

[1]*University of Ljubljana, Faculty of Computer and Information Science*
[2]*Graz University of Technology, Institute of Computer Graphics and Vision*
E-mail: jaka.cikac@gmail.com

Abstract

Quadcopters are becoming more popular and integrated into modern society. Currently we are familiar with controlling quadcopters via our mobile phones. In this work we develop a gesture control system that can be used on a low-cost quadcopter equipped with a simple RGB camera and a powerful embedded computer. The system is split into three modules - action detection with optical flow, human pose estimation with CNNs and gesture classification with relational features computed on the human pose. We assembled our own dataset called DS2017, in which 640 gestures are performed by 20 people. We show that action detection can detect actions sufficiently well, the human pose estimation works very well at high speed and gesture classification achieves high accuracy.

1 Introduction

As quadcopters get smaller and more affordable there are many new possibilities for their use. It is not uncommon to find drones at many university laboratories, research institutes and now even normal households. This opens up the possibility of making drones a part of people's lives to perform simple tasks or simply be very interesting toys. There are countless enthusiasts using drones equipped with the latest video capturing technology to provide stunning aerial views. However, there is always a needed component - *control*. Without the ability to control drones in a safe and efficient manner they are not really useful. There are now various ways of controlling the drones from manual flight (remote controller) to mobile application control. Manufacturers are even using VR to provide first person views of drone's flight using special goggles (FPV) and on-board cameras.

This work will focus on controlling the drones by using gestures. Hands free control would enable the operators to focus more on what they are doing, be it an activity that requires their physical engagement or rather just enjoying the scenery.

1.1 Problem definition

We set out to lay the ground work on how drones could be controlled using full-body gestures without any external devices, such as motion detectors and RGB-D cameras. Instead, the system is computer vision based. In order to be able to "see" the gestures, a drone should be equipped with an RGB monocular camera, from which a video feed of the user can be processed using computer vision, to estimate the human pose and further recognize and classify the gestures that are performed.

1.2 Related work

The field of gesture recognition has, in recent years, been explored to many depths, which can be seen from the many surveys that have appeared, such as [1].

One example of gesture controlling drones is [2]. In their work authors researched, which gestures seem natural for users to interact with the drone and what kind of modality (gestures or voice) the users used. They found that users used the same gestures as people use for conveying information to their pets or even interpersonal gestures such as come here, stop, come closer, move left, etc.

The main limitations of gesture control on drones come from computational power required on the on-board processor, it's camera system and it's payload capacity. For that reason most gesture control systems that were developed, require external devices and an off-line computer that could run the recognition algorithms. One of such examples is [3] where authors used a Parrot AR Drone and Microsoft Kinect to evaluate different metaphors for conveying controls for a variety of flying operations supported by the UAV.

Researchers in [4] created an implementation of gesture control using an inexpensive drone, the Parrot AR, and it's on-board RGB camera. The system was based on face tracking and hand gestures. Since the Parrot AR is lacking an on-board computer, everything needed to be processed on an off-line PC. The user would also have to wear colored gloves, so that the drone would be able to detect and track them. It also limited drone's movements in a very restricted space near the user.

One example of gesture control where, in theory, the UAV is not limited to the area near the user is [5], where authors wanted to track and recognize gestures for signaling aircraft. They created a dataset of NATOPS aircraft handling signals and a unified framework for body and hand tracking. Our work was partially inspired by this framework, but there are major differences. Authors of [5] use a 3D camera that is stationed on the ground. However we aim to implement a similar framework directly on the UAV with only a monochrome camera. In

this paper we present the main functionalities of the developed system and summarize the experimental results; further details are given in [6].

2 System design

The gesture control system is composed of three main components. The first important component of the system is *action detection*. We use action detection in order to improve computational efficiency of the whole system. The second component is *pose estimation* in order to obtain important features for *gesture classification* which infers the performed gesture. This final component provides the command to the drone. We need to detect when the action or gesture is taking place in a video sequence so that we can focus the processing power on a particular segment on which we use complex pose estimation algorithms that provide us with important features. The result of pose estimation is a set of joint positions as coordinates in an image. We use a state of the art method for pose estimation [7]. This method is based on deep convolutional neural networks (DCNNs). This choice obliges us to use a specific powerful GPU, that is not available on the drone. Despite this the system is still ready for real-time and on-line deployment. Gesture classification has been implemented on top of pose features in a bag of words approach. We learn gestures as words and then try to match a newly detected action to these words using an SVM classifier. In the following subsections we describe in more detail each component of the gesture control system and its underlying methods. A system overview diagram is shown on Figure 1.

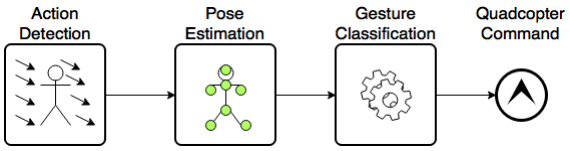


Figure 1: Three modules of the gesture control system: action detection, pose estimation, and gesture classification.

3 Methodology

3.1 Action Detection with person tracking

Action detection is split into three sub modules. We first detect the person with a *person detector* based on Histograms of oriented gradients [8], which provides initialization coordinates for a short-term tracker Adaptive Scale Mean Shift (ASMS) [9] that keeps track of a person through subsequent frames. We chose this particular tracker because it is extremely fast and performs well due to its awareness of background appearance and scale adaptation. Since tracking is a hard task, we make it easier by re-detecting a person every number of frames. After we are sure that we have the person localized we employ a dense grid around the person in which we calculate the optical flow using the Lucas-Kanade method. We filter the optical flow using RANSAC [10], so that only the person generated optical flow points remain. After filtering we count the optical flow points in 5 consecutive frames that form a chunk and classify it with a linear SVM, which decides if there was enough points, or not, to consider the chunk as an action. Chunks are buffered into circular buffers, which provide continuous functionality and additional filtering of false positives and false negatives. A flow chart of action detection with its submodules is shown on Figure 2. Frames on which action was detected are forwarded to a real-time pose estimation module.

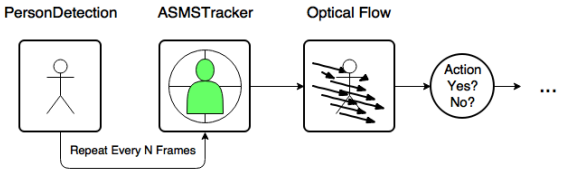


Figure 2: Overview of the action detection sub modules: person detection, visual tracking, optical flow estimation, determining action presence.

3.2 Human Pose Estimation

We chose a method called Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields [7] (MPE-PAF) as the main method for real-time person estimation. It was the first method to deliver real-time performance for pose estimation and it works for multiple people at the same time. Like CPMs [11] it is a method based on deep convolutional neural networks and inherits many ideas of CPMs, which were proven to have a very accurate pose estimation performance. MPE-PAF is a bottom-up approach that does not require a person detector. It has an architecture for jointly learning parts detection and parts association, which are then parsed with a greedy parsing algorithm to produce human pose estimation. This approach is not as accurate as CPMs but it does not perform much worse. Its main advantage is of course speed. Benchmarks were ran on a laptop version of NVIDIA GeForce 1080-GTX GPU, with which the authors obtain real-time performance of 8.8 FPS for a video with 19 people, making it an ideal choice for a real-time gesture control system.

3.3 Gesture Classification

After the pose is estimated, we are given a set of body joints for each frame in the detected action, which allows for certain features to be computed. The descriptor, which is a combined set of various features, computed on the body joint positions, is referred to as *Pose Features*. This descriptor combines positional features and relational features that describe geometry between combinations of joints. It was shown that a combination of these two types of features produce the best results for gesture classification [12]. All joint positions are normalized to $[0, 1]$. After normalization of joint positions, pose features are computed. In this work we compute features on a set, which is annotated with 15 joints per frame. We

calculate position features, distance features, orientation features, angle features and temporal features.

Position features are simply the normalized joint coordinates. *Distance features* are computed between combinations of pairs of joints. For joint pairs the distance between the joints is computed. *Orientation features* between joint pairs are computed as the orientation of the vector connecting two joints with respect to the neck-to-belly orientation. Then *angle features* are computed, represented by the inner angle of two vectors that span triplets of joints. Finally *temporal features* are computed. Temporal features consist of differences of features between two or more adjacent frames. In total this gives us 3180 features per frame, which make up the Pose Features frame descriptor.

After computing Pose Features we perform K-means clustering to quantize the descriptors into clusters. After the centers of clusters are returned we calculate the distance of each frame descriptor to each cluster and find the closest cluster center. After this is done for each frame, we end up with a histogram representing the whole video (action). The histogram is then normalized. Histograms are collected into a database and used as features for training the SVM. We use a radial basis function for the kernel.

For feature classification we therefore use a Bag of Words method trained on such descriptors. In the inference stage a similar procedure is used. After the pose is estimated on the frames that were deemed to contain an action, Pose Features are computed on each of them and the closest cluster center is found for the assembled descriptor. Then a histogram is assembled, normalized and then sent to the SVM for classification, which produces the result - a given action by the user.

4 Experimental results

4.1 DS2017 dataset

In our experiments we used 4 main gestures, namely *up*, *down*, *left* and *right*. For these 4 gestures we have created a *controlled gestures* dataset and an *intuitive* dataset. Each of these two categories is further split in a set of videos with a *steady camera* and the second with a *moving camera*, simulating the movements of the drone. Therefore we have 4 sub-sets in total, *controlled-steady set*, *controlled-unsteady set*, *intuitive-steady set* and *intuitive-unsteady set*.

The DS2017 consists of videos of 20 people, performing 640 gestures in 80 videos. Every video is annotated for action detection with 5 frames granularity and split into gestures then categorized in one of the 4 main categories. We limited the scenes to relatively homogeneous backgrounds, with equal illumination throughout the video. Although there are some examples where this does not hold. There is always a single person in the video, which is performing a gesture and in general the person does not move around on the scene. Two examples of video sequences showing the gesture *up* are presented in Fig. 3.

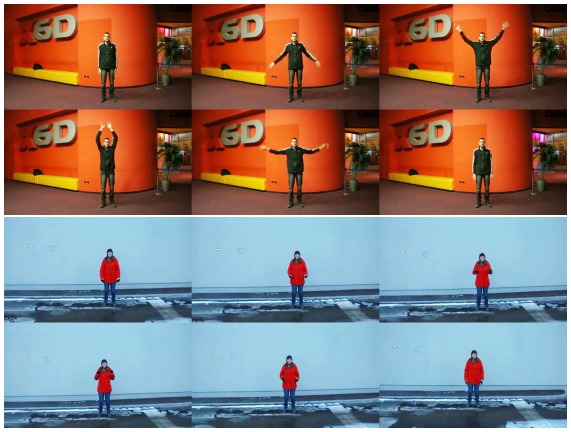


Figure 3: DS2017 dataset; two examples of gesture *up*: controlled (top) and intuitive (bottom).

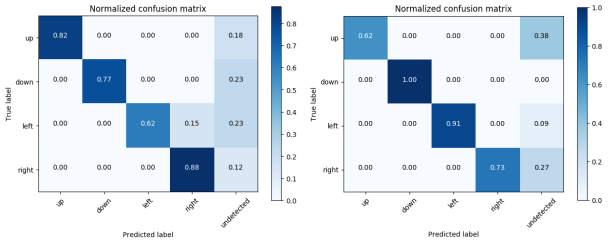
The purpose of a *controlled gesture* set is to have a very distinct set of gestures, that are very exuberantly performed in a standardized manner, potentially resulting in more optical flow points and have similar Pose Features descriptors, even when different people perform the gesture. It is meant to serve as a standardized controlled experiment, on which the algorithms would be evaluated. A more challenging part of the dataset are *intuitive gestures*, which are more natural to how a human would gesture a drone, intuitive and more fluent. These actions were not thoroughly regulated or explained to the people who would perform them in order to encourage variability of gestures from person to person.

We created two types of such datasets considering different acquisition conditions. *Steady* sub-sets were taken on a stabilized camera, so that there is no movement of the background. *Unsteady* videos were taken without any stabilization and random movements of the camera were introduced. These movements are meant to simulate a flying drone, with issues with stabilization or influence from the windy environment.

4.2 Performance evaluation

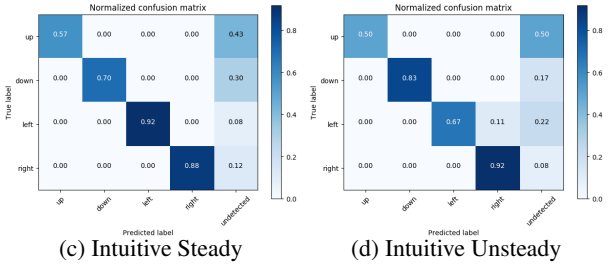
We evaluated the proposed system on our dataset DS2017. Here we report the classification accuracy after the video is sent through the whole pipeline. We evaluate the gesture sets separated on *steady* and *unsteady* sets. The action detection algorithm detected the action in **83,13%** of cases, or on 133 out of 160 test gesture videos in total. In the pose estimation model most issues occur with gestures *intuitive down* and *controlled down* due to the crisscrossing of the person's hands in front of their body. The pose estimation also struggles when there are hard shadows and under exposed frames. The movement of the drone does not impact the quality of the pose estimation algorithm.

Gesture classification accuracy is considered with undetected gestures taken into account, to show the complete system performance. In other words, if the action detection module failed to detect an on-going action, it is treated as miss-classification. For each category the classification accuracy is presented in a confusion matrix,



(a) Control Steady

(b) Control Unsteady



(c) Intuitive Steady

(d) Intuitive Unsteady

Figure 4: Confusion matrices for the evaluation of the gesture control system on DS2017 categories, which include the undetected gesture rates. If the gesture was not detected with the action detection module, it is treated as miss-classified and presented as category "undetected".

which also includes undetected gestures, on Figure 4. If action detection successfully detects the action however, the classification accuracy is 96.8%.

The highest overall accuracy is achieved on the *Controlled unsteady* gesture set, with an overall accuracy of 81%, as shown on Figure 4b. The lowest classification accuracy is achieved on *Intuitive unsteady* gesture set, as shown on Figure 4d, with an overall accuracy of 73%. The results confirm that the action detection module is struggling with detecting gesture *up*, which decreases the classification accuracy in all categories, except for category *Control steady*, where the gesture *right* has the lowest accuracy. The overall accuracy achieved for *Intuitive steady* gesture set is 76%. Overall accuracy for *Controlled steady* gesture set is 77%.

4.3 Runtime evaluation

From a detected action to the final gesture prediction it takes **2,14 s**, measured on a PC with a Core i7 4770K (3.5 GHz). The whole pipeline would run slower on a quadcopter, due to a less powerful CPU. However, most time is spent for pose estimation (*2,03 s* for a gesture on 25 frames), run on an external PC, so the total time would not be much longer. The entire pipeline would work in real-time on an on-board computer, if equipped with a sufficiently powerful GPU, such as a future version of NVIDIA Jetson TX 2.

5 Conclusion

We developed a three-phase gesture control system, by first detecting when the action is happening on the video, with a fast person detector, a fast tracker and optical flow. After detecting the action, we use a state of the art method to estimate the human pose on the frames that contain an action. The method uses DCNNs and is able to achieve real-time performance. After getting the locations of joints in the human pose, we compute relational Pose Features, that provide features for SVM classification, which predicts the final gesture.

We present the dataset DS2017 that features 640 gestures performed by 20 people. The system is able to detect actions in 83% of cases, having more success with the controlled gestures, while intuitive gestures are a harder challenge. We also found that the method [7] is extremely good in estimating the human pose on our dataset, struggling only when the hands criss-cross, as seen in gestures "down". We are also able to reach a high classification accuracy of 96.8% on our DS2017 dataset for the final gesture prediction.

References

[1] H. Cheng, L. Yang, Z. Liu, Survey on 3D hand gesture recognition, IEEE Transactions on Circuits and Systems for Video Technology 26 (9) (2016) 1659–1673.

[2] J. R. Cauchard, K. Y. Zhai, J. A. Landay, et al., Drone & me: an exploration into natural human-drone interaction, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 361–365.

[3] K. Pfeil, S. L. Koh, J. LaViola, Exploring 3D gesture metaphors for interaction with unmanned aerial vehicles, in: Proceedings of the 2013 international conference on Intelligent user interfaces, ACM, 2013, pp. 257–266.

[4] J. Nagi, A. Giusti, G. A. Di Caro, L. M. Gambardella, Hri in the sky: controlling uavs using face poses and hand gestures.

[5] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 500–506.

[6] J. Cikač, Drone control using gestures, Master's thesis, University of Ljubljana, Faculty of Computer and Information Science, Slovenia (2017).

[7] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: CVPR, 2017.

[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.

[9] T. Vojir, J. Noskova, J. Matas, Robust scale-adaptive mean-shift for tracking, Pattern Recognition Letters 49 (2014) 250–258.

[10] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[11] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[12] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black, Towards understanding action recognition, in: International Conf. on Computer Vision (ICCV), 2013, pp. 3192–3199.

Drone segmentation and tracking in grounded sensor scanned LiDAR datasets

Kristijan Mirčeta[1], Ciril Bohak[1], Byeong Hak Kim[2], Min Young Kim[2], Matija Marolt[1]

[1]*University of Ljubljana, Faculty of Computer and Information Science*
[2]*Kyungpook National University (Daegu, South Korea)*
E-mail: kristijan.mirceta@gmail.com, ciril.bohak@fri.uni-lj.si, durumy98@hanmail.net, minykim@knu.ac.kr, matija.marolt@fri.uni-lj.si

Segmentacija in sledenje dronom v podatkih prizemljenega senzorja LiDAR

V tem članku se spopadamo s problemom sledenja dronom v zunanjih scenah, ki jih skenira prizemljen senzor LiDAR. Problem razdelimo v 3 podprobleme: (1) segmentacija scene, (2) detekcija dronov, (3) sledenje dronom. Za segmentacijo scene uporabimo metodo radialno omejenih najbližjih sosedov. Za detekcijo dronov uporabljamo konvolucijske nevronske mreže. Za učenje mreže in za generiranje učne množice uporabljamo orodje Blensor, s katerim sintetizeramo sceno, ki želimo, da dobro simulira realnost. Na koncu, po detekciji dronov, le-tem tudi sledimo skozi več časovno zaporednih okvirjev s pomočjo Kalmanovega filtra, ki uporablja križno-korelacijski filter kot opazovalni model, skoraj konstantno hitrost kot model gibanja, ter vokseliziran deskriptor kot vizualni model.

Abstract

In this paper, we tackle the problem of tracking drones in outdoor scenes, scanned by a grounded LiDAR sensor. The problem is split into 3 subproblems: (1) segmentation of the scene, (2) drone detection and (3) drone tracking. For segmenting the scene, we use a Radially Bounded Nearest Neighbors method. For drone detection we use a Convolutional neural network (CNN). To train the CNN and produce a dataset of drones we use the Blensor framework, with which we synthesize drone examples, with the goal of them resembling realistic conditions. Finally after detecting the drones, we track them throughout multiple time frames using a Kalman filter, which uses a cross-correlation filter as the observation model, Nearly constant velocity as the motion model and a voxelized 3D descriptor for the visual model.

1 Introduction

In the cities of the future, Unmanned aerial vehicles (UAV) will have an important role in many functions of society, such as transport, delivery of goods, and surveillance systems [6]. But just as this technology could be very useful for society, it may also be used with malicious intent, for activities such as unauthorized access to off limits areas, spying on unsuspecting victims or disruption of other good-intented UAVs. Because of this, building a system for combating against malicious drones is an interesting and worthwhile problem. We want to detect and track drones to be able to avoid them (as other drones), destroy them (when they are in off limits areas) or inspect their activity to reveal patterns signifying malicious intent. In this paper, we are working with UAVs in LiDAR scenes, which are sparse 3D point cloud scenes. The sparsity of the LiDAR data is a big limitation, since it holds that the further away from the sensor an object is, the more sparsely it is sampled. This means that the extracted features will be less informative for detectors or trackers. In the following sections, we first present the data we are working on in Section 2, then we decompose the problem into 3 stages: segmentation in Section 3, detection in Section 4 and tracking in Section 5, and elaborate on each of them. In the final Section 6 we present conclusions and give pointers for future work.

2 Data

The driving force of today's AI algorithms is data. This puts large firms like Google, Amazon and Microsoft in a strongly advantageous position compared with their competition in the form of start-ups which may have great ideas, but no data to bring them to fruition. The recently emerging trend of synthetic data generation may be a way to democratize AI. The idea is that, instead of gathering massive amounts of data from the real world and labeling it by hand, we should create artificial datasets with which we may train our AI. Some researchers have already shown how synthetic data leverages the performance of their models when applied to real world cases. An example is [8], where the authors generated objects in the Unity game engine[1], and Blender[2] where they labeled key points on the objects in the virtual world and learned to predict the orientation of the object. This let them predict orientations of objects in the real world when filmed with a camera as well. There are also other packages like [17], which implements a wide range of sensors to scan these virtual worlds, and learn to predict various targets from this generated data in the hope of accurately simulating such data sensors in the real world. Because the generated data should be as close to real world data as possible, the Unreal Engine 4[3] is a great option due to

[1] https://unity3d.com/
[2] https://www.blender.org/
[3] https://www.unrealengine.com/

its Physically based rendering (PBR) capabilities which are unparalleled at the moment in similar software. Some packages like [18] work as a plugin to Unreal Engine 4, where both data is synthesized and simulation tests for self-driving cars can be made, and there are also packages higher in abstraction which provide capability for general computer vision tasks. An example is UnrealCV [15].

In this fashion, we are using Blender with the Blensor framework [5], to simulate virtual worlds and scan them using a LiDAR sensor. For our scenes, we are scanning with a simulated Velodyne HDL-64e grounded sensor.

We use a scene generator, which generates a scene with a small number of trees, primitive shapes (cubes, cylinders and prisms) and drones. These shapes are uniformly distributed around a 50×50 meter field, and the drones can fly on heights of up to 12.5 meters. Meanwhile we enforce the constraint that no objects can overlap with each other. We do not include the ground, as removing it is a very well researched problem and has been shown empirically to improve segmentation results by [2].

We generated 1000 such scenes, and scanned them using Blensor, where the scanner was located in the coordinate origin $(0, 0, 0)$, which resulted in our final input dataset - a point cloud with (x, y, z) values.

When testing out our methods, the scenes would be split as per standard practice in machine learning, where there would be a training set and a test set, roughly in the proportion of $4 : 1$.

Because the scenes are generated, we basically had the labels and bounding boxes for each item in the scene, and were therefore fully able to measure the accuracy of our methods using standard procedures. An example picture of a scanned scene can be seen in Fig. 1.

3 Segmentation

For the segmentation of the scene, we used a simple method called Radially Bounded Nearest Neighbours (RBNN) [7]. This method basically performs a clustering of the points, by declaring all points that are transitively less than some defined threshold (radius r) distant from one another to belong to the same object. Because there is no overlapping of objects in our scenes, and the objects are not very close to one another, this method was able to provide a sufficient segmentation for our needs. Most of the time the result of applying this method to our scene, was that each object was its own cluster. This provided a good basis for the detection of drones.

4 Detection

A bigger problem to tackle was the detection of the drones within the cluster. While we now know which points represent which objects, we need to find out which of the objects are in fact drones, and then track them.

The first thing to do was to find a descriptor for each cluster of points, then decide on whether it is a drone on the basis of the descriptor. Researchers have tried various different descriptors for various point cloud classification

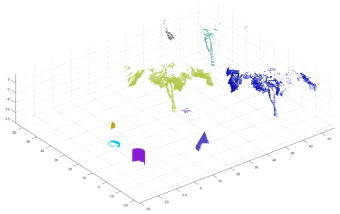


Figure 1: Plot of a segmented synthetic LiDAR scene. The gray object in the middle is the drone.

tasks, most prevalent among which fit into the following categories:

- **Multi-view descriptors**, where 2D projections (or picture) are taken from various angles of the point cloud . An example is in [19], where they classify office furniture via feeding multiview pictures into a CNN.

- **Voxelized descriptors**, where the point cloud is transformed into a regularized voxel grid. This can be done so, that if a point falls into a bin within the voxel grid, that voxel is set to 1. Then in some cases a Gaussian filter can be run through this grid, to give the voxels more volume. Such an approach is useful with methods that use the notion of similarity between descriptors to aid in their prediction. An example is in [12], which uses such a descriptor for registration. Even a descriptor without Gaussian filtering is useful, as it can be fed into a CNN, and the CNN will find its own filters which may prove to be even more useful. Examples of these are [16] and [11], where the latter was considered state-of-the-art in 2015. A plot of an example voxelized descriptor of a drone is show in Fig. 2.

- **Raw point clouds**, where the point cloud is not preprocessed, but used directly. Some deep neural nets have been researched, where this method produces very good results such as [14, 13]. A thing to watch out for here is permutational invariance, where it is ensured that the result of the prediction is independent of the ordering of the points in the point cloud.

- **Contextual**, where not only the point cloud segment itself is used, but its relations with the other parts of the scene are also accounted for. Such methods can be used as a leverage to other ones, and are currently producing state-of-the-art results as can be seen in [9].

Many previously listed methods employed CNNs to solve the problem of detection. We used a similar architecture as was used by VoxNet [11], where only the last layer was changed into a dense layer with a softmax activation, meant to classify only into two classes - *drone* and *not a drone*. The architecture defines a convolutional layer with 32 5×5×5 filters with $stride = 2$ for the first

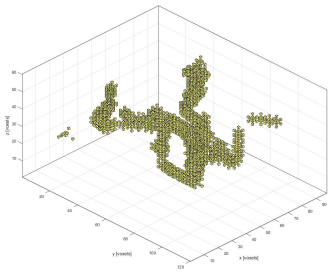


Figure 2: A visualization of the voxelized descriptor of a drone. The drone is dense because a Gaussian filter has been applied on it to give it more volume.

layer, another convolutional with 32 3×3×3 filters in the second layer, then a 2×2×2 max pooling layer, a 128 neuron fully connected layer which is used for the final prediction. A picture of the architecture can be seen in Fig.3.

The input is a voxelized descriptor, convolved with a Gaussian kernel with σ proportional to the distance from the sensor origin, because there is no way to inform the CNN about the variance in sampling density due to varying distance from the sensor origin. With applying the gaussian filter in such a way, we have encoded this information into the descriptor.

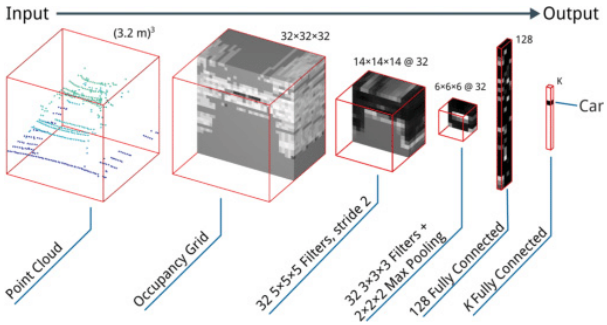


Figure 3: The original VoxNet CNN architecture. The point cloud is first mapped on to an occupancy grid (voxelized descriptor), flows through two convolutional layers, and finally through two fully connected layers to arrive at the prediction.

4.1 Detection Results

To train the CNN we used 2700 samples, of which there were roughly 300 drones, and the other objects were tree, cube, torus, cylinder, pyramid. The test set was 1000 samples, with roughly the same proportion (100 drones).

Through 10 random shufflings of the training and test set, the average AUC was above 0.99 when using a ROC curve, indicating that the CNN is a reliable classifier, while still being very fast. Though this result is very good, the reason is probably that the scenes we are predicting on are very simple. Nonetheless, it provides a good indicator that our direction of research is promising, and our next step would be to try this on more complex scenes.

5 Tracking

For the tracking, we were inspired by [12] where to track an object in a sequence of frames, an object was first segmented, then as stated before, a voxelized descriptor was used to match the objects in two subsequent frames using Bhattacharya similarity. If the similarity passed a certain threshold, then it was declared that the set of points in frame $t + 1$ was the same object as in frame t. The general idea is the same as the correlation filter [10].

In addition to this, we recognized a problem for the tracker, if the tracked object either came too close to another object or momentarily overlapped with it. For example if the drone passed through the edge of a tree canopy. In this case the tracker gets confused and loses track of the drone due to the similarity quickly dropping below a defined threshold.

To solve this, we used a Kalman filter [1] where the used observation model was the correlation filter, the visual model was the voxel grid, and the motion model was Nearly constant velocity (NCV). This provided our method with robustness to the stated problem. It is worthy to note here, that we assume the movement of the drone to be nearly linear, which is not a realistic assumption. Though research on how to relax this constraint has already been made with the extended Kalman filter [3] and unscented Kalman filter [20], which can track nonlinear movement as well.

5.1 Tracking Results

We tried tracking a drone following a linear path, but would pass through the edge of a tree canopy on its way. We tried applying only a correlation filter at first, where the tracker would get stuck in the tree canopy, unable to track the drone onwards. But as shown on Fig. 4, the Kalman filter was successful in overcoming its obstacle and tracking the drone all the way.

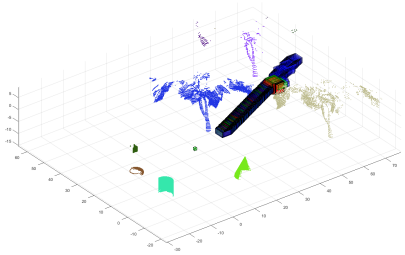


Figure 4: The picture shows multiple overlapped frames in a LiDAR scene, where the linear path of the drone is shown. The blue sequence of bounding boxes are the intermediate steps of the Kalman filter tracker. It can be seen that the Kalman filter was able to follow the linear path of the drone right through the tree canopy.

6 Conclusion

We have developed a pipeline of subproblems which defines a method for tracking drones in a LiDAR dataset. There are many opportunities to refine our method in more

domains. We could do the segmentation and detection steps in one single step instead of two, just as the current state-of-the-art method SPGraph [9]. We could also refine our neural network architecture to use the new information we are accounting for with relationships among objects and object parts. We believe that much more work should be done regarding the data generation, where we could use better simulations to extract RGB values from scenes as well, and perhaps use some modern techniques like Generative Adversarial Networks [4] to generate scenes that are more believable and realistic than our randomly crafted ones. By optimizing for realism in virtual scenes, we are also optimizing for performance in real-world scenes.

Lastly, regarding tracking, we could upgrade the method by using an unscented Kalman filter for tracking, to model nonlinear motion, which is more usable in reality. As tracking is also a bottleneck to real-time performance due to slow computation of cross-correlation, we could replace both the visual and observational models with faster ones, perhaps in RGB data, we would use color histograms instead of regularized voxel grids for similarity computation.

References

[1] Gary Bishop and Greg Welch. An Introduction to the Kalman Filter: SIGGRAPH 2001 Course 8. In *Computer Graphics, Annual Conference on Computer Graphics & Interactive Techniques*, pages 12–17, 2001.

[2] B. Douillard, J. Underwood, N. Kuntz, V. Vlaskine, A. Quadros, P. Morton, and A. Frenkel. On the Segmentation of 3D LIDAR Point Clouds. In *2011 IEEE International Conference on Robotics and Automation*, pages 2798–2805, May 2011.

[3] Keisuke Fujii. Extended Kalman Filter. *Refernce Manual*, 2013.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Michael Gschwandtner, Roland Kwitt, Andreas Uhl, and Wolfgang Pree. BlenSor: Blender Sensor Simulation Toolbox. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Song Wang, Kim Kyungnam, Bedrich Benes, Kenneth Moreland, Christoph Borst, Stephen DiVerdi, Chiang Yi-Jen, and Jiang Ming, editors, *Advances in Visual Computing*, pages 199–208, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[6] Ismail Guvenc, Ozgur Ozdemir, Yavuz Yapici, Hani Mehrpouyan, and David W. Matolak. Detection, Localization, and Tracking of Anauthorized UAS and Jammers. In *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, pages 1–10, 2017.

[7] Klaas Klasing, Dirk Wollherr, and Martin Buss. A Clustering Method for Efficient Segmentation of 3D Laser Data. *2008 IEEE International Conference on Robotics and Automation*, pages 4043–4048, 2008.

[8] Marcel Klomann, Michael Englert, Kai Weber, Paul Grimm, and Yvonne Jung. Improving Mobile MR Applications Using a Cloud-based Image Segmentation Approach with Synthetic Training Data. In *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, pages 4:1–4:7, New York, NY, USA, 2018. ACM.

[9] Loïc Landrieu and Martin Simonovsky. Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs. *CoRR*, abs/1711.09869, 2017.

[10] Yang Li and Jianke Zhu. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, pages 254–265, Cham, 2015. Springer International Publishing.

[11] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015.

[12] Daniel D. Morris, Brian R. Colonna, and Paul H. Haley. LADAR-Based Mover Detection from Moving Vehicles. *CoRR*, abs/1709.08515, 2017.

[13] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv preprint arXiv:1706.02413*, 2017.

[14] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[15] Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *Proceedings of European Conference on Computer Vision*, pages 909–916. Springer, 2016.

[16] Xavier Roynard, Jean-Emmanuel Deschaud, and François Goulette. Classification of Point Cloud Scenes with Multiscale Voxel Deep Network. *CoRR*, abs/1804.03583, 2018.

[17] Raffaele Schiavullo. SYNCITY virtual hyper realistic simulator for machine learning to train ADAS systems. *Virtual Reality*, 4:06, 2018.

[18] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and service robotics*, pages 621–635. Springer, 2018.

[19] Vladeta Stojanovic, Matthias Trapp, Rico Richter, and Jürgen Döllner. A Service-oriented Approach for Classifying 3D Points Clouds by Example of Office Furniture Classification. In *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, pages 2:1–2:9, New York, NY, USA, 2018. ACM.

[20] Eric A Wan and Rudolph Van Der Merwe. The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 153–158. IEEE, 2000.

Detecting hands with a convolutional neural network trained using synthetic data

Barbara Aljaž, Luka Čehovin Zajc

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-pošta: ba4918@student.uni-lj.si, luka.cehovin@fri.uni-lj.si

Abstract

Deep learning, more precisely convolutional neural networks, are known for the large amount of data they require for training. One of the new approaches of providing a sufficient amount of training samples is their artificial generation using computer graphics. In this paper we present a use-case of training a convolutional neural network detector using generated images of hands. We have evaluated the detector on a dataset of real images for a touch-less interface human-computer interaction scenario and compared it with a detector trained on real images. Results of the experiment are promising and present many opportunities for further development of such training technique.

1 Introduction

Deep learning, more specifically, the use of multi-layer convolutional neural networks (CNN) has made a significant impact in the computer vision field in the last decade, performance on tasks like detection, recognition, classification, and tracking has been greatly improved. This was primarily made possible by new methodological discoveries, but also by the rapid increase in computing power and storage capacities, as well as availability of data.

But despite its abundance, collecting suitable data for various domains is challenging and expensive. Training and testing images or videos have to be collected, filtered, and annotated. Collected images may also include bias and may not represent the appearance variations uniformly. Since deep learning methods are notably data-hungry with training sets well above 1000 samples, a need for easier and more flexible ways of data collection has emerged. Beyond data augmentation that increases dataset size using simple image processing operations (affine transformations, blur, noise), artificially generated samples can be used as well in some cases. Modern computer graphics approaches can produce realistic images of objects. Moreover, the properties of an object and its surroundings can be controlled during data generation process, reducing bias and giving full control over annotation process. More importantly, the process is automatic, meaning that we can generate large quantity of training samples and quickly adapt to new learning scenarios.

In this paper we investigate the role of synthetically generated data in training a CNN-based detector for human hands. This kind of detector can be used as a component in various human-computer (HCI) and human-robot (HRI) interaction scenarios. Hands are also very suitable test domain for such approach because of high DoF and reasonably low variability between different hands. We have developed a system that generates a large amount of synthetic training images based on a parameterized 3D model of a hand and combine them with an arbitrary background, as shown in Figure 2. These samples were then used to train a CNN detector that we have evaluated on a dataset of real images of hand poses.

2 Related work

Hand localization and gesture recognition has a long history of research. Traditionally, color cues were used [1] since such approach is computationally effective, but the performance of such method suffers in case of color similarity or difficult lighting conditions. With the advent of low-cost depth cameras (e.g. Kinect), hand detection has become easier in constrained usage scenarios and some gestures could be detected with simple image processing operations [2]. More recently, machine learning has also been used, e.g. AdaBoost [3], Random forest [4], and even more recently CNNs [5, 6]. All these approaches relied on datasets of images of real hands which are challenging to build in an unbiased manner and difficult to extend to different training scenarios.

On the other side, synthetic imagery has been successfully utilized before in various, predominantly HCI, scenarios. In one of the initial attempts generated depth images were used to train a random-forest-based body-part detector [7], and more recently to determine hand posture parameters using CNNs [8]. These methods achieve good performance also due to lower amount of appearance variation of depth images in comparison to color images.

Computer-generated color images have been utilized in deep learning to do 3D face reconstruction [9] and to detect various objects in an autonomous driving scenario [10]. Recently several attempts have used generated hand images for hand pose recognition. In [11] authors used synthetic images to classify pre-localized hand images into several gestures classes. In [12], a full hand

pose is recovered from a pre-localized hand. In [13], the approach is extended into temporal domain. In contrast, our presented work is focused on solving hand localization using synthetic images which could be seen as complementary step to posture recognition.

3 The hand detector

Our process of creating a real-time hand detector consists of generation of artificial training samples, described in Section 3.1, training a suitable CNN network, described in Section 3.2.

3.1 Dataset generation

For synthetic data generation we have used an open-source 3D model [1], seen in Figure 1. The model consists of 2590 faces, 59 bones, and has skin-like texture enhanced using sub-surface scattering technique. We have used a Python script within the Blender 3D modelling application [2] to generate images with random skeleton configurations, hand global rotation and color hue as well as position and intensity of environmental lights.

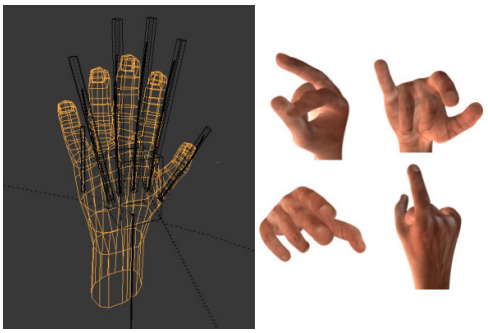


Figure 1: Wire-frame representation of a hand model (left) and some generated poses rendered on transparent background (right).

For the current experiments we have generated 100 different hand poses with 20 instances of each, some examples are shown in Figure 1. All images were rendered with transparent background. In the next step we have combined hand images with random real images that were verified that they did not contain visible hands. Background images were randomly cropped and the overlaying hand images were augmented with combination of flipping, rotating, blurring as well as shifting hue and saturation components. Some examples of generated learning images are presented in Figure 2.

3.2 Detector

For the detector have used a modified YOLO network architecture [14] that is known for its real-time performance. The architecture is different from commonly used R-CNN because it uses a single network for extracting potential foreground candidates and their categorization. The network is consequently trained on an entire image

[1] free3d.com/3d-model/rigged-hands-28855.html
[2] www.blender.org



Figure 2: Examples of artificially generated images used in learning process with bounding boxes that mark groundtruth.

and can therefore get more information about the context around the object. The network architecture is inspired by the GoogLeNet model [15], it has 24 convolutional layers followed by two fully connected layers. The learning process and inference consist of dividing an image into a grid where every cell can predict a certain number of bounding boxes, their scores and class probabilities. The loss function is composed of 5 parts - error of predicted bounding boxes' coordinates, error of their dimensions, error of object presence, error of object absence and classification error. This approach makes inference very fast, but it is also prone to having lower detection accuracy and recall because of a limited number of potential bounding boxes.

We have started from the original YOLO architecture used for general object detection. We have lowered the number of filters in first fully connected layer from 4096 to 1024 since our problem is easier and we can sacrifice some complexity in order to achieve higher inference speed, the final architecture is presented in Figure 3. We have also added batch normalization layers used during training. We have trained the network from scratch and, based on our preliminary experiments, used Adam optimizer with more than two hundred times smaller learning rate in comparison to the original network, to avoid divergence during training. The training was performed using batch size of 5 and trained for a maximum of 100 epochs.

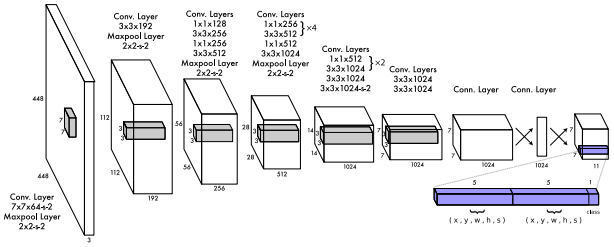


Figure 3: Network architecture of our detector.

Since detector can provide multiple detections for a single object, we perform a non-maxima suppression step

to reduce the number of overlapping proposals. We have clustered all detection with mutual overlap of 0.1 together and retained only the one with the highest detection score.

4 Experimental evaluation

In our evaluation we have focused on using a detector to detect hands in camera-facing HCI scenario. The parameter distribution of hands was therefore adapted for such view-point. We have evaluated our detector on a hand gesture dataset [16, 17], that consists of RGB, depth images, and information from a Leap Motion sensor (in contrast to other work that uses the dataset we have used only RGB images because of the nature of our study). The dataset contains 1400 images of 14 different people showing 10 different gestures with 10 samples per gesture. Some examples of test images can be seen in Figure 4. Since bounding-box annotations were not provided, we annotated images using LabelImg [3] utility and cropped them from the left and right side in order to obtain a square image suitable for the YOLO network (which expects images of 448×448 pixels). In the evaluation we have followed a standard PASCAL evaluation procedure for object detection [18]. We present results for the evaluation in form of precision-recall curves and best F-scores in Figure 5.



Figure 4: Examples from the testing dataset.

We begun our evaluation with a model trained exclusively on synthetic images, denoted as *only_syn*. This model was trained on 5000 synthetic samples. This model achieved F-score of 0.758 when evaluated on the entire testing dataset. The performance of the model did not increased if it was trained on more images, which could signify that the variability of synthetic training data was not sufficient. Indeed, our current data generation approach does not take into account background-foreground contrast which can be observed as a problem in Figure 6. One of the main problems of the model was also poor accuracy of the YOLO architecture, which results in many detection windows with overlap a bit below the 0.5 PASCAL threshold. The shape of the precision-recall plot for the *only_syn* model shows that some wrong detections have very high scores. Upon a closer investigation we have discovered that most of these detections were indeed positioned on a hand, had a good detection score but with a very poor overlap with the groundtruth.

We have also evaluated the model trained on synthetic images on another synthetically generated dataset of 5000 images, we denote this result as *syn_on_syn*. The model achieved better performance with F-score of 0.88. This was expected, but the score is still low enough that

[3] github.com/tzutalin/labelImg

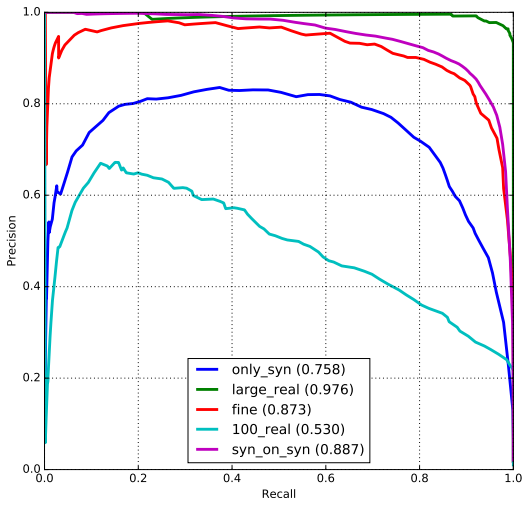


Figure 5: Precision-recall curve. The optimal F-score for each curve is provided in brackets next to the corresponding legend entry.

we cannot talk about over-fitting of the model to the synthetic data and shows potential for mining hard samples for more efficient training.

In order to see the limit of training the same network on realistic data, we have trained the detector from scratch on a random 80% of our realistic dataset (1120 images) and tested on the remaining 20% (280 images). The results are denoted as *large_real*, the best achieved F-score was very high, i.e. 0.97. This was expected since images were chosen randomly and the similarity of 10 instances of the same person showing the same gesture is very high, which means that the model is likely very overfitted. We would also like to emphasize that, in contrast to this detector, that was trained on a selected number of testing gestures, the detector trained on synthetic images is more general and that its performance would probably not degrade even if new gestures would be added to the testing dataset.

The last experiment that we have performed addresses fine-tuning of a model trained with synthetic data with some real samples. Similar approach has been investigated in [11] and was confirmed to improve performance of their model. We have used 100 images (two people showing five gestures) from the realistic dataset to fine-tune the model. The evaluation of the model, denoted as *fine*, was performed on a subset of the testing dataset that excluded both people and all five gestures to avoid overfitting. The best F-score rose to 0.873, a 10% improvement in comparison to detector learned only on synthetic data. A reasonable explanation for this increase could be the role of the context in convolutional neural networks. A strong cue for the presence of a hand could also be a wrist, part of an arm or even face. Since our synthetic images only contain hands the appropriate context is missing and was only introduced during fine-tuning on real

data. We have also trained the model denoted *100_real* only on 100 real images and achieved best F-score of 0.53 on the testing subset. This number is still quite high, which probably means that the testing dataset is not very diverse.

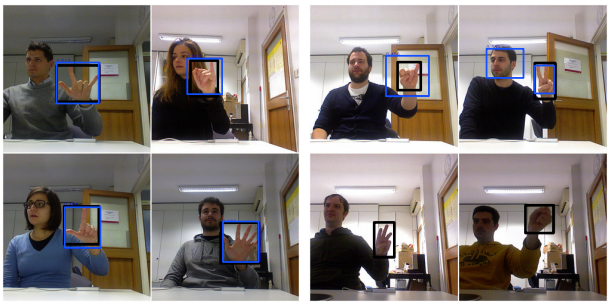


Figure 6: Examples of successful detection (left) and various failure cases (right) that occurred either because of a small overlap with groundtruth, detection of face due to skin color similarity, false negative because of image contrast and false negative due poor lightning conditions and motion blur. Black rectangles represent ground-truth, blue rectangles represent detections.

5 Conclusion

In this paper we have presented our study on suitability of using computer generated synthetic data to train a CNN detector. We have developed a synthetic data generation system for images of human hands, trained a CNN detector and evaluated it on realistic dataset. We have achieved encouraging performance that shows the potential of using synthetic images for deep learning in computer vision.

We have also discovered several potential directions for further work. While the current model of a hand is quite realistic, it does not acknowledge the role of context in CNNs, that could improve detection accuracy. We also intend to extend training dataset with more variability in terms of hand pose and other factors (e.g. hand hue, lights). In terms of accuracy we will evaluate other fast CNN architectures that improve localization. We also plan to evaluate our model on more challenging datasets and compare it with existing approaches. Finally, our system enables effortless extension into recognition of hand postures and well as hand gestures. Such a system could be a building block of many HCI or HRI applications.

References

[1] Xiaojin Zhu, Jie Yang, and Alex Waibel. Segmenting hands of arbitrary color. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 446–453. IEEE, 2000.

[2] Yi Li. Hand gesture recognition using kinect. In *2012 IEEE International Conference on Computer Science and Automation Engineering*, pages 196–199, June 2012.

[3] Mathias Kölsch and Matthew Turk. Robust hand detection. In *FGR*, pages 614–619, 2004.

[4] Xian Zhao, Zhan Song, Jian Guo, Yanguo Zhao, and Feng Zheng. Real-time hand gesture detection and recognition by random forest. In Maotai Zhao and Junpin Sha, editors, *Communications and Information Processing*, pages 747–755. Springer, 2012.

[5] Akshay Rangesh, Eshed Ohn-Bar, and Mohan M Trivedi. Hidden hands: Tracking hands with an occlusion aware tracker. In *CVPRW 2016*, pages 19–26, 2016.

[6] Shiyang Yan, Jeremy S Smith, Yizhang Xia, Wenjin Lu, and Bailing Zhang. Multi-scale convolutional neural networks for hand detection. In *Applied Computational Intelligence and Soft Computing*, volume 2017, 04 2017.

[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, CVPR '11, pages 1297–1304, Washington, DC, USA, 2011. IEEE Computer Society.

[8] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. *CoRR*, abs/1609.09698, 2016.

[9] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469, Oct 2016.

[10] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV 2016*, pages 102–118. Springer, 2016.

[11] C. J. Tsai, Y. W. Tsai, S. L. Hsu, and Y. C. Wu. Synthetic training of deep cnn for 3d hand gesture identification. In *2017 International Conference on Control, Artificial Intelligence, Robotics Optimization (ICCAIRO)*, pages 165–170, May 2017.

[12] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single RGB images. *CoRR*, abs/1705.01389, 2017.

[13] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. *CoRR*, abs/1704.02201, 2017.

[14] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[16] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. Hand gesture recognition with leap motion and kinect devices. *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1565–1569, 2014.

[17] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. Hand gesture recognition with jointly calibrated leap motion and depth sensor. *Multimedia Tools Appl.*, 75(22):14991–15015, November 2016.

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Skladiščni sistem za verifikacijo predmetov s pametnimi očali

Janez Križaj, Janez Perš in Vitomir Štruc

Fakulteta za elektrotehniko, Univerza v Ljubljani
E-pošta: janez.krizaj@fe.uni-lj.si

Warehouse system for object verification with smart glasses

This paper presents a warehouse product verification system that helps warehouse staff in performing their work. The proposed system sequentially captures the pictures of the objects that the warehouse worker encounters in his work and verifies if the object in the picture is the one that the worker is supposed to find in the warehouse. The system uses Android-powered smart glasses to capture image data, while the computationally intensive verification tasks are carried out on the server computer and are implemented using recent deep learning techniques. By doing so, the system is able to process images in near real-time and achieves high verification accuracy as is demonstrated by the performed experiments.

1 Uvod

V članku predstavimo podporni sistem, ki skladiščnim delavcem omogoča zanesljivejše in po možnosti hitrejše izvajanje svojih nalog. Pri obstoječem načinu dela v skladiščih delavec v informacijskem sistemu najprej poišče lokacijo določenega produkta v skladišču, na dotični lokaciji pa nato z ustreznim čitalnikom preveri ali gre res za iskani produkt. Razviti podporni sistem pa s pomočjo pametnih očal zajema sliko objektov, s katerimi se srečuje skladiščni delavec pri svojem delu, in preveri ali gre za pravi objekt, t.j. objekt, ki naj bi ga delavec v skladišču poiskal. Z načrtovano tehnologijo je torej v postopek vpeljan avtomatiziran proces verifikacije produktov, ki je izveden z računalniško-podprto tehnologijo razpoznavanja objektov iz slikovnih podatkov, zajetih s pametnimi očali.

Glavni izziv in razvojni problem, s katerim smo se srečali pri udejanjenju podpornega sistema je problem verifikacije objektov na podlagi slikovnih podatkov, zajetih s pametnimi očali. Procesno verigo verifikacijskega postopka smo sestavili iz klasičnih gradnikov, kot je detektor objektov, modul za izračun značilk in modul za razvrščanje, pri čemer smo za detekcijo objektov in izračun značilk uporabili sodobne postopke, ki temeljijo na uporabi globokih nevronskih mrež.

Zaradi relativno visoke računske zahtevnosti teh postopkov smo večji del procesne verige prenesli na strežnik z grafično kartico, ki podpira hitro izvajanje in učenje globokih mrež, pametna očala Vuzix, na katerih teče operacijski sistem Android, pa nam služijo le za zajem in prenos slike na strežnik ter prikaz rezultata verifikacije. Ob predpostavki dovolj hitre internetne povezave smo na ta način zagotovili sprejem povratne informacije, v obliki rezultata verifikacije in koordinat detekcijskega okna, v okvirno 0,2 s po tem, ko smo poslali sliko v obdelavo na strežnik.

2 Implementacija podpornega sistema

To poglavje vsebuje opis gradnikov in delovanje podpornega sistema za skladiščne delavce.

2.1 Način delovanja

Demonstracijski sistem za predmet, ki ga demonstrator prime v roke izvede verifikacijo. Sistem prične s poskusi razpoznavanja predmetov na zajetih slikah, ko skladiščnik z izbiro želenega objekta na skladiščnem računalniku sproži postopek verifikacije. Od začetka razpoznavanja se z določenim intervalom vzorčenja zajemajo slike in pošljejo v obdelavo na strežnik. V primeru, da je istovetnost predmeta potrjena, se sistem ne odzove, v primeru, da sistemu istovetnosti predmeta ne uspe potrditi, pa sistem sproži “alarm” in uporabnika opozori, da trenutno aktualen predmet, ne ustreza predmetu, ki naj bi ga skladiščnik poiskal. Ob vsakem “alarmu” demonstracijski sistem shrani tudi sliko zajetega predmeta za namene arhiviranja.

2.2 Zgradba sistema

Kot je prikazano na sliki 1 je podporni sistem v grobem sestavljen iz treh delov:

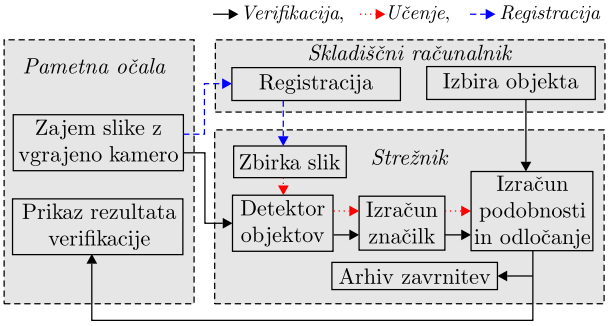
- *Pametna očala*: Na pametnih očalih teče spletna aplikacija, ki omogoča zajem slike z vgrajeno kamero očal, prikaz slike na zaslonu očal, pošiljanje slike v obdelavo na strežnik v formatu JPEG, sprejem rezultata verifikacije iz strežnika in prikaz rezultata vključno z detekcijskim oknom na zaslonu očal;

- *Strežnik*: Ob sprejemu slike, strežnik pošlje dano sliko skozi procesno verigo, zgrajeno iz klasičnih gradnikov razpoznavalnega sistema, kot so modul za predobdelavo slik, modul za izračun značilk, modul za merjenje podobnosti in odločanje. Strežnik nato izračunan rezultat verifikacije in koordinate detekcijskega okna za vsako dobljeno sliko pošlje nazaj do pametnih očal in

obenem arhivira slike pri katerih je prišlo do zavrnitve pri verifikaciji;

- *Skladiščni računalnik* na katerem se izvaja registracija novih objektov v sistem in izbira objekta, ki ga želimo verificirati.

Sledeča podpoglavja vsebujejo podrobnejši opis naštetih sklopov razvitega sistema.



Slika 1: Shematski prikaz podpornega sistema.

2.3 Registracija novih objektov v sistem

Pri registraciji objekta v sistem, se predpostavlja, da imamo na voljo 6 slik vsake od stranic tega objekta. Ob prenosu vseh 6 slik stranic objekta na strežnik, se na strežniškem delu sistema izvede še učenje detektorja in razvrščevalnika na novo dograjeni učni zbirki. Za učinkovitejše učenje smo dodali postopek samodejnega tvorjenja umetnih slik, ki temelji na izgradnji 3D modela objekta iz njegovih slik stranic in nam omogoča enostavno povečanje učnega materiala. Postopek izgradnje 3D modela iz slik stranic zahteva, da pri postopku registracije v grafičnem vmesniku ročno označimo vse štiri kote stranice na posamezni sliki, preostali koraki tvorjenja slik in učenja pa so samodejni. Na osnovi lokacije označenih kotov se s perspektivično transformacijo izvede poravnava stranice v frontalno lego. Model danega objekta nato sestavimo iz frontaliziranih slik vseh 6 stranic objeta. Umetne slike, ki jih uporabimo za učenje detektorja in razvrščevalnika, končno tvorimo z rotacijo modela v 3D prostoru in dodajanjem različnih ozadij, kot prikazuje slika 2.

2.4 Detekcija objektov

Za detekcijo objektov smo uporabili postopek Faster-RCNN [1], implementiran v knjižnici CNTK [2]. Omenjeni postopek uporablja za detekcijo objektov konvolucijske nevronske mreže (implementacija trenutno podpira mreži VGG16 in AlexNet) in za učenje potrebuje veliko učnega materiala. Ker take količine učnih slik nimamo na razpolago smo vzeli globoko mrežo, ki je bila vnaprej naučena na zbirki slik ImageNet in nato le doučili parametre mreže na lastni zbirki slik. Postopek se je izkazal za hitrega in učinkovitega, z deležem detektiranih objektov preko 96% in povprečnim IOU (ang. intersection over union) enakim 0,78. Detekcijsko okno nam kasneje služi za kvadratni izrez področja, ki ga zahteva modul za izračun značilk.



(a) Slike stranic objekta z označenimi robnimi točkami.



(b) Frontalizirane slike stranic objekta.



(c) Umetno tvorjene slike

Slika 2: Postopek registracije in tvorjenja umetnih slik.

2.5 Izračun značilk

Ker postopki globokega učenja v zadnjih letih dosegajo visoke rezultate na področju računalniškega vida, smo se odločili, da izračun značilk izvedemo s pomočjo globokih mrež. V eksperimentalni primerjavi (poglavje 3.2) se je izkazalo, da največjo učinkovitost verifikacije zagotavljajo značilke, ki jih dobimo z mrežo Resnet kot izhode iz predzadnje plasti nevronov (izpustimo zadnjo polno povezano plast). Iz tega razloga smo se odločili, da za izračun značilk nadalje uporabljamo mrežo ResNet. Značilke izračunamo iz kvadratnega področja, ki zajema detekcijsko okno objekta, pri čemer smo preizkusili kakšen vpliv ima na končno uspešnost verifikacije ohranitev oz. odstranitev ozadja izven detekcijskega okna (poglavje 3.3).

2.6 Izračun podobnosti in odločanje

Dani objekt, ki ga posnamemo s kamero pametnih očal, primerjamo z izbranim registriranim objektom, ki smo ga predhodno izbrali v uporabniškem vmesniku. Na podlagi mere podobnosti med značilkami danega objekta in značilkami izbranega registriranega objekta se izvede sprejem oz. zavrnitev objekta, glede na to ali je izračunana podobnost večja ali manjša od v naprej nastavljenega praga. Določitev praga in preizkus različnih vrst mer podobnosti eksperimentalno ocenimo v poglavju 3.4.

2.7 Uporabniški vmesnik

Komunikacija uporabnika s sistemom je omogočena preko več uporabniških vmesnikov. Uporabniški vmesnik na pametnih očalih je izveden v obliki spletne aplikacije, ki jo zaženemo v internetnem brskalniku očal in omogoča tako zajem slike kot prikaz rezultata verifikacije in prikaz detekcijskega okna na dani sliki objekta. Na ta način je zajem neodvisen od operacijskega sistema in razpoložljivosti razvojnega kompleta (SDK). Na skladiščnem računalniku dostopamo do več vmesnikov, od katerih eden omogoča registracijo objektov v sistem, drugi vmesnik nam omogoča izbiro objekta za katerega

želimo izvesti verifikacijo, tretji vmesnik pa predstavlja spletni arhiv slik z zavrnjenimi objekti. Vmesniki so implementirani v jeziku HTML5, CSS in JavaScript.

2.8 Komunikacijski kanali

Demonstracijski sistem deluje na treh enotah, ki vključujejo strežniški računalnik, kjer se izvaja glavnina procesa verifikacije, pametna očala s pripadajočo spletno kamero ter skladiščni računalnik na katerem dostopamo do vmesnika za registracijo in izbiro objekta. Ko sistem deluje v načinu verifikacije je potrebno sprotno pošiljanje podatkov med pametnimi očali in skladiščnim računalnikom na eni strani ter strežnikom na drugi strani. V ta namen smo z uporabo spletnih vtičnic (ang. websockets) implementirali komunikacijska kanala, ki omogočata obojesmerni prenos podatkov. Komunikacijski kanal med vmesnikom pametnih očal in strežnikom skrbi za prenos slike v smeri do strežnika in za prenos rezultata verifikacije ter koordinat detekcijskega okna v obratni smeri. Na komunikacijskem kanalu med vmesnikom za izbiro objekta in strežnikom pa se v smeri proti strežniku vrši prenos oznake izbranega objekta, v obratni smeri pa teče prenos rezultata verifikacije.

2.9 Programska oprema

Jedro podpornega sistema, t.j. proces verifikacije, ki teče na strežniku, smo implementirali v programskem jeziku Python, pri čemer smo si pomagali z različnimi programskimi knjižnicami. S pomočjo knjižnice CNTK smo izvedli detekcijo objektov, knjižnici Tensorflow in Keras smo uporabili pri izračunu značilk, knjižnica Mayavi nam služi za 3D modeliranje objektov v procesu registracije, knjižnico Sklearn pa smo uporabili pri merjenju podobnosti med vektorji značilk. Pri udejanjenju komunikacijskih kanalov s strežnikom pa smo uporabili knjižnico Websockets. Spletni uporabniški vmesniki so implementirani s pomočjo programskih jezikov HTML5, CSS in JavaScript.

3 Eksperimenti

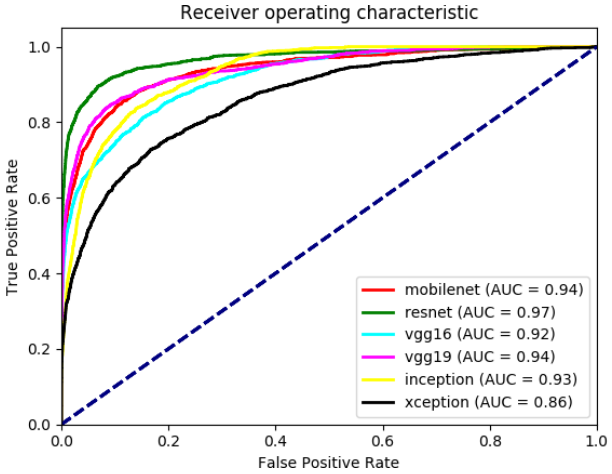
V tem poglavju predstavimo rezultate izvedenih eksperimentov. S pomočjo pridobljenih eksperimentalnih rezultatov smo nastavili oz. izbrali določene odprte komponente sistema, pri čemer smo težili k čim večji učinkovitost verifikacije, ki smo jo merili v obliki deleža TPR pri 0,1% FPR. V okviru testiranja smo preverili tudi učinkovitost delovanja posameznih postopkov, ki smo jih integrirali v sistem, kot tudi delovanje prototipnega sistema kot celote. Eksperimente smo izvedli na lastni zbirki slik kosmičev, nekatere preliminarne eksperimente pa smo naredili na zbirki kosmičev, ki je rezultat diplomskega dela iz [3]. Ker sistem deluje v načinu verifikacije, smo uspešnost delovanja prikazali v obliki krivulj ROC (ang. receiver operating characteristic).

3.1 Snemanje lastne zbirke slik

Za potrebe evalvacije zanesljivosti razvitega sistema smo zbrali lastno zbirko slik 8 različnih škatel kosmičev, s katerimi smo poskusili imitirati okolje v katerem deluje skladiščni delavec. Zbirka vsebuje testno množico in galerijsko množico. Testna množica vsebuje približno sto slik vsake od škatel, pri čemer smo ob njihovem zajemu težili k temu, da orientacija škatel na slikah čim bolj variira. Galerijsko množico poleg 6 slik stranic vsakega od 8 registriranih predmetov, sestavlja 1000 umetnih slik vsakega od objektov, ki smo jih tvorili po postopku opisanem v poglavju 2.3.

3.2 Izbira globoke mreže za izračunu značilk

V nizu eksperimentov, opisanem v tem razdelku, smo preizkusili uspešnost delovanja pri uporabi različnih prednaučenih globokih mrež, ki smo jih uporabili pri izračunu značilk. Globoke mreže potrebujejo velike količine podatkov za učinkovito učenje, ki je praviloma računsko zahtevno in posledično časovno potratno. Težavam pri učenju mrež smo se zato izognili z uporabo mrež, ki so bile predhodno naučene na velikem številu generičnih slik. Osredotočili smo se na primerjavo šestih pred-naučenih mrež iz knjižnice Keras, ki so bile naučene na zbirki ImageNet [4]. Te mreže so VGG16 in VGG19 [5], Inception [6], Xception [7], ResNet [8] in MobileNet [9]. Rezultati teh eksperimentov so prikazani na sliki 3 iz katere lahko razberemo, da najvišjo uspešnost verifikacije dosežemo z mrežo ResNet. V vseh nadaljnjih eksperimentih smo zato za izračun značilk uporabili globoko mrežo ResNet.



Slika 3: Uspešnost verifikacije ob uporabi različnih vrst prednaučenih globokih mrež.

3.3 Definicija območja pri izračun značilk

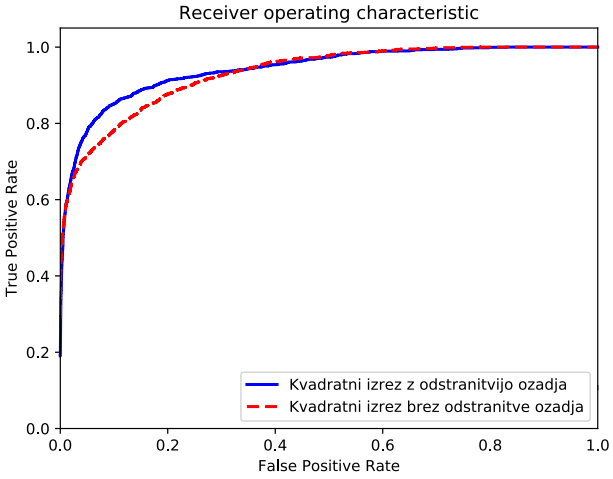
Globoka mreža ResNet, ki jo uporabljamo pri izračunu značilk, zahteva na svojem vhodu sliko kvadratne oblike, dimenzije 229×229. Detekcijski pravokotnik praviloma ni kvadratne oblike, ampak ima poljubno razmerje med višino in dolžino, ki je odvisno od oblike objekta na sliki. Obravnavali smo dve možnosti izvedbe kvadratnega izreza pravokotnega področja detekcije: *i*) kvadratni izrez, kjer ohranimo ozadje in ga po potrebi ekstrapoliramo, če je detekcijsko okno blizu roba slike (slika 4b), *ii*) kvadratni izrez, kjer vrednosti slikovnih elementov izven detekcijskega okna nastavimo na določeno konstantno vrednost (črno področje na sliki 4c). Izkazalo se je, da

slednji postopek prinese boljšo učinkovitost verifikacije (slika 5) v primerjavi s postopkom, kjer ohranimo ozadje.



(a) (b) (c)

Slika 4: Določitev kvadratnega območja pri izračunu značilk: a) Vhodna slika z detekcijskim oknom, b) kvadratni izrez brez odstranitve ozadja, c) kvadratni izrez z odstranitvijo ozadja.



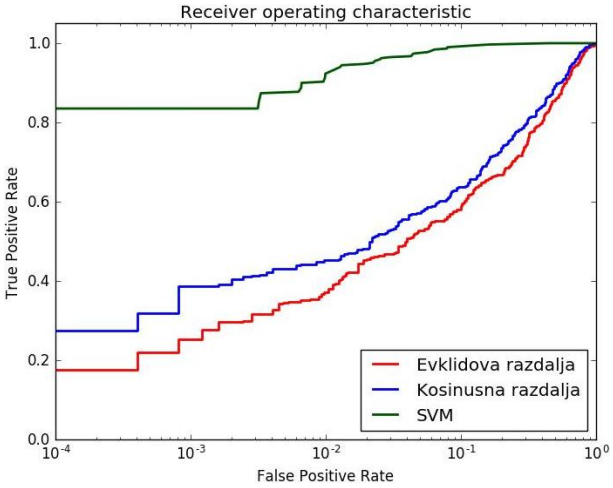
Slika 5: Uspešnost verifikacije glede na uporabljen postopek izreza.

3.4 Izbira mere podobnosti in razvrščevalnika

Pri izbiri mere podobnosti v modulu za odločanje smo, poleg iskanja najbližjega soseda iz vsakega razreda z uporabo Evklidove in kosinusne razdalje, preizkusili še merjenje podobnosti z metodo podpornih vektorjev [10] (ang. support vector machine, SVM), s katero smo, kot je razvidno iz slike 6, dosegli znatno izboljšanje uspešnosti verifikacije. Odločitveni prag smo določili glede na operacijsko točko, ki smo jo izbrali pri 0,1% napačnih sprejemov (ang. false positives). V tej točki je z uporabo metode podpornih vektorjev uspešnost verifikacije presegla 80% pravilnih sprejemov (ang. true positives).

4 Zaključek

V članku smo predstavili demonstracijski sistem za skladiščne delavce, ki smo ga razvili v okviru projekta GOSTOP. Podporni sistem omogoča pohitritev opravil skladiščnih delavcev in enostavnejše preverjanje istovetnosti produktov v skladiščih. Opisali smo osnovne značilnosti razvite storitve, orisali njeno arhitekturo in funkcionalnost ter eksperimentalno ovrednotili njeno delovanje. Pomembna lastnost omenjenega sistema je, da ga je možno uporabiti z vsako napravo, ki podpira novejše spletne brskalnike (v našem primeru Firefox).



Slika 6: Uspešnost verifikacije pri različnih postopkih merjenja podobnosti.

Zahvala

Raziskovalno delo, ki je pripeljalo do predstavljenih rezultatov, je bilo delno financirano s strani ARRS (raziskovalna programa Metrologija in biometrični sistemi, P2-0250, in Računalniški vid, P2-0214) ter programa GOSTOP (št. pogodbe C3330-16-529000), financiranega s strani RS-MIZŠ in EU-ESRR.

Literatura

[1] R. Girshick, "Fast R-CNN," v *Proc. of the 2015 IEEE Int. Conf. on Comp. Vis. (ICCV)*, ser. ICCV '15. Washington, DC, ZDA: IEEE Computer Society, 2015, str. 1440–1448.

[2] F. Seide in A. Agarwal, "CNTK: Microsoft's Open-Source Deep-Learning Toolkit," v *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, ZDA: ACM, 2016, str. 2135–2135.

[3] U. Brodnik, "Detekcija artiklov na prodajnih policah z računalniškim vidom," *Univerza v Ljubljani, Fakulteta za elektrotehniko*, zv. Diplomsko delo, 2012.

[4] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," v *2009 IEEE Conf. on Comp. Vis. and Pat. Rec.*, Jun. 2009, str. 248–255.

[5] K. Simonyan in A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, 2014.

[6] C. Szegedy *et al.*, "Going Deeper with Convolutions," v *2015 IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, Jun. 2015, str. 1–9.

[7] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," v *2017 IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, Jul. 2017, str. 1800–1807.

[8] K. He, X. Zhang, S. Ren, in J. Sun, "Deep Residual Learning for Image Recognition," v *2016 IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, June 2016, str. 770–778.

[9] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *CoRR*, 2017.

[10] C. Cortes in V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, zv. 20, št. 3, str. 273–297, Sept. 1995.

Primerjava delovanja kratkoročnih in dolgoročnih sledilnikov v vodnem okolju

Martin Pernuš[1], Borja Bovcon[2], Matej Kristan[2] in Janez Perš[1]

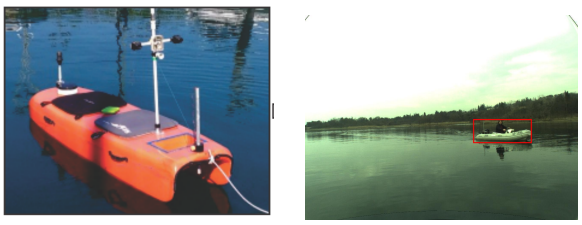
[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
[2]*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*
E-pošta: janez.pers@fe.uni-lj.si

Comparison of Short-Term and Long-Term Trackers in the Waterborne Environment

An analysis of short-term and long-term trackers performance on marine-based dataset of video sequences is presented. We used existing marine datasets to create two new specialized datasets, targeted towards short-term and long-term tracking analysis, respectively. We perform experiments on those datasets using several of the top-performing short-term and long-term trackers based on their Visual Object Tracking competitions performance. Trackers are evaluated based on their processing speed as well. The results are compared to results in similar studies.

1 Uvod

V zadnjem desetletju so raziskave v robotiki prinesle vrsto majhnih avtonomnih površinskih plovil (angl. *Unmanned Surface Vehicles, USV*). Primer USV kaže Slika 1. Zaradi njihovih zmožnosti navigiranja v marinah in plitkih vodah ta plovila predstavljajo velik potencial v avtomatizaciji del, kot je nadzor obal in pregled umetnih struktur pod in nad vodnim površjem. Vodimo jih lahko ročno ali pa imajo neko stopnjo avtonomije. Za dosego visoke stopnje avtonomije plovilo potrebuje sposobnost zaznavanja in sledenja objektov. Za ti nalogi lahko uporabimo kamere, kar omogoča uporabo metod računalniškega vida. Metode računalniškega vida, ki obravnavajo detekcijo in sledenje objektom so lahko specializirane za uporabo na vodni površini, ali pa splošne, namenjene uporabi v kakršnemkoli okolju.



Slika 1: Levo: Plovilo podjetja Harpha d.o.o., s katerega so bili zajeti posnetki. Desno: primer sličice v videoposnetku, ki vsebuje tarčo, ki ji mora sledilnik slediti in pravokotnikom, ki predstavlja resnično pozicijo tarče.

V našem delu smo se osredotočili na analizo sledenja že detektiranih objektov. Detektiran objekt, ki mu sledimo, v tem delu imenujemo tarča. Analizirali smo delovanje sledilnikov, ki so namenjeni splošni uporabi in ne specifično sledenju na morski površini. Sledilniki so bili izbrani glede na rezultate, ki so jih dosegli na priznanem izzivu Visual Object Tracking (VOT) [2]. Te smo testirali na dveh novih podatkovnih zbirkah, ki za osnovo uporabljata podatkovni zbirki Marine Obstacle Detection Dataset (MODD) 1 [14] in 2 [15], namenjeni pa sta testiranju kratkoročnih in dolgoročnih sledilnikov.

2 Pregled področja

2.1 Sledenje

Sledenje lahko v grobem razvrstimo v dva večja sklopa, na kratkoročno in dolgoročno sledenje. Kratkoročno sledenje se nanaša na sledenje tarčam v videoposnetkih, kjer je tarča vedno v zornem kotu kamere, dolgoročno sledenje pa dovoljuje izginotje in zakrivanje tarče. Večina raziskovalnega fokusa se je v preteklosti osredotočila na kratkoročno sledenje. Pri kratkoročnem sledenju se izkaže, da zaradi postopne degradacije vizualnega modela sledilnika (algoritma, ki izvaja sledenje), prej ali slej pride do odpovedi sledenja [1]. Zaradi tega se pojavlja težnja k razvoju dolgoročnih sledilnikov, ki so sposobni ponastavitve sledenja.

Rezultati izziva VOT 2017 [2] kažejo, da trenutno najboljši kratkoročni sledilniki temeljijo na metodi diskriminativnih korelacijskih filtrov, katerega odlikuje nizka računska zahtevnost, saj za reševanje optimizacijskega problema uporablja hitro Fourierovo transformacijo.

Dolgoročni sledilniki so običajno sestavljeni iz kombinacije kratkoročnega sledilnika in detektorja tarč [5]. Detektor tarč lahko v primeru neuspešnosti kratkoročnega sledilnika le-tega ponastavi.

2.2 Evalvacijski protokoli

Najbolj znana metodologija ocenjevanja kratkoročnih sledilnikov je metodologija izziva VOT [2]. Za testiranje uporablja vizualno raznoliko podatkovno zbirko, končni rezultati pa so ocenjeni na ločeni podatkovni zbirki, do katere tekmovalci nimajo dostopa. V nasprotju z drugimi metodologijami [3] ob odpovedi sledilnika le-tega ponastavi čez nekaj sličic.

V nasprotju z metodologijami kratkoročnega sledenja, metodologij in podatkovnih zbirk za dolgoročno sledenje še nedavno ni bilo. V [5] so v okviru metodologije VOT predstavili novo metodologijo ocenjevanja dolgoročnih sledilnikov in vizualno raznoliko podatkovno zbirko. Primarna mera kvalitete sledilnika je maksimalna F-ocena glede na prag razvrščanja tarče. Na videoposnetkih so testirali tako dolgoročne, kot tudi kratkoročne sledilnike. Metrike ocenjevanja sledilnikov so osnovane na metodologiji VOT in so podrobneje opisane v poglavju 3.

2.3 Detekcija objektov in sledenje v vodnem okolju

Za zaznavo in sledenje objektov na USV lahko uporabimo različne senzorje. Tipični senzorji so senzorji za merjenje razdalje, kot so radar, lidar in sonar. V zadnjem času se pojavlja vse več študij, v katerih objekte zaznavajo in jim sledijo s kamerami [4]. V primerjavi s senzorji razdalje kamere nimajo problemov z nizko kotno resolucijo in nizko hitrostjo zajema podatkov. Obenem imajo zmožnost detekcije objektov, ki so z drugimi senzorji težje zaznavne, kot na primer potapljač na vodni površini ali plavajoče razbitine. Slabosti uporabe kamer pa izvirajo predvsem iz težav modeliranja vodne dinamike, različnih zornih kotov pogleda na objekte, velike variabilnosti v osvetljenosti objektov in odbojnosti žarkov na vodni površini.

Najbolj sorodno delo našemu je [4], v katerem analizirajo delovanje različnih detektorjev in sledilnikov na njihovi podatkovni zbirki. Podatkovno zbirko sestavljajo videoposnetki, ki so zajeti iz plovila ali iz obale. Sledilnike so testirali le na videoposnetkih, ki so bili zajeti iz obale, medtem ko v našem delu analiziramo posnetke, ki so bili zajeti izključno iz plovila. Primer ene sličice iz nabora posnetkov prikazuje Slika 1.

3 Metode

V nadaljevanju so našteti uporabljeni kratkoročni in dolgoročni sledilniki, ki so v tem delu uporabljeni, nato pa opisane še metode ocenjevanja teh sledilnikov.

3.1 Kratkoročni sledilniki

Kratkoročni sledilniki so bili izbrani glede na rezultate pričakovanega povprečnega prekrivanja izziva VOT 2017 [2]. Najboljši peti sledilniki, ki smo jih testirali na naši podatkovni zbirki, so Continuous Convolution Operator Tracker (CCOT) [6], Efficient Convolution Operators for Tracking (ECO) [7], Convolutional Features for Correlation Filters (CFCF) [8], Correlation Filter with Weighted Convolution Responses (CFWCR) [9], in gNetTracker (gNet).

3.2 Dolgoročni sledilniki

Na izbor dolgoročnih sledilnikov je vplival članek [5], kjer so identificirali primerne sledilnike za dolgoročno sledenje. Testirani dolgoročni sledilniki so Tracking, Learning, Detection (TLD) [10], MUlti-Store TrackER (MUSTER) [11], Clustering of Static-Adaptive Correspondence for Deformable Object Tracking (CMT) [12] in Fully-Correlational Long-term Tracker (FCLT) [13].

3.3 Metode ocenjevanja

3.3.1 Ocenjevanje kratkoročnih sledilnikov

Poznamo več mer ocenjevanja kratkoročnih sledilnikov. Najbolj pogosto uporabljene so povprečna točnost, robustnost in pričakovana povprečna točnost sledilnika. Povprečna točnost meri povprečni Jaccardov indeks med napovedjo in resnično pozicijo tarče. Jaccardov indeks je ob času t definiran, kot

$$\Omega(A_t, G_t) = \frac{A_t \bigcap G_t}{A_t \bigcup G_t}, \tag{1}$$

kjer je A_t napovedana, G_t pa resnična pozicija tarče.

Robustnost meri število odpovedi sledilnika. Odpoved sledilnika se zgodi takrat, ko je Jaccardov indeks $\Omega(A_t, G_t) = 0$. Ti dve meri nam povesta veliko o kvaliteti sledilnika, potrebujemo pa še neko enoznačno mero, ki pove, kateri sledilnik je boljši od drugega. S pričakovano povprečno točnost (angl. *Expected Average Overlap, EAO*) ocenimo povprečno prekrivanje preko najbolj pogostih dolžin videoposnetkov v podatkovni zbirki, pri čemer upoštevamo tudi odpovedi sledilnika. Dodatno smo testirali še hitrost izvajanja sledilnikov.

3.3.2 Ocenjevanje dolgoročnih sledilnikov

Metoda ocenjevanja dolgoročnih sledilnikov je povzeta po članku [5]. V primerjavi s kratkoročnimi sledilniki morajo biti dolgoročni sledilniki sposobni tako sporočiti odsotnost tarče, kot tudi ponovno detekcijo tarče. Dolgoročni sledilnik mora tako poleg pozicije tarče sporočiti tudi lastno stopnjo zaupanja prisotnosti tarče. Napoved sledilnika se upošteva le takrat, ko je njegova stopnja zaupanja večja od praga za razvrščanje. Za ocenjevanje kvalitete lahko uporabimo mero maksimalne F-ocene glede na parameter praga za razvrščanje. Podrobnosti so opisane v [5]. Tudi pri dolgoročnih sledilnikih analiziramo njihovo hitrost procesiranja sličic.

4 Eksperimenti

4.1 Podatkovna zbirka

V naših eksperimentih uporabljamo dve podatkovni zbirki, ki sta služili kot podlaga za kreacijo podatkovnih zbirk, primernih za testiranje kratkoročnih in dolgoročnih sledilnikov. To sta podatkovni zbirki Marine Obstacle Detection Dataset 1, (MODD1) [14] in 2 (MODD2) [15]. Oba vsebujeta videoposnetke, posnete na USV. Posnetki so bili pridobljeni v Koprskem zalivu. Objekti na vodi so označeni s pravokotniki. Objekti obsegajo boje, zasidrano jadrnico, potapljača na morski gladini, premikajoč se čoln ipd.

Podatkovni zbirki MODD1 in MODD2 je bilo potrebno za ocenjevanje kvalitete sledilnikov spremeniti v ustrezno obliko. V okviru kratkoročnega sledenja lahko hkrati sledimo le eni tarči, videoposnetki pa včasih vsebujejo več tarč. Take videoposnetke smo obdelali tako, da smo iz njih naredili več videoposnetkov, primernih za testiranje kratkoročnih sledilnikov. Podobno smo naredili pri gradnji podatkovne zbirke za testiranje dolgoročnih sledilnikov. Pri izboru sekvenc za kratkoročne sledilnike smo

pazili, da tarče med videoposnetkom nikoli popolnoma ne zaidejo izven zornega kota kamere. Na drugi strani pa je bil prav to pogoj za izbor dolgoročnega videoposnetka. Videoposnetki niso bili v nobenem primeru manipulirani z z umetnim zakrivanjem ali dodajanjem tarč. Ustvarili smo 68 videoposnetkov, primernih za testiranje kratkoročnih in 16 videoposnetkov, primernih za testiranje dolgoročnih sledilnikov. Podatkovna zbirka videoposnetkov za dolgoročne sledilnike je obsegala videoposnetke, kjer se zgodi eno zakrivanje ali izginotje objekta, v enem posnetku pa se izginotje objekta zgodi dvakrat. Vsakič se tarča pojavi približno na istem mestu, kot izgine, vendar se ji ob tem ponekod spremeni velikost in osvetljenost. Primer sličic videoposnetka, primernega za testiranje dolgoročnih sledilnikov, prikazuje Slika 2.

4.2 Rezultati

Rezultati so bili pridobljeni s pomočjo orodja VOT[1], ki skrbi za komunikacijo s kodo sledilnikov, podpira pa tudi podrobno analizo rezultatov.

4.2.1 Rezultati kratkoročnih sledilnikov

Tabela 1 prikazuje rezultate kratkoročnih sledilnikov na podatkovni zbirki, primerni za kratkoročno sledenje, v obliki pričakovanega povprečnega prekrivanja (angl. *Expected Average Overlap, EAO*) in sličic na sekundo (angl. *Frames per Second, FPS*). Vsi algoritmi so bili testirani na isti računalniški opremi.

Tabela 1: Rezultati kratkoročnih sledilnikov

Sledilnik	EAO	FPS
CFWCR	0,65	1,71
ECO	0,61	3,47
CFCF	0,61	0,77
CCOT	0,60	0,41
gNet	0,59	2,01

Rezultati kratkoročnih sledilnikov se razlikujejo od rezultatov, pridobljenih na izzivu VOT 2017. Sledilnik CFWCR, ki je bil v naših testih najboljši, je na izzivu VOT 2017 izmed vseh petih naših testiranih sledilnikov dosegel najslabše, peto mesto, najboljši sledilnik glede na VOT 2017 CCOT pa je dosegel četrto mesto. S primerjavo vrednosti EAO z izzivom VOT 2017 lahko opazimo, da so naši videoposnetki za sledenje veliko lažji za preizkušene sledilnike. Najboljši sledilnik na izzivu VOT 2017, CCOT, je namreč dosegel le 0,203 vrednosti EAO. Pridobljeni rezultati kažejo na izredno dober uspeh sledilnikov na danih videoposnetkih v vodnem okolju.

S podrobno analizo kratkoročnih sledilnikov na posameznih posnetkih lahko vidimo, v kakšnih videoposnetkih imajo sledilniki največ težav. Večina videoposnetkov, kjer imajo sledilniki težave, je karakteriziranih z visoko hitrostjo rotacije plovila v kombinaciji z zmernim valovanjem, kar povzroči velike nenadne spremembe pozicije tarče. Sledilniki robustno sledijo ob spremembah velikosti in osvetljenosti tarče.

[1] https://github.com/votchallenge/vot-toolkit

4.2.2 Rezultati dolgoročnih sledilnikov

Tabela 2 prikazuje rezultate dolgoročnih sledilnikov na podatkovni zbirki, primerni za dolgoročno sledenje. Poleg opisanih dolgoročnih sledilnikov smo za primerjavo testirali še kratkoročni sledilnik ECO, ki smo mu prav tako pripisali lastno mero zaupanja. Primarna mera je tu maksimalna F-ocena, prav tako pa analiziramo tudi hitrost sledilnikov v enotah FPS.

Tabela 2: Rezultati dolgoročnih sledilnikov

Sledilnik	Maksimalna F-ocena	FPS
TLD	0,52	2,96
ECO	0,45	4,81
MUSTER	0,39	1,11
FCLT	0,39	2,40
CMT	0,31	71,43

Rezultati dolgoročnih sledilnikov kažejo na uspešnost sledilnika TLD v primerjavi z ostalimi sledilniki. Sledilnik ECO je presenetljivo drugi, čeprav je smatran kot kratkoročni sledilnik. Sledilnika MUSTER in FCLT imata enak rezultat, najslabši pa je sledilnik CMT.

Dobljeni rezultati dolgoročnih sledilnikov na morski gladini so nasprotni od rezultatov, pridobljenih v [4]. Tam sicer neposredne primerjave s sledilniki, testiranimi v našem delu, niso opravili, so pa ugotovili, da je sledilnik TLD na njihovi morski podatkovni zbirki slabši od sledilnika Kernelized Correlation Filter (KCF) v natančnosti, priklicu in stopnji lažnega alarma. KCF je hiter sledilnik, vendar na izzivu VOT 2017 dosega manj kot polovico vrednosti EAO vseh tukaj testiranih kratkoročnih sledilnikov.

Prav tako se naši rezultati bistveno razlikujejo od rezultatov v [5], vendar se tam ne omejujejo na videoposnetke morske površine, kar omogoča izbiro širšega nabora tarč. Sledilnik TLD je v njihovih rezultatih dosegel le 0,27 F-ocene, najboljši pa je bil sledilnik FCLT z 0,48 F-ocene. Tako kot v našem testu, je tudi tam sledilnik CMT dosegel najslabše rezultate. Eden od razlogov za slabe rezultate na naši podatkovni zbirki je najbrž tudi ta, da CMT deluje popolnoma na podlagi značilnih točk, te pa potrebujejo tarče s teksturo. Veliko naših tarč je boj, ki odražajo le malo teksture.

5 Zaključek

V članku smo predstavili analizo kratkoročnih in dolgoročnih sledilnikov na videoposnetkih morskega okolja, posnetih iz USV. V ta namen smo iz obstoječih podatkovnih zbirk MODD1 in MODD2 zgradili dve podatkovni zbirki. Prva je namenjena testiranju kratkoročnih sledilnikov, druga pa dolgoročnih sledilnikov. Na njih smo testirali nekaj trenutno najboljših odprtokodnih kratkoročnih in dolgoročnih sledilnikov in primerjali njihove rezultate.

Rezultati kratkoročnih sledilnikov so na naši podatkovni zbirki zelo dobri, saj sledilniki nimajo večjih problemov s sledenjem v vodnem okolju. Za najboljši kratkoročni sledilnik se je izkazal sledilnik CFWCR [9]. Eks-



Slika 2: Prikazane so tri sličice videoposnetka, primernega za testiranje dolgoročnih sledilnikov. Na levi zgornji strani posamezne sličice je označena njena zaporedna številka, z rumeno barvo pa je označena resnična pozicija tarče. Tarča je na tem videoposnetku jadrnica, ki zaradi rotacije USV izgine iz vidnega polja, kasneje pa se znova pojavi.

perimenti z dolgoročnimi sledilniki so pokazali na večje razlike v kvaliteti sledenja med posameznimi sledilniki. Za najboljši dolgoročni sledilnik se je izkazal sledilnik TLD.

Možnosti nadaljnega dela je veliko. Spremenili bi lahko obstoječe sledilnike ali ustvarili nove. Podatkovna zbirka MODD2 vsebuje posnetke stereo kamer, kar bi lahko uporabili za izdelavo ali adaptacijo obstoječega sledilnika, ki za detekcijo in sledenje uporablja informacije globinske slike, pridobljene s kalibrirano stereo kamero.

6 Zahvala

Raziskava, ki jo predstavlja ta članek je bila deloma financirana iz programov ARRS P2-0095 in P2-0214 ter projekta ARRS J2-8175.

Literatura

[1] A. Moudgil in V. Gandhi, "Long-term visual object tracking benchmark", arXiv preprint arXiv:1712.01358, 2017.

[2] M. Kristan et al., "The Visual Object Tracking VOT2017 Challenge Results", 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, 2017, pp. 1949-1972.

[3] Y. Wu, J. Lim in M. H. Yang, "Object tracking benchmark", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, str. 1834-1848, Sept 2015.

[4] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally in C. Quek, "Videoprocessing from electro-optical sensors for object detection and tracking in maritime environment: A survey," IEEE Transactions on Intelligent Trans-portation Systems, vol. 18, str. 1993–2016, Aug 2017.

[5] A. Lukežič, L. Č. Zajc, T. Vojirr, J. Matas in M. Kristan, "Now you see me: evaluating performance in long-term visual tracking", *arXiv preprint arXiv:1804.07056*, 2018.

[6] M. Danelljan, A. Robinson, F. S. Khan in M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking", v *European Conference on Computer Vision*, str. 472–488, Springer, 2016.

[7] M. Danelljan, G. Bhat, F. S. Khan in M. Felsberg, "Eco: Efficient convolution operators for tracking", v *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, str. 6931-6939, July 2017.

[8] E. Gundogdu in A. A. Alatan, "Good features to correlate for visual tracking", *IEEE Transactions on Image Processing*, vol. 27, str. 2526-2540, May 2018.

[9] Z. He, Y. Fan, J. Zhuang, Y. Dong in H. Bai, "Correlation filters with weighted convolution responses", v *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, str. 1992-2000, Oct 2017.

[10] Z. Kalal, K. Mikolajczyk in J. Matas, "Tracking-learning-detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, str. 1409-1422, July 2012.

[11] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov in D. Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking", v *2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, str. 749–758, June 2015.

[12] G. Nebehay in R. Pflugfelder, "Clustering of static-adaptive correspondences for deformable object tracking", v 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), str. 2784–2791, June 2015.

[13] A. Lukežič, L. Č. Zajc, T. Vojir, J. Matas in M. Kristan, "Fclt- a fully correlational long-term tracker", *arXiv preprint arXiv:1711.09594*, 2017.

[14] M. Kristan, V. S. Kenk, S. Kovačič in J. Perš, "Fast image-based obstacle detection from unmanned surface vehicles", *IEEE Transactions on Cybernetics*, vol. 46, str. 641-654, March 2016.

[15] B. Bovcon, J. Perš, M. Kristan *et al.*, "Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation", *Robotics and Autonomous Systems*, vol. 104, str. 1-13, 2018.

Unconstrained Ear Recognition Using Residual Learning and Attention Mechanisms

Tim Oblak, Blaž Meden, Peter Peer, Žiga Emeršič

Faculty of Computer and Information Science, University of Ljubljana, Slovenia
E-mail: to1702@student.uni-lj.si, {blaz.meden, peter.peer, ziga.emersic}@fri.uni-lj.si

Abstract

With the recent popularity of deep convolutional neural networks, image-based biometrics is one of many domains, that consequently gained new progress in solving previously incomplete or unanswered challenges. While some biometric modalities, like the fingerprint, are already considered mature, others are still in need of more reliable approaches. Ears can be used for person identification since they have the necessary properties of a biometric modality. High accuracy ear identification systems do exist but mostly focus on a controlled environment. In this paper, we try to improve the current state-of-the-art in ear recognition by using residual learning and attention mechanisms. By stacking residual building blocks, we find the optimal architecture to be ResNet with 18 convolutional layers. We achieve a Rank-1 score of 54.46% with full model learning, which is a 5.35 percentage point improvement from the previous best trained on the VGG architecture, however, the model still underperforms against those, trained with selective learning. We observe that aggressive data augmentation is needed when dealing with a small dataset. We also conclude that the Attention Model performance is subpar compared to other architectures.

1 Introduction

In recent years, the use of deep neural networks in the broader field of machine learning brought us new achievements in a diverse range of challenges. Especially, the field of computer vision gained significant improvements with image recognition and object detection using deep *convolutional neural networks (CNN)*. They are known to have good generalization properties in an unconstrained environment. This includes images captured in the wild or on the go, which often contain noise, occlusion, different object rotations and lighting conditions. Such a property is especially useful in forensic, security or surveillance scenarios, where environment impact on the captured image is nondeterministic.

The application of deep learning is also progressively being used in the domain of biometrics, specifically with identification using certain biometric modalities, where until recently, expert knowledge or conventional computer vision local descriptors were used to tackle the problem. A biometric modality can be represented as a biological trait of a living being, which is unique and can be attributed to the being's identity. While recognition of some modalities, such as the fingerprint, are already considered mature, there are still those, that need better approaches to be reliable in practice.

In this paper, our main focus is ear recognition. As stated in [8], ears are known to have rigid biometric properties, such as uniqueness, longevity and a persistent shape, which does not change drastically throughout our lifetimes. Even though ear recognition systems have achieved some maturity, their success is still dependant on a controlled indoor environment.

In comparison to modalities such as eyes, face or fingerprints, there is still a lack of large-scale datasets available to the public. Even existing ones are usually small or only contain images captured in a constrained environment. With this in mind, we make use of data augmentation techniques to artificially inflate our training data.

To summarize, our first and foremost goal is to implement and test two specific deep neural network architectures that proved to achieve state-of-the-art results on some of the biggest public datasets, such as ImageNet [5], and compare them to the current best-performing architectures in the domain of ear recognition. In the process, we also want to measure the impact of data augmentation aggressiveness on the model performance.

2 Related work

A few extensive surveys have been made in this domain. In *A Survey on Ear Biometrics* [8], the authors make an extensive overview of detection techniques, feature extractors and available datasets, while providing experimental results and listing other open research questions. The report *Rank-1* identification rate ranges from 72.7% to 100%. Note, that all mentioned datasets are acquired in a controlled or semi-controlled environment. No deep neural network approaches were used. In *Ear Recognition: More Than a Survey* [18], the authors mostly focus on the recognition aspect. Their summary of known approaches lists *Rank-1* scores similar to the previously mentioned survey [8]. They also propose the *Annotated Web Ears (AWE)* dataset, which contains 1000 images of 100 persons captured on the internet and thoroughly annotated. It is an unconstrained dataset which represents real-world conditions. Again, none of the mentioned ap-

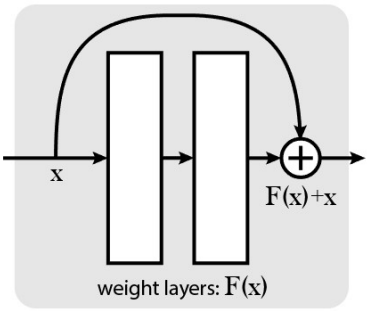


Figure 1: A residual building block used in ResNet and Residual attention network architectures.

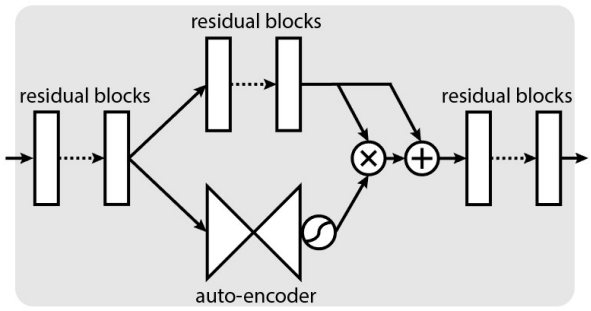


Figure 2: The attention module building block of the Residual Attention Network.

proaches outlined in the paper use any sort of deep learning. Roughly, the usual ear recognition approaches can be divided into 4 groups:

Geometric approaches focus on the extraction and analysis of the geometrical features, such as outer curves and helix point location. They rely on edge detection as their pre-processing step, so all of the textual information is therefore lost and not used for recognition. They are, however, usually scale and rotation invariant.

Holistic approaches encode the input image ear representations as a whole. They are prone to error when varying illumination or different pose of the subject is presented in the input image.

Local approaches make use of local neighborhood descriptors. Opposed to geometric approaches, they do not rely on relations between different points on the image, but between neighboring pixels in a small area. The descriptors can be calculated densely across the image or be selected beforehand for their uniqueness.

Deep-learning-based approaches mostly use CNNs to build the recognition model. They are usually learned end-to-end, which simplifies the recognition pipeline, however, the number of trainable parameters is higher in comparison to other machine learning algorithms. In [16], the authors achieved state-of-the-art results using deep learning with CNNs. They used full model learning and selective learning on top of features learned from the ImageNet dataset. A *Rank-1* score of 62.0% and a *Rank-5* score of 80.46% were achieved on the AWE dataset.

Lastly, different hybrid approaches also exist. These represent any combination of all above-listed approaches.

3 Methodology

In this section, we make an overview of the techniques we use to design and test new architectures. First, we present the main ideas behind existing architectures and then describe our workflow.

3.1 Deep Residual Networks

The residual learning approach was first published by Microsoft Research Asia [11, 12]. The main idea behind this architecture lies in its core building block, called the residual block. The block is shown in figure 1. It has two convolutional layers, where each of them also includes a batch normalization layer and an activation layer. The input and output of the block are connected with a "shortcut connection" which performs an identity mapping and its output is added to the output of the stacked layers. With this mechanism, the blocks that contribute to the increasing loss can be skipped. This allows the network to be even deeper without suffering from accuracy oversaturation and rapid degradation. These "shortcut connections" do not add any new parameters or computational complexity and can be used alongside common optimization techniques without changing the backpropagation calculations. Similar ideas were found previously in LSTM [10] and Highway Networks [9] and mostly used in a *recurrent neural network (RNN)* settings. The authors have published multiple versions, which include 50, 101, 152, 200 and even 1001 layer architectures.

3.2 Residual Attention Network

The Residual Attention Network is a CNN which is mostly made from previously mentioned residual block but also includes a truncated branch, which implements the attention mechanism. This mechanism was until recently mostly popular with RNN architectures in natural language processing and machine translation domains, where new samples are weighted based on the occurrence in previous iterations. The authors propose the Attention Module structure, shown in figure 2, which generates attention-aware features. The network is then constructed by stacking multiple Attention Modules.

Each of these modules can be divided into two branches; mask branch and trunk branch. The function of the trunk branch is to perform the usual feature processing and can be adopted into any network structure. The authors chose to use the residual block as their structure of choice. The mask branch is structured in form of an auto-encoder, also found in other architectures [3, 4]. The output of the mask branch is a soft mask, which, when joined with trunk branch output, acts as sort of a control gate function for every neuron output of the trunk branch.

The main motivation for using attention mechanisms lies in their ability to focus on a subset of inputs. This could be helpful in case of occlusion by hair or ear accessories.

3.3 Experiments

We devise a plan to test the described architectures in full. First, we test the effect of data augmentation. Best per-

forming augmentation rate will be used as the baseline. Then, we implement and test the residual attention network. There, some of the residual blocks are substituted with the Attention Modules in a way, that the final depth remains unchanged.

As of now, there is no pre-trained ImageNet model available for any of these architectures. Because of the lack of resources we are therefore not able to do selective learning and are only comparing the full model scores.

4 Experimental Setup

Here, we define our work environment including tools, datasets, performance metrics and the choice of hyper-parameters.

4.1 Dataset selection & augmentation

For our dataset, we use the images provided by the *Unconstrained Ear Recognition Challenge (UERC)* [17]. We divide the dataset into groups of 40% and 60% of images for the test and train data, respectively.

Since the size of the dataset is relatively small, generic data augmentation [7] is used to artificially inflate our training set. We make multiple augmentation configurations, which include no augmentation, $10\times$ and $100\times$ the augmentation rate for each image in the training set.

Images are first resized to a shape of $224 \times 224 \times 3$, which is the shape of the network input. Then, we follow the randomized augmentation process described in [16]. Before entering the network, images are normalized to a range between -1 and 1 and individual channel mean values are subtracted.

4.2 Performance metrics

To evaluate and compare trained models, we adopt the standard recognition performance metrics. *Rank-1* and *Rank-5* recognition rates are computed. For *Rank-1*, we check if the prediction with the highest probability matches the ground truth label. For *Rank-5*, we check if predictions with top five probabilities contain the ground truth label.

We evaluate the scores graphically by plotting the *Cumulative Match Curve (CMC)*. The graph shows the recognition rate of all possible ranks. We also compute the *Area Under Cumulative Match Curve (AUCMC)*, which summarizes the visual CMC numerically.

4.3 Research environment

Our experiments are made in *Python 3.4* programming language, and we use *Keras 2.0.5* [2] with a *TensorFlow 1.4.0* [14] backend as our main neural network library. We train and test the setup on a *ubuntu 16.04* computer, with a *GeForce® GTX 1080 Ti* graphics card, *Intel® Core™ i7-7700K* processor and *16 GB* physical memory.

4.4 Hyper-parameters

For the optimization algorithm, we use Adam [1]. Compared to stochastic optimizers such as SGD, it converges faster and has more momentum configuration options. It also requires less fine-tuning and can be set by only specifying the learning rate. We use a learning rate of 0.001, but reduce it by half when the loss function starts to converge. We use categorical cross entropy for computing the loss of the last softmax layer.

A custom data generator is used to supply images to the network in batches of 64. We do not set a fixed epoch number, but implement an early stopping feature, which stops the learning phase when loss function converges for some time.

Table 1: Effect of different data augmentation rates applied to dataset. Models are trained using the ResNet-18 architecture.

	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
No aug.	31.32	51.77	89.06
10x	36.70	56.62	91.18
100x	**54.46**	**73.19**	**95.24**

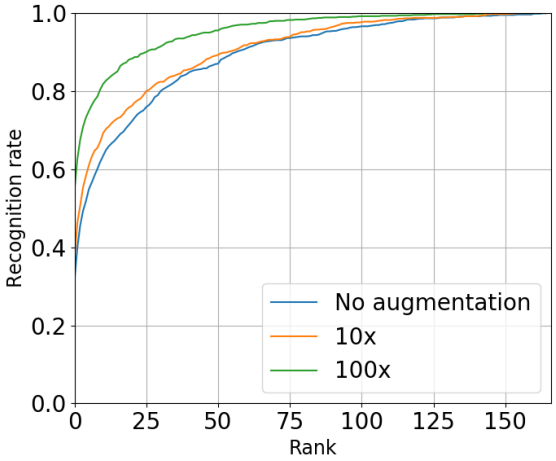


Figure 3: Cumulative Match Curves of ResNet-18 models with different data augmentation rates applied.

5 Results & Discussion

In this section, we present and discuss our results regarding architecture depth, type and the amount of data augmentation present.

5.1 Residual Network & data augmentation

The preliminary results have shown that, when comparing network depth, the 18-layer ResNet architecture performed best on the *UERC* dataset. We, therefore, used this configuration as our baseline. Also, data augmentation has a significant impact on the final model performance, as it is presented in table 1. The $100\times$ augmentation rate outperforms non-augmented dataset by 23.14 Rank-1 percentage points achieving a score of 54.46%. Cumulative match curves are shown in figure 3. With data augmentation, we fill the intra-class variance gaps caused by images gathered in the wild. This leads to better generalization, even in different environmental conditions.

5.2 Residual Attention Network

For our last experiment, we implemented the Attention Module from [13]. The architecture was then designed empirically by modifying the baseline ResNet-18. Here, we describe the best performing model. Again, we test the architecture with multiple data augmentation rates, but the configuration with $100\times$ data augmentation rate produced best results.

As presented in table 2, the model performed subpar compared to ResNet-18. It achieved a Rank-1 score of 37.67%. The best *Rank-1* score would only be comparable to ResNet-18 model with no data augmentation applied.

Table 2: Performance comparison of different ear recognition CNN architectures, trained with full model learning. We reference VGG as the previous-best model.

	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
VGG	49.08	66.67	92.99
ResNet	**54.46**	**73.19**	**95.24**
RAN	37.67	60.27	92.66

5.3 Final comparison

To summarize our findings, the results in table 2 illustrate, that ResNet-18 architecture outperforms other models, trained with full model learning. In [16], best results were achieved by using SqueezeNet [6] architecture, which achieved a *Rank-1* score of 62.00%, but was trained using selective learning, thus not comparable to our methods. We do, however, include the results achieved by VGG [15] architecture, as it performed best when trained by full model learning.

6 Conclusion

There are still many domain-specific challenges, that could be improved with the use of deep learning and ear recognition is definitely one of them, as we demonstrate with our experiments. The performance of standard computer vision and expert knowledge approaches is well documented in multiple extensive surveys, while there are still various possible improvements to be made with the use of CNNs.

We have applied two novel CNN structures to the area of ear recognition, ResNet and its modification using the Attention Module. We first illustrated the importance of aggressive data augmentation rate. Significant improvements in performance were measured when such an augmentation was applied.

We improved the previous-best unconstrained ear recognition model, which was also trained using full model learning. Note, that training models with selective learning using pre-trained weights still outperforms our methods by a margin of $\tilde{8}$ percentage points. On a dataset with 166 subjects and 2,304 cropped ear images, our best model achieved *Rank-1* and *Rank-5* scores of 54.46% and 73.19%, respectively.

In the future, we will train the ResNet-18 architecture using selective learning, as we believe that state-of-the-art unconstrained ear recognition results could be achieved.

References

[1] D. P. Kingma, J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[2] F. Chollet. Keras, 2015. Available at: https://github.com/fchollet/keras.

[3] V. Badrinarayanan, A. Kendall, R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[4] A. Newell, K. Yang, J. Deng. Stacked hourglass networks for human pose estimation. *European Conference on Computer Vision*, pages 483–499, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. ImageNet: A large-scale hierarchical image database. *Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, K. Keutzer. Squeezenet: Alexnet-level accuracy with $50\times$ fewer parameters and 1 MiB model size. *CoRR*, abs/1602.07360, 2016.

[7] L. Taylor, G. Nitschke. Improving deep learning using generic data augmentation. *CoRR*, abs/1708.06020, 2017.

[8] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, M. S. Nixon. A survey on ear biometrics. *ACM Computing Surveys*, 45(2):1–35, 2013.

[9] R. Srivastava, K. Greff, J. Schmidhuber. Training very deep networks. *Advances in Neural Information Processing Systems*, pages 2377—2385, 2015.

[10] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural Computation*, pages 1735—-1780, 1997.

[11] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] K. He, X. Zhang, S. Ren, J. Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision*, pages 630–645, 2016.

[13] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang. Residual attention network for image classification. *Conference on Computer Vision and Pattern Recognition*, pages 6450–6458, 2017.

[14] 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, X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Available at: https://www.tensorflow.org/.

[15] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[16] Ž. Emeršič, D. Štepec, V. Štruc, P. Peer. Training convolutional neural networks with limited training data for ear recognition in the wild. *International Conference on Automatic Face Gesture Recognition*, pages 987–994, 2017.

[17] Ž. EmerŠic, D. Štepec, V. Štruc, P. Peer, A. George, A. Ahmad, E. Omar, T. E. Boult, R. Safdari, Y. Zhou, S. Zafeiriou, D. Yaman, F. I. Eyiokur, H. K. Ekenel. The unconstrained ear recognition challenge. *CoRR*, abs/1708.06997, 2017.

[18] Ž. Emersic, V. Štruc, P. Peer. Ear Recognition: More Than a Survey. *Neurocomputing*, 255:26–39, 2017.

Experimental design for activity prediction in smart home environment

Gregor Donaj[1], **Mirjam Sepesy Maučec**[1], **Sašo Karakatič**[1], **Matej Borko**[2], **Andrej Žgank**[1]

[1] *Faculty of Electrical Engineering and Computer Science, University of Maribor*
[2]*A1 Slovenia*

E-mail: gregor.donaj@um.si, mirjam.sepesy@um.si, saso.karakatic@um.si, matej.borko@a1.si, andrej.zgank@um.si

Abstract

In this paper, we design an activity prediction framework to be used in a smart home environment. Because the data from target environments is not yet available, we propose an approach of using an open dataset and adopting it for the target environment. Two machine learning techniques were studied, Random Forest and HMM. Based on preliminary experiments, HMM provide better results and we decided to study it further in the future.

1 Introduction

Progress of technologies in machine learning and pervasive computing has generated interest in the development of smart environments that will assist humans with many valuable functions to improve living. A smart home is the most widely studied smart environment [1]. It has been researched for nearly a couple of decades, and it is even more intensive researched today. The main reason for increasing amounts of attention is the rising aging population. The concept of smart homes gives elder people hope to extend the period living independently within their home environment.

The idea of a smart home, stated more broadly, is to support the well-being of the residents of the home. Our goal, presented in this paper, is a prediction of residents' activity in the smart home environment. To develop the prediction system large and varied data sets need to be collected over time, as they are necessary to uncover hidden patterns and unknown correlations between data. Collecting them is expensive and time-consuming. In this paper, we propose an approach of using an open dataset and adopting it for the target environment.

The paper is organized as follows: Section 2 discusses related work. In Section 3 we describe the general features of CASAS open dataset that was chosen for transformation to our target environment. In Section 4 we describe the target environment and specify how the data from CASAS dataset was mapped to the foreseen data in the target environment as well as enriched with third-party environmental data. In Section 5 an overview of methods, commonly used for activity prediction is given. For our practical work, we selected two of them and performed initial experiments. We report the preliminary results. Section 6 concludes the paper. We give our plan for future work on the project.

2 Related work

Currently, numerous Smart Home projects intend to make daily life comfortable. Some of them were developed for proof-of-concept demonstration, and others were integrated into real life environments.

In [2] learning techniques were developed to discover patterns in the resident's daily activities and to generate automation policies that mimic these patterns. Ontological modeling and semantic reasoning were integrated into an agent-based system for activity recognition [3]. Data-driven approaches for predicting future activities shown that regression learners could be useful to learn an activity predictor that can reason about relational and temporal structure among the activities and to forecast future activity occurrences [4].

A prerequisite for the research on activity prediction is the availability of datasets which are large enough and ready to be used and analyzed. Because no data has been collected in our target environment yet, we have searched for open datasets available for public use. GCDC [5] is the collection of data about users cooking meals. It does not contain information about other activities. In d-WAR [6] datasets, users were wearing sensors, and the data was sent by the mobile phone to the base station. There were no stationary sensors. DRED [7] includes only sensor data, whereas activities are missing. The MIT [8], CASAS [9] and ARAS [10] datasets contain data about stationary sensors and activities. We decided to use CASAS data collection because the experimental environment in this project is most similar to our target environment.

3 CASAS dataset

The CASAS project took place at Washington State University [11]. The idea was to design a "smart home in a box" that is ready to be used to perform key functions out of the box. The CASAS data collection contains 32 smart home datasets: 19 datasets represent single-resident sites, 4 represent locations with two residents, and the rest house larger families or residents with pets. Most activity recognition algorithms have been tested in situations where a single individual is performing activities in a continuous, sequential manner. CASAS project focuses on a different complexity issue for home behavior, that of recognizing

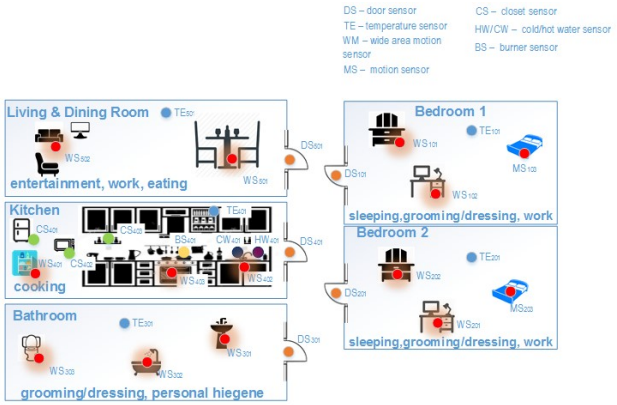


Figure 1: The target environment.

activities of more than one resident. The data collection includes 20 residents aged 21 to 62 years (mean 33 years), 8 males and 12 females, with a variety of background and technological familiarity [11].

In the CASAS project, the status of the residents and their physical surroundings are perceived using sensors. The examples of sensors are motion sensors, door sensors, temperature sensors, etc. Sensors generated events that consisted of a date, a time, a sensor identifier, and a sensor message. Sensor data were captured using a customized sensor network and stored in a database. Data was mapped to corresponding activity labels by using an activity recognition software, capable to label activities in real-time as sensor events arrive in a stream. The examples of activities are: bed-toilet transition, cook, eat, enter home, leave home, personal hygiene, etc. To maintain privacy the participant names and identifying information were removed and the collected data was encrypted before it was transmitted over the network.

In our experiments, only the following CASAS datasets were used: 7, 8, 11, 14, 15, 17 and 35. In datasets 7, 8, 11 and 14 one resident was living in an apartment, and in the rest, two of them.

4 Mapping from CASAS to target environment

The first step in building the dataset was to collect the appropriate files in the selected CASAS datasets, transform them into a consistent format, and make a preliminary analysis of the datasets. This analysis was necessary to define an appropriate target environment onto which the CASAS datasets could be mapped.

We also corrected some obvious errors (e.g., correcting the year from 22009 to 2009) and removed inconsistencies between the datasets (e.g., different column delimiter characters in datasets) to be later able to build a concise single-file dataset.

The mapping procedure is intended to construct a modified and combined dataset, which we call the IQdatabase, as this work is part of the IQ Home Program (http://www.iq-home.si/en/).

Table 1: Excerpt from the sensor mapping rules.

ID	Original sensors	Target sensors
7,8,11	D005, M37	DS301
7,8,11	M38	WS301
7,8,11	M39	WS302
7,8,11	M40,M41	WS303

Table 2: Excerpt from the activity mapping rules.

Original activity	Target activity
R1_Meal_Preparation	meal_preparation
R1_prepare_dinner	meal_preparation
R1_prepare_lunch	meal_preparation
R2_Meal_Preparation	meal_preparation

4.1 The target environment

After examining the selected datasets, we designed a custom target environment into which all data from the used CASAS datasets were mapped. The target environment is presented in Figure 1. The number and type of rooms are determined to fit the CASAS datasets. The most significant difference from the CASAS datasets is the use of a hallway with doors to all other rooms, while in the CASAS apartments some rooms were directly connected.

The sensors in the target environment have two-letter designations based on the type of sensor, eg. DS for door sensor (see figure), followed by a three digit number, where the first digit identifies the room. This way we were able to define a more consistent designation system compared to the original CASAS datasets.

4.2 Mapping procedures

The next step was to determine the mapping rules for sensors and activities. Examining the environments from the original CASAS datasets, the sensor data associated with different activities, the position of sensors, and the most likely rooms for activities, a mapping scheme was determined. A part of it is shown in Table 1 for sensor mappings and Table 2 for activity mappings.

In Table 1 we see that some rules are defined for several CASAS datasets, as they were collected in the same apartment. In Table 2 we see that some activities were combined into one, as not all original datasets had the same activities. Also, activities for both residents (R1 and R2) were mapped into one activity. Some activities were not mapped at all as they appear only in one of the datasets.

The original datasets, as well as our target environment dataset, are all text files. We, therefore, used simple Perl wrappers to perform all preprocessing and mapping.

4.3 Enriching data with environmental information

The context of the activities are the external data (besides the actions) that can be gathered from various sensors around or inside the home. CASAS datasets do not have

any external information in their provided form, so we were forced to use third-party sources for this task. The literature of research on this topic shows that there is a connection between subject actions and the weather and environment data inside and outside the monitored home.

For this reason, we wanted to enrich the CASAS data with external environment data from third party sources, to provide more valuable information, that can later be used in the pattern recognition process with artificial intelligence.

After extensive research, we decided to use weather information from Dark Sky service [12], where they offer free API to gather historical weather information for almost any location in the world. Their limitations on the API calls were inside of our call numbers, so there was no problem with this, and their terms of using their data also suited our type of work.

For gathering data from Dark Sky API, we used Python 3, and we used Python wrapper by Ze'ev Gilovitz [13], which provides an easy way to communicate with Dark Sky API with Python programming language.

After surveying the literature on the topic, we concluded to enrich CASAS data with the following environment data: Temperature (outside temperature in Celsius unit), summary (written summary of the weather conditions), icon (icon code used for the weather maps – can be used as an indicator of weather conditions), humidity (percentage of the outside humidity), air pressure (in millibars) and wind speed (in miles per hour).

4.4 The IQdatabase

After mapping, all data were combined into a final dataset, which we called IQdatabase (currently version 2.0). The database is formatted as a single text file with 48 columns per line. The columns indicate the original CASAS dataset number, date, time, the current status of 32 sensors, the state of 6 environmental data, and the current status of 7 activities, namely: work, dressing, personal hygiene, eating, meal preparation, sleeping, and TV.

To be able to perform experiments, the database was divided into a train set, a development set, and a test set. Both the development and train sets contain about a month of data and are taken at the end of the original dataset number 11, which is the largest original dataset.

The train set was further divided into 3 parts. The first part includes all other data from dataset 11. The second includes all data from other datasets from the same apartment as dataset 11. The third dataset includes all data from other locations. The sizes in terms of lines of data for all sets are given in Table 3.

Table 3: The data sizes in the IQdatabase.

Set	Size
Test	211,722
Development	194,220
Train 1	1,210,515
Train 2	663,363
Train 3	2,255,638
Train - total	4,129,516
Total	4,535,458

5 Activity prediction

Models, used for activity prediction, can be classified as hand-crafted, logic based or probabilistic. In hand-crafted models, training data is manually examined to find patterns that are used to predict activities. Logic-based models automatically learn rules from training data. The learning system is based on the principles of inductive logic programming (especially search, representations, operators, background knowledge) with transformation-based tagging (e.g., error-driven search).

Hand-crafted and logic-based models are deterministic. As activities are commonly performed in a non-deterministic way, probabilistic models are the most popular approaches for activity prediction. These models take into account that sensor readings in a real smart home are sometimes noisy. Several probabilistic models have been proposed to model the sequence classification problem of activity prediction: naïve Bayes classifiers, hidden Markov models, support vector machine, artificial neural networks, conditional random fields, random forest classifier, etc. In our research, we decided to use Random Forest (RF) and hidden Markov models (HHM). Both of them were successfully used in other researches [14].

5.1 Predicting with Random Forest

The Random Forest was proposed in [15] and further developed in [16] and is an ensemble learning algorithm, used for classification and regression problems. An RF classifier consists of a group or ensemble of decision tree predictors, each capable of producing a result when presented with a set of previously unknown data. The result of the decision tree takes the form of one of the predefined classes. The ensemble of decision tree results are aggregated, and the most popular class is chosen.

The main difference between the bag of decision trees and the RF method is that the bag of decision trees combines the result of independently constructed decision trees each on their own randomly selected subset of training data, but RF method also changes the way each decision tree in the ensemble is constructed. In the construction of one decision tree for the RF ensemble, each process of making a decision tree node consist of an additional step - the independent predictor value choosing for that particular node is done on a random subset of predictor values. Besides the random sampling of training data for each decision tree present in the creation of the decision tree, this step of randomly subsetting of available predictors ensures that more diverse trees are constructed and widen the search space for the classification prediction.

The problem of activity prediction is a time series problem, where prediction is made on the current information and the past data. Thus, we had to preprocess the available data in such way that every instance should also include information from the past, and so regular classi-

Table 4: Results of activity prediction in the IQdatabase.

	Precision	Recall	F1-score
Random Forest			
Train 1	0.75	0.26	0.27
Train 1+2	0.35	0.28	0.19
Train 1+2+3	0.25	0.20	0.14
HMM			
Train 1	0.47	0.21	0.15
Train 1+2	0.42	0.16	0.11
Train 1+2+3	0.45	0.18	0.14

fiers could perform prediction. This was done with the sliding windows procedure.

For our experiment, we used the following settings: window size: 10 past events; RF size: an ensemble of 100 decision trees in the forest. We used the default classifier output.

5.2 Predicting with Hidden Markov Models

Hidden Markov Models (HMM) are an extension of a Markov chain random process, hidden from the observer, and an observable random variable depended only on the current state of the Markov chain. HMMs are used to model processes, where the system has only input data, available through a noise channel. The most known example of using HMMs is acoustic modeling in speech recognition [17].

In our system, the hidden states represent the current activity in the environment. One possible state is also the absence of any activity. The Markov chain is the model for transitions between activities. The observable random variable is the state of all sensors and environmental data.

Since our training sets contain all activity and sensor data, it is simple to train the HMM. The Markov chain is directly trained based on all transitions between activities. The observable variable is trained as a set of independent variables. Most of them have discrete values, e.g., a sensor is ON in OFF.

5.3 Results

We performed two series of experiments. The first was based on Random Forest and the second one on HMM. In both series, the experiments were repeated three times, each time with a different training set. The results of preliminary experiments are collected in Table 4. Precision was selected to be the key metric. When Using all training material, HMM brought more promising results. We will continue the research with HMM in the future.

6 Conclusion

In this paper, we have proposed an approach to adopting an open dataset for activity prediction in smart home environments, for which the data is not yet available. HMMs have shown promising results for activity prediction and will be studied in more details in the future. Our next step is also to conduct Wizard of Oz experiments in real-life environments.

The application of these methods will be further applied in smart home environments for the IQ Home project, which is meant for establishing a supportive environment for smart home residents.

Acknowledgment

This work was partially funded by the Republic of Slovenia and the European Union under the European Regional Development Fund – IQ Home Program “Intelligent home of the new generation designed on smart appliances and wood”.

References

[1] L. C. De Silva, C. Morikawa and I. M. Petra. State of the art of smart homes. Engineering Applications of Artificial Intelligence, 25(7), 1313-1321, 2012

[2] P. Rashidi and D. J. Cook. Keeping the resident in the loop: Adapting the smart home to the user. IEEE Transactions on systems, man, and cybernetics-part A: systems and humans, 39(5), 949-959, 2009.

[3] L. Chen, C. D. Nugent, and H. Wang. A knowledge-driven approach to activity recognition in smart homes. IEEE Transactions on Knowledge and Data Engineering, 24(6), 961-974, 2012.

[4] B. D. Minor, J. R. Doppa, D. J. Cook. ”Learning Activity Predictors from Sensor Data: Algorithms, Evaluation, and Applications”, IEEE Transactions on Knowledge and Data Engineering, vol 29, no. 12, 2744-2757, 2017.

[5] GCDC, http://kitchen.cs.cmu.edu/ [Dostopano: 21.8.2018]

[6] d-WAR, https://people.eecs.berkeley.edu/˜yang/software/WAR/index.html [Dostopano: 21.8.2018]

[7] DRED, http://www.st.ewi.tudelft.nl/˜akshay/dred/ [Dostopano: 15.2.2018]

[8] MIT DB, http://courses.media.mit.edu/2004fall/mas622j/04.projects/home/ [Dostopano: 21.8.2018]

[9] CASAS, http://ailab.wsu.edu/casas/datasets/ [Dostopano: 21.8.2018]

[10] ARAS, https://cmpe.boun.edu.tr/aras/download.php [Dostopano: 21.8.2018]

[11] D. J. Cook, A. Crandall, B. Thomas, and N. Krishnan. CASAS: A smart home in a box. IEEE Computer, 46(6):26-33, 2013.

[12] DarkSky, https://darksky.net/dev/ [Dostopano: 21.8.2018]

[13] Python API, http://zeevgilovitz.com/python-forecast.io/ [Dostopano: 21.8.2018]

[14] E. Nazerfard, B. Das, L. B. Holder, and D. J. Cook. Conditional random fields for activity recognition in smart environments. In Proceedings of the 1st ACM International Health Informatics Symposium, pp. 282-286. ACM, 2010.

[15] T.K. Ho. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-44, 1998

[16] L. Breiman. Random forests. Machine learning, 45(1), 5-32, 2001

[17] M. Gales, S. Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing, 1(3), 195-304, 2007

Detekcija napak na materialih, ki izkazujejo periodično strukturo

Bilal Krupić[1], **Janez Perš**[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
E-pošta: janez.pers@fe.uni-lj.si

Detecting Defects in Materials That Exhibit Periodic Structure

We present a novel machine vision method, which is able to detect defects in materials that exhibit periodical structure. The method is based on 2D Fourier transform of the image. Our assumption is, that Fourier transforms of images showing periodical structure contain higher number of peaks than images in general. The peaks are detected using widely known MSER detector, and counted. The number of peaks represents one-dimensional image feature, which is proportional to the amount of periodic structure the image contains. We built a novel database of textile sample images, containing artificially made defects (cuts, tears). Database is fully annotated, and was used to evaluate the proposed algorithm.

1 Uvod

Avtomatsko zaznavanje površinskih napak je eden od pomembnejših problemov strojnega vida. V sodobni proizvodnji morajo biti izdelki brezhibni tako funkcionalno, kot tudi vizualno, saj vsako odstopanje od popolnosti zniža vrednost izdelka v očeh potrošnika. Metode strojnega vida za zaznavanje površinskih nepravilnosti so raznolike in segajo od preprostega upragovljanja sivinskih slik homogenih površin (kjer so napake jasno vidne) do metod, pri katerih uporabimo posebej zasnovano osvetlitev (npr. temno polje, angl. dark field illumination). Na ta način lahko postanejo jasno vidne tudi praske, razpoke in druge poškodbe, tudi če barvno ali svetlostno ne odstopajo od ostale površine.

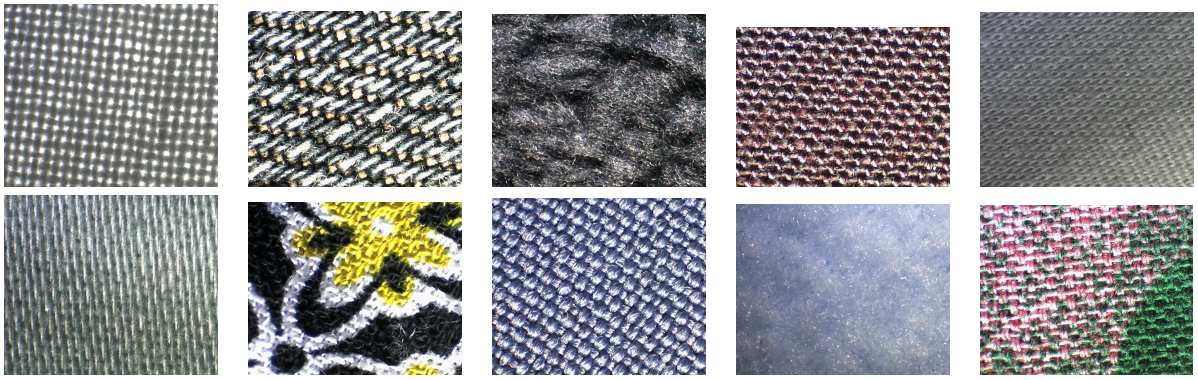
Poseben izziv so površine oziroma materiali, za katere ne moremo zgraditi točnega modela, ker je površinska struktura prekompleksna, naključna, ali pa bi izgradnja modela zahtevala nesorazmerno veliko dela (npr. teksturirani ali unikatni izdelki). V tem članku poskušamo nasloviti problem površin, ki vsebujejo *periodične vzorce*. Osnovna predpostavka našega pristopa je, da se napaka na vzorcu (površini), ki izkazuje periodičnost, izrazi kot porušitev (odsotnost) periodičnosti v tistem delu površine, kjer se napaka pojavi. Naš cilj je bil torej zasnovati metodo, ki bi sicer različne vzorce površin preslikala v takšen prostor, kjer je periodičnost močno izražena, in se jo da zaznati enostavneje kot na izvornih slikah. V nadaljevanju članka najprej predstavimo sorodna dela, potem našo metodo ter bazo slik, ki smo jo zajeli za testiranje našega pristopa. Prikažemo tudi rezultate kvantitativne ocene delovanja problema binarnega razvrščanja vzorcev na tiste brez napake in z napako, ter le kvalitativno ilustriramo delovanje detektorja napak na tekstilu, ki uporablja predstavljen razvrščevalnik ter metodo drsečega okna.

2 Sorodna dela

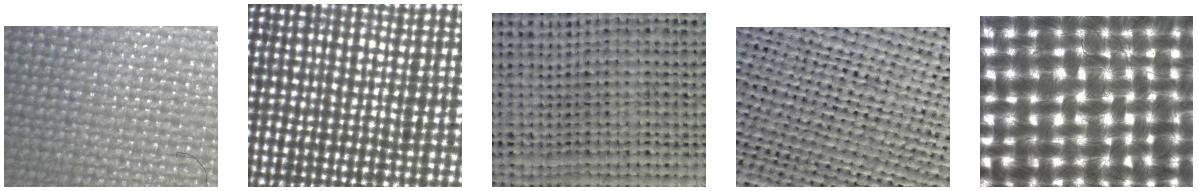
Osnovna ideja, ki se pojavi pri zaznavanju odstopanj od periodičnih vzorcev je seveda Fourierjeva transformacija, ki jo lahko na digitalnih računalnikih izvedemo v hitri diskretni različici (FFT, angl. Fast Fourier Transform). Idejo najdemo že v članku iz leta 1977 [1], kjer so avtorji predstavili metodo za zaznavo napak v integriranih vezjih, ki predpostavlja periodičnost opazovanega vzorca. Podrobno analizo izzivov, ki jih predstavlja analiza periodičnih vzorcev najdemo v [2], kjer avtorji predstavijo pristop z banko filtrov, ki bi naj naslovila te izzive.

Skoraj vsa ostala objavljena dela na tematiko analize napak v periodičnih vzorcih se nanašajo na tkane ali pletene tekstilne izdelke, ki so najpogostejši in najbolj množično proizvajan artikel, v katerem najdemo takšno strukturo materiala. Presenetljivo je, da uporabe Fourierjeve transformacije v teh člankih na način, podoben naši ideji, ne zasledimo. V članku [3] je opisan pristop, ki uporablja lokalno izračunane statistične lastnosti slike, klasifikacijo pa izvede s pomočjo perceptrona. Podoben pristop z analizo statističnih značilnosti in klasifikacijo z Bayesovim klasifikatorjem najdemo v [4]. V [5] uporabijo analizo področij (angl. blob) in njihovih lastnosti. Zanimivo je, da članek navaja, da je uspešnost človeških pregledovalcev tekstila pri odkrivanju napak le okrog 80%. Članek [6] predstavi algoritem, ki se najprej nauči lokalnih ponavljajočih se značilnosti, in to znanje uporablja pri detekciji odstopanj od teh značilnosti, podoben pristop z učenjem motivov na tkanini pa je uporabljen tudi v [7]. V [8] je opisana metoda, ki uporablja Gaborjeve filtre, odziv pa predstavi z binarno sliko, kjer je za vsak slikovni element podana odločitev, ali vsebuje napako, ali ne. Članek [9] predlaga uporabo medsebojne informacije (angl. mutual information), ki jo tudi sicer zasledimo pri analizi tekstur.

Tako [10] kot [11] se osredotočata na uporabo mehke logike za klasifikacijo napak na tekstilih. V [12] najdemo zanimiv pristop, ki uporablja diskretno kosinusno transformacijo (DCT) v okviru standarda za kompresijo slik JPEG za zaznavanje napak. Sodobni pristopi so med apli-



Slika 1: Vzorci tekstila, ki je bil uporabljen pri zajemu naše baze slik. Zaradi boljšega izkoristka ločljivosti prikazujemo 1/4 vsakega od vzorcev. Vzorci na sliki nimajo napak. Opazimo lahko, da tretji vzorec v prvi vrsti in četrti vzorec v drugi vrsti pravzaprav ne izkazujeta periodične teksture (material ni tkan ali pleten). Dva od vzorcev sta potiskana. Osvetlitve vzorcev na prikazanih slikah so različne in podrejene temu, da je vsak vzorec kar najbolje viden.



Slika 2: Primerjava štirih različnih načinov osvetlitve in dveh povečav prvega tekstilnega vzorca. Vse slike so brez napak. Prve štiri slike prikazujejo 10x povečavo in po vrsti osvetlitev od zgoraj, osvetlitev od spodaj, osvetlitev od strani (kot približno 30 stopinj) z 60 cm fluorescentno svetilko ter osvetlitev od strani z 30 cm fluorescentno svetilko. Peta slika kaže primer prvega vzorca, zajetega pri 60x povečavi z osvetlitvijo od spodaj.

kativnimi članki redki, izjema je [13], kjer je uporabljen avtoenkoder za rekonstrukcijo tekstilnega vzorca in posledično zaznavanje napak.

3 Zajeta baza slik

Precej sorodnih del testira algoritme na majhni bazi slik, kar po eni strani ne omogoča verodostojnega preizkušanja predlaganih algoritmov, po drugi strani pa ne omogoča uporabe globokega učenja. Zaradi obeh razlogov smo izvedli sistematičen zajem baze slikovnih vzorcev 10 različnih tekstilnih materialov, ki so prikazani na sliki 1.

3.1 Oprema in osvetlitev

Slike so bile zajete z digitalnim USB mikroskopom znamke Traveler, v ločljivosti 1280×1024 slikovnih elementov. Slike so barvne, programska oprema mikroskopa pa jih shrani v formatu JPEG, kar pomeni da že pride do določene stopnje degradacije, ki pa s prostim očesom ni opazna. Mikroskop omogoča 10x, 60x in 200x povečavo. Izkazalo se je, da sta za naše potrebe primerni le prvi dve povečavi, primerjavo pa vidimo na sliki 2. Mikroskop omogoča osvetlitev od zgoraj ali spodaj (presvetlitev), dodali pa smo še osvetlitev s strani, s krajšo (30 cm) in daljšo (60 cm) fluorescentno cevjo, ki poudarita globinski relief materiala (sence). Tudi primerjava med različnimi načini osvetlitve je prikazana na sliki 2.

3.2 Referenčne oznake in vzorčenje izrezov slik

Za vsakega od 10 materialov smo posneli različice z vsemi štirimi možnimi načini osvetlitve, dvema povečavama, ter tremi možnimi tipi napak (cefranje, zbadanje, rezanje, kot prikazuje sliki 3). Dodali smo četrto množico slik, ki ni vsebovala napak, vsak vzorec pa smo pod mikroskopom premikali in ga zajeli petkrat. Tako pridemo do množice 1600 zajetih slik. V povprečju je na slikah več kot ena napaka, tako da imamo na razpolago nekaj tisoč vzorcev z napakami.

Človeški operater je vse napake označil z zaključenimi poligoni, kot to prikazuje slika 3), naključno pa smo izbrali množico kontrolnih izrezov brez napak. Opozoriti je treba, da v tem članku nismo uporabljali tako podrobnih oznak za potrebe klasifikacije, ampak smo vsakemu poligonu izračunali očrtan pravokotnik, ki je bil vodilo za izrez vsakega od vzorcev, ki smo jih uporabili pri klasifikaciji.

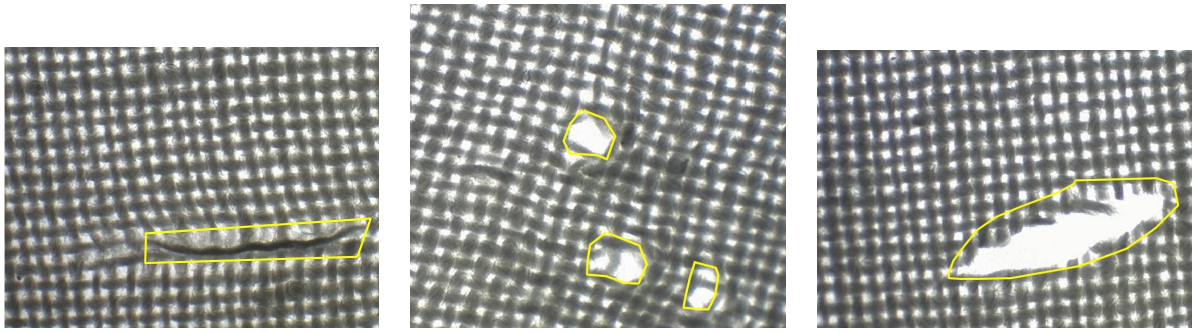
4 Algoritem razvrščanja vzorcev

Cilj algoritma je oceniti, kakšno stopnjo periodičnosti vsebuje vzorec. Izrez slike, ki ga želimo oceniti, pretvorimo v sivinsko sliko in na njem izvedemo dvodimenzionalno (2D) Fourierjevo transformacijo, s pomočjo algoritma FFT. Rezultat je 2D polje kompleksnih vrednosti, katerih absolutno amplitudo izračunamo, logaritmiramo in tako dobimo sivinsko sliko spektra, v kateri je frekvenca 0 v sredini slike, pojavi pa se več izrazito visokih vrednosti (vrhov), ki pripadajo periodičnim vzorcem v sliki.

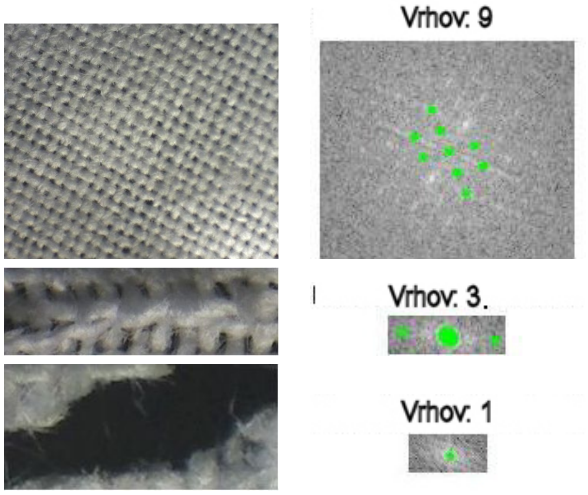
Te vrhove zaznamo z detektorjem MSER [14]. Izhod algoritma je skalarna vrednost, to je število zaznanih vrhov. Pričakujemo, da bo vrednost višja, če v periodičnem vzorcu ni nepravilnosti, in nižja, če je v vzorcu napaka. Delovanje algoritma prikazuje slika 4.

5 Eksperimenti in rezultati

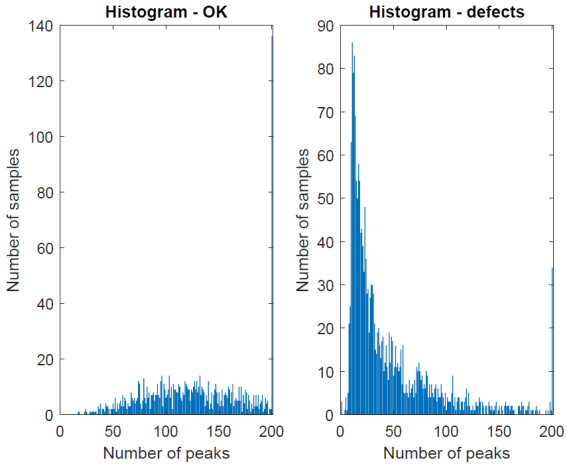
Delovanje algoritma je seveda močno odvisno od parametrov ter od pravilne izbire zajema slik (povečava, način



Slika 3: Primerjava treh različnih tipov napak. Po vrsti si sledijo: potrgana in izvlečena vlakna, prebadanje z ostrim predmetom ter rez v materialu. Z rumeno barvo so prikazane referenčne oznake napak (poligoni), ki jih je prispeval človeški opazovalec. V tem članku smo sicer kot referenčne oznake uporabili poligonom očrtana pravokotna okna, zaradi omejitve metode (2D Fourierjeva transformacija).



Slika 4: Primer delovanja algoritma. Na levi so izrezi slik, na desni pa amplitudni spektri 2D Fourierjeve transformacije z označenimi in preštetimi detekcijami MSER. V prvi vrsti je prikazan primer slike brez napak, algoritem najde veliko število vrhov. V drugi in tretji vrsti sta prikazana izreza slik z dvema tipoma napake, in rezultat detekcije vrhov v njunih amplitudnih spektrih.



Slika 5: Analiza delovanja predlaganega algoritma štetja vrhov na celotni bazi zajetih slik. Histograma prikazujeta porazdelitev števila zaznanih vrhov v amplitudnih slikah 2D Fourierjevih transformacij s pomočjo MSER detektorja z optimalnimi parametri. Na levi strani je prikazana porazdelitev za izreze slik vzorcev tekstila brez napak, na desni pa za izreze slik tekstila, ki vsebujejo napake. Vidimo, da imajo v splošnem vzorci z napakami bistveno manj detektiranih vrhov, kar je bistvo predlagane metode.

osvetitve, itd). Ugotovili smo, da je za tekstilne vzorce, ki so bili nam na voljo, 10x povečava optimalna, zato smo testiranje izvedli le na tem delu baze. Tudi v primeru dejanske industrijske aplikacije bi seveda ustrezno povečavo izbrali, zato je takšen način testiranja smiseln.

V našem primeru so parametri algoritma kar parametri detektorja MSER. 3/5 naše baze slik (2 od 5 premikov tekstila pod mikroskopom) smo uporabili za optimizacijo parametrov, rezultate pa prikazujemo na preostalih 2/5 baze. Optimizacijo parametrov smo izvedli z iskanjem po mreži (angl. grid search), pri čemer smo maksimizirali ploščino pod krivuljo ROC (ang. Receiver Operating Characteristic), ki jo sicer prikazuje slika 6. Delovanje algoritma smo preverili v večih korakih.

5.1 Ločljivost obeh kategorij vzorcev

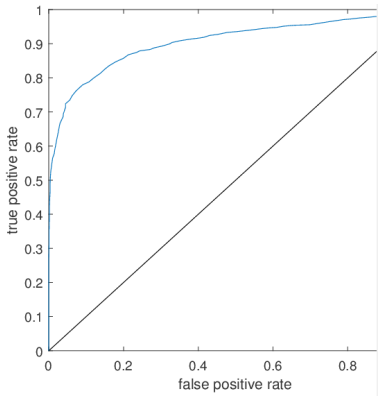
Če algoritem deluje, pričakujemo, da bo v vzorcih brez napak število vrhov pomembno višje, kot na vzorcih, ki vsebujejo napake. Izračunali smo histograma obeh kategorij vzorcev, ki sta prikazana v sliki 5. Vidimo da porazdelitvi ustrezata pričakovanjem.

5.2 Učinkovitost razpoznavanja

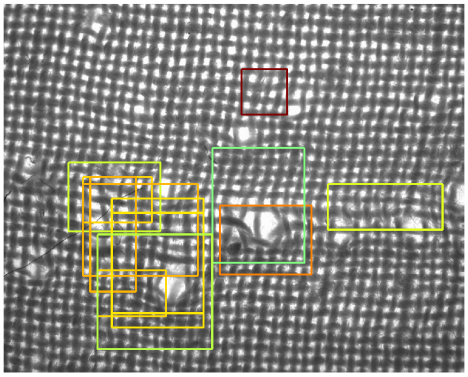
V drugem koraku smo problem obravnavali kot binarni razvrščevalnik. V našem primeru so pozitivni vzorci tisti, ki vsebujejo napake, negativni pa tisti, ki napak v tekstilu ne vsebujejo. Rezultat preizkušanja takšnega razvrščevalnika v obliki krivulje ROC prikazuje slika 6.

5.3 Primer detekcije

Sam razvrščevalnik še ne omogoča *iskanja* točk v tekstilu, kjer sumimo na napako. Za to ga moramo uporabiti v okviru koncepta *drsečih oken* (angl. sliding window). Ob tem je treba izbrati optimalno delovno točko – prag razvrščevalnika (ki leži na krivulji ROC), ter številne druge parametre (velikost okna, delež prekrivanja oken, itd). Izbira vseh parametrov je močno odvisna od tehnoloških parametrov proizvodnje (glede na to, kakšne napake pričakujemo), cene ročnega pregledovanja lažno pozitivnih detekcij napak, in tako naprej. Zato prikazujemo le primer detekcij na sliki 7, za hevristično izbran nabor parametrov, brez podrobne evaluacije.



Slika 6: Uspešnost binarne klasifikacije izrezov slik, prikazana kot krivulja ROC. Klasifikacija je bila zastavljena tako, da pozitivni primeri vsebujejo napake v tekstilu, negativni primeri pa napak ne vsebujejo.



Slika 7: Primer detekcije napak z drsečim rotiranim oknom. Detekcije so obarvane glede na razdaljo od pragu - zelene detekcije malo presežejo prag, rdeče ga močno presežejo, ostale barve so vmes.

Za primerjavo, nekateri avtorji [5] navajajo, da je človeški pregledovalec tekstila sposoben zaznati 80%napak.

6 Zaključek

Predstavili smo metodo razvrščanja tekstilnih vzorcev z napakami in brez napak, ki izkorišča dejstvo, da večina tekstilnih materialov izkazuje periodične lastnosti. Treba je upoštevati, da so prikazani rezultati (ROC krivulja) rezultat testiranja na celotni bazi, v kateri imamo tudi materiale, kjer je periodičnost slabo izražena, ter vse možne načine osvetlitve za vsak material. V praktični aplikaciji bi seveda uporabili le en tip osvetlitve, ki bi ga izbrali glede na lastnosti materiala. Podobno velja tudi za parametre algoritma MSER.

Naše nadaljnje delo bo uporaba metod globokega učenja, pri čemer bomo analizirali, ali je delovanje na slikah spektra boljše kot na surovih slikah materialov. Sklepamo, da bo učenje precej hitrejše in zahtevalo manj vzorcev, če se globoki nevronski mreži ni treba naučiti periodičnosti (ampak le-to ocenimo v okviru predobdelave).

Zahvala

Raziskava, ki jo predstavlja ta članek je bila deloma financirana iz programa ARRS P2-0095, projekta ARRS J2-9433, ter programa GOSTOP (št. pogodbe C3330-16-529000), financiranega s strani RS-MIZŠ in EU-ESRR.

Literatura

[1] L.U. Almi and J. Shamir. One-dimensional fourier transform for the inspection of photomasks. In E. MAROM, A.A. FRIESEM, and E. WIENER-AVNEAR, editors, *Applications of Holography and Optical Data Processing*, pages 541 – 547. Pergamon, 1977.

[2] Zong-Da Tsai and Ming-Hwei Perng. Defect detection in periodic patterns using a multi-band-pass filter. *Machine Vision and Applications*, 24(3):551–565, Apr 2013.

[3] C. Neubauer. Segmentation of defects in textile fabric. In *[1992] Proceedings. 11th IAPR International Conference on Pattern Recognition*, pages 688–691, Aug 1992.

[4] M. T. Habib, S. B. Shuvo, M. S. Uddin, and F. Ahmed. Automated textile defect classification by bayesian classifier based on statistical features. In *2016 International Workshop on Computational Intelligence (IWCI)*, pages 101–105, Dec 2016.

[5] A. Bodnarova, M. Bennamoun, and K. K. Kubik. Defect detection in textile materials based on aspects of the hvs. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 5, pages 4423–4428 vol.5, Oct 1998.

[6] D. Chetverikov. Structural defects: general approach and application to textile inspection. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 1, pages 521–524 vol.1, Sept 2000.

[7] H. Y. T. Ngan, G. K. H. Pang, and N. H. C. Yung. Patterned fabric defect detection using a motif-based approach. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II – 33–II – 36, Sept 2007.

[8] Hao Liu and Jiuqiang Han. Defect detection in textiles using optimal gabor wavelet filter. In *2006 6th World Congress on Intelligent Control and Automation*, volume 2, pages 10005–10007, June 2006.

[9] G. Abdel-Azim and S. Nasri. Textile defects identification based on neural networks and mutual information. In *2013 International Conference on Computer Applications Technology (ICCAT)*, pages 1–6, Jan 2013.

[10] N. A. Khalifa, S. M. Darwish, and M. A. El-Iskandarani. Automated textile defects recognition system using computer vision and interval type-2 fuzzy logic. In *2012 First International Conference on Innovative Engineering Systems*, pages 148–153, Dec 2012.

[11] S. M. Darwish. Soft computing applied to the build of textile defects inspection system. *IET Computer Vision*, 7(5):373–381, October 2013.

[12] F. Arnia and K. Munadi. Real time textile defect detection using glcm in dct-based compressed images. In *2015 6th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO)*, pages 1–6, May 2015.

[13] Y. Li, W. Zhao, and J. Pan. Deformable patterned fabric defect detection with fisher criterion-based deep learning. *IEEE Transactions on Automation Science and Engineering*, 14(2):1256–1264, April 2017.

[14] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and A. David Marshall, editors, *BMVC*. British Machine Vision Association, 2002.

Biomedicinska tehnika

Biomedical Engineering

Effect of electropermeabilization of cell membrane on metformin action on MDA-MB-231 breast cancer cells

Mojca Pavlin[1,2*], Jernej Repas[1,2], Maša Kandušer[1]

[1]*Group for nano and biotechnological applications, Faculty of electrical engineering, University of Ljubljana*
[1]*Institute of Biophysics, Faculty of Medicine, University of Ljubljana*
E-pošta: mojca.pavlin@fe.uni-lj.si

Abstract. *Interest in cancer metabolism opens entirely new area of research in cancer biology. Metabolic alterations in cancer cells enable novel targeted therapeutic approaches. Metformin, an established metabolic drug can be used as anti-cancer treatment and might act against cancer indirectly or systemically. However high discrepancies between achievable physiological concentrations in vivo and the effect of half maximal inhibitory concentration in vitro are not adequately explained. The drug enter cell cytoplasm by organic cation transporter and its effectiveness could limit intracellular concentration of the drug. The aim of our study was to enable direct access of metformin into cytosol by cell membrane permeabilization. We applied external electric pulses that cause transient cell membrane permeabilitation by electroporation and enable free diffusion of metformin into the cell. We tested the effect of electroporation on metformin activity on breast cancer cell line MDA-MB-231 in vitro. Our results indicate that metformin in combination with electroporation does not increases cytotoxicity of metformin in MDA-MB-231 cells.*

1 Introduction

Cancer cells differ from normal cells in several key biological characteristics as described in the renowned paper "The Hallmarks of Cancer" [1]. Recently, metabolic reprogramming in cancer cells has been recognized as a new hallmark of cancer cells. Metabolic alterations in cancer cells were already recognized in 1920s by Otto von Warburg who was first to observe that tumor cells maintain glycolysis even under normoxic conditions – known as Warburg effect. The re-emergence of interest in cancer metabolism opened entirely new area of research in cancer biology and oncology since metabolic alterations in cancer cells offer the possibility for novel therapeutic approaches that selectively target cancer cell metabolism and use of already established metabolic drug for anti-cancer treatment [2].

Epidemiological studies suggest that metformin, one of widely used type 2 diabetes drugs, might reduce the risk and mortality of cancer in type 2 diabetes [3]. On the one hand, metformin might protect against cancer indirectly by its systemic lowering of insulin level and glucose homeostasis. In addition of its systemic effects metformin targets cancer cells directly as an inhibitor of complex I of electron transport chain [4]. Breast cancer is one of most common cancers, among which triple-negative cancers represent most aggressive types. MDA-MB-231 cell line is a model cell line of triple negative human breast cancer cells that is very often used to study different anti-cancer agents.

Direct anti-cancer effects of metformin have been thoroughly examined in cultured MDA-MB-231 cells, but its effectiveness as cytotoxic effects remains questionable due to inconsistent *in vitro* results. Clearly, mechanisms that may link metformin to direct anti-cancer effects require further characterization. Metformin ameliorates systemic glucose homeostasis via at least two mechanisms. One mechanism involves activation of the AMP-activated protein kinase (AMPK) [5]. However, the mechanism of metformin action are not completely understood.

In vitro some studies have shown reduced viability after metformin treatment of MDA-MB-231 cells in concentrations as low as 30-500 μM while others have shown its effects on proliferation and viability only at very high concentrations (10-40 mM). However, in patients with type 2 diabetes peak plasma concentrations of metformin are 10–30 μM. While inconsistent *in vitro* results suggest that differences in experimental design might modulate sensitivity of MDA-MB-231 cells to metformin. Recently we have shown [6] that glucose level and protocol of medium renewal strongly affects metformin action *in vitro*. Still, high discrepancies between achievable physiological concentrations and the effect of half maximal inhibitory concentration (IC50) *in vitro* are not adequately explained. Based on those discrepancies it was hypothsed that due to its charge metformin can be concentrated inside mitochondria thus achieving much higher local concentration do to the electrochemical potential gradient. In addition, the effectiveness also depends on metformin ability to cross the plasma membrane. It is known that metformin needs organic cation transporter OCT1 for active transport across the plasma membrane [4]. Therefore, its cytoplasmic concentration could be increased by cell membrane permeabilization.

So far none of the studies examined the effects of metformin *in vitro* in combination with electroporation. Electroporation is a physical method that uses electric pulses to permeabilise cell membrane. It has been used to transfer small molecules or macromolecules into cells

in vitro and *in vivo* [7]–[9] and is currently used in clinical applications of electrochemotherapy [10], [11].

In this study, we examined the effect of metformin in combination with electroporation on cell survival in comparison with classical treatment with metformin alone. We used MDA-MB-231 cell line cultured *in vitro* and performed preliminary experiments on cell suspension.

2 Materials & Methods

2.1 MDA-MB-231 cells

Triple negative MDA-MB-231 breast cancer cell line was from ATCC (USA). MDA-MB-231 cells were maintained in high-glucose (4.5 g/l) RPMI-1640 medium (Genaxxon bioscience, Germany) without pyruvate, supplemented with 2 mM L-glutamine (Sigma-Aldrich) and 10% fetal bovine serum (FBS; Sigma-Aldrich). They were incubated in humidified atmosphere (95% air/5% CO2) at 37˚C. Experiments were performed in RPMI-1640 with 10% FBS and with 1 g/l (5.6 mM) glucose (Sigma-Aldrich). Cell culture medium was renewed every 24 hours.

2.2 Pulse protocols

Electric pulses were generated by electric pulse generator Electro cell B10 HVLV (BetaTech, France). Cells in suspension were prepared with tripsinization. Tripsinized cells were centrifuged at 1000 RPMI for 5 minutes and pelet was resuspended in KPB puffer (10 mM KH_2PO_4/K_2HPO_4 1 mM $MgCl_2$ and 250 mM sucrose) to obtain 10^6 cells/ml. 100 µl of the sample was electroporated in 4 mm Eppendorff cuvettes and 25 vol % of FBS was added to all cells after electroporation. The effect of electroporation in combination with four concentrations of metformin, 30 µM, 300 µM, 1 mM and 5 mM [6] was tested. Pulsing protocols consisted of 4 consecutive square pulses applied at repetition frequency 1 Hz. Three amplitudes of electric fields (E) were used: E = 0.6 kV/cm, 0.8 kV/cm 1, 1.2 kV/cm and 1.4 kV/cm. Pulse durations (t) was 200 µs. Control samples were not electroporated (Control 0 – negative control, E = 0 kV/cm and without metformin). Electric field strength (E) was calculated by the formula $E = U / d$, where U denotes applied voltage and d the electrode distance.

2.3 Determination of MDA-MB-231 cell viability by Hoechst and propidium iodide staining

The effect of metformin and electroporation on cell survival and morphology was inspected visually at inverted microscope (Zeiss Axiovert 200, Germany) under bight field at 10× objective magnification.

The survival of adherent MDA-MB-231 cells was determined 48h after experiment using Hoechst 33342 (Thermo Fisher Scientific) Cells treated in suspension were plated in 24 well plates with the cell density 4×10^4 per well and fluorescence was determined at 350 nm/461 nm (excitation/emission) using Tecan Infinite 200 (Tecan, Männedorf, Switzerland). To lyse cells, 0.03% SDS was added for 30 minutes at room temperature. To stain dsDNA, 100 mM NaCl, 50 mM TRIS-HCl, (pH=8.25) buffer, containing 5 µg/ml Hoechst 33342 was added. Background fluorescence intensity was subtracted and relative cell number in each sample was presented as the percentage of fluorescence intensity of treated samples relative to control sample (untreated cells). All experiments were performed at least in duplicates.

Cell membrane permeability was determined by propidium iodide (PI) staining. PI is cell membrane impermeant stain that can be used to detect cell membrane permeabilization. PI was added to KPB imediately before electric pulse application and cell membrane permeabilization was determined by flow cytometry (Attune NxT, Life Technologies). For each sample 10.000 events were recorded and analyzed.

Effects of metformin on cell energy metabolism was analyzed by measuring oxygen consumption rates (OCR; a measure of oxidative phosphorylation) and extracellular acidification rates (ECAR; a measure of glycolysis) of cells with Seahorse XFp Analyser. Basal and stressed (with 1 µM oligomycin and 1 µM FCCP injection) metabolic phenotypes of MDA-MB-231 cells were determined with Seahorse Energy Phenotype Test according to manufacturer instructions.

3 Results & Discussion

Electroporation experiments were performed to determine the effect of metformin on cell survival after exposure of MDA cells to different electric pulse amplitudes and metformin concentrations. Electroporation transiently permeabilize cell membrane allowing direct acces of metformin to the cytoplasm.
The effect of electroporation that causes cell membrane permeabilization on plated cells is shown in Figure 1. A bright field images of cells treated with different concentrations of metformin (the left column) and cell treated with metformin and permeabilized by electroporation (4×200 µs 1.2 kV/cm, 1 Hz, the right column) are presented.

Our results indicate that electroporation in combination with metformin does not have a significant impact on treated cells as can be observed in selected micrographs (Figure 1). The combined treatment does not affect cell survival quantified spectrofluorometrically (Figure 2). Both electroporation alone and metformin alone reduce cell viability to 80% and 60%. Those effects are consistent with previously published results [6], [7]. The combination of both does not result in further reduction of cell viability.

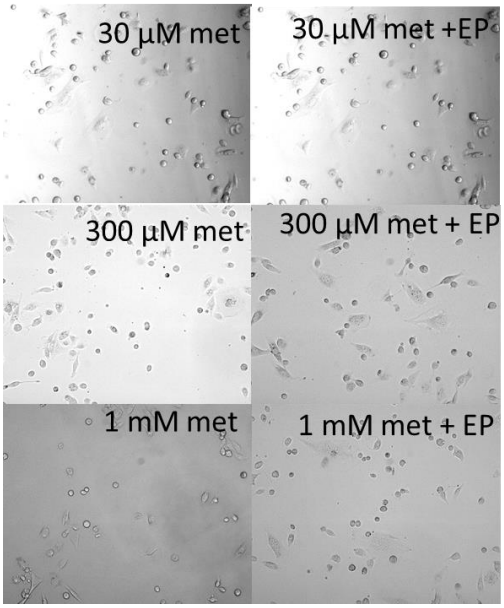


Figure 1: Bright field images of MDA-MB-231 cells treated with different concentrations of metformin alone or in combination with electoporation 48h after treatment. Electric pulse parameters were 4 x 200 µs, 1 Hz, 1.2 kV/cm. Metformine concentrations were 30 µM, 300 µM and 1 mM. Objective magnification 10×.

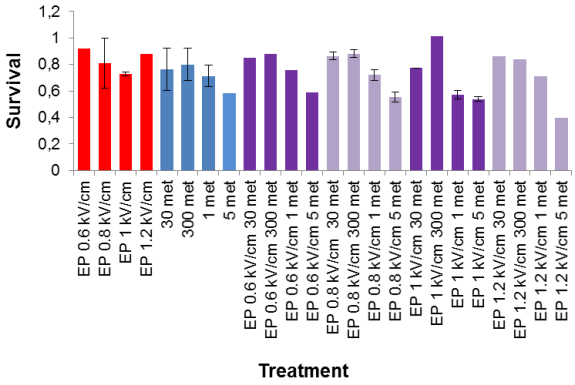


Figure 2: Cells treated in suspension: effect of increasing electric field strength and metformin concentration or combination of electric pulse parameters and concentrations of metformin on survival of MDA-MB-231 cells 48h after treatment

The effect of electric pulses on cell membrane permeability of MDA cell line was detected in a separate experiment and is presented in Figure 3 and Table 1.

The results indicate that almost all population of cells is efficiently permeabilized at 1 kV/cm allowing entrance of small molecules, as determined by PI staining (PI molecular weight 668.39 g/mol). Metformin is a small molecule with molecular weight 129.16 g/mol and can therefore enter the cells permeable to PI. The detected permeabilization of MDA cell line is similar to other cell lines of similar sizes [7].

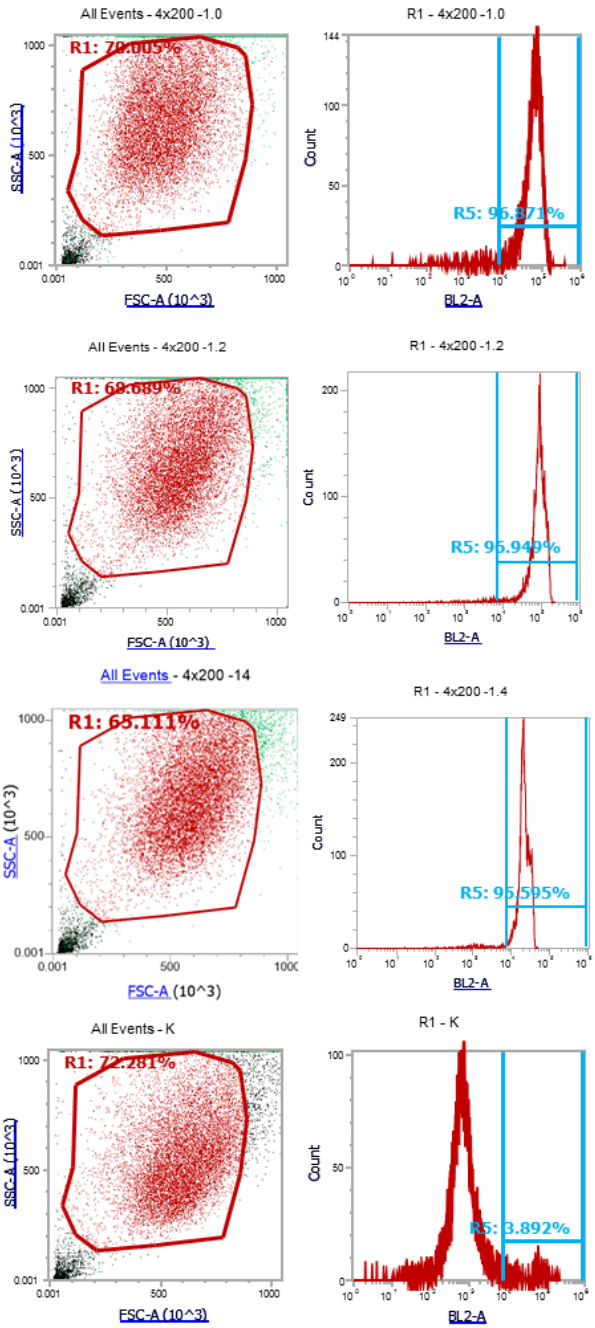


Figure 3: Cell permeabilization of MDA-MB-231 measured with PI uptake. On the left side scater and forward scatter of the samples, viable cells were gated in R1 region. On the right (R5) percentage of PI positive cells in the gated region determined on fluorescence channel BL2. Electric pulse parameters were 4 pulses of 200 µs duration, amplitude of 1, 1.2 and 1.4 kV/cm and repetition frequency 1 HZ. Control cells (at the bottom) were not exposed to electric pulses.

Table 1: The percentage of PI uptake by electroporated MDA-MB-231 cell in suspension. NK, negative control represents cells not exposed to electric pulses. For electroporation we used 4 electric pulses of 200µs duration amplitude of 1 kV/cm, 1.2 kV/cm and 1.4 kV/cm and repetition frequency 1Hz.

	NK	1 kV/cm	1.2 kV/cm	1.4 kV/cm
PI	4 %	97 %	97 %	96 %

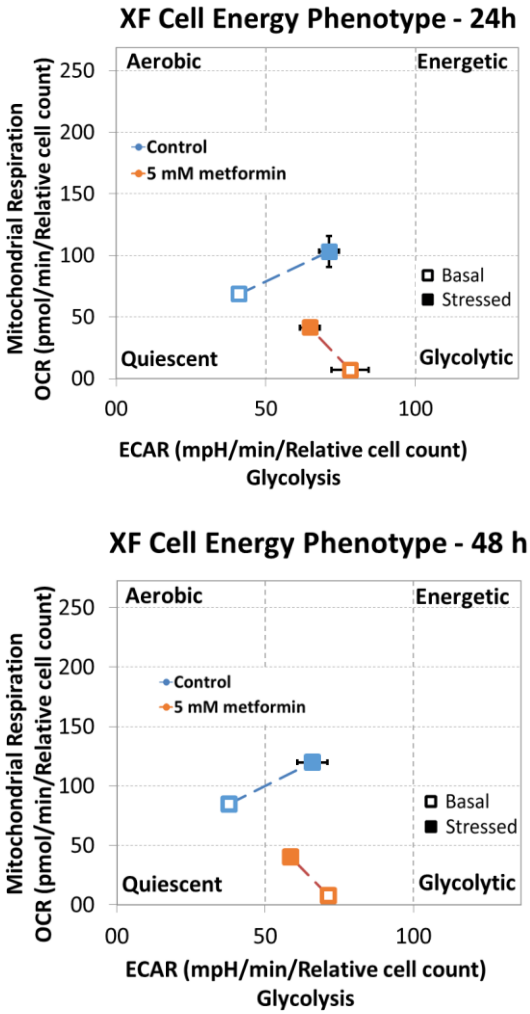


Figure 4: Cell energy metabolism of MDA-MB-231 cells treated with metformin 48h after treatment. Empty squares represent basal, while full squares represent stressed energy phenotype. T-bars indicate standard deviations between the wells. Results were normalized to relative cell counts.

Combination of electroporation and metformin in one treatment (Figure 1 and 2) does not further reduce cell viability indicating that OCT1 transporters [4] responsible for active transport of metformin across cell membrane are sufficient to obtain efficient intracellular concentrations of metformin. Namely, the increased cell membrane permeability caused by electric pulses of different amplitudes (Table 1, Figure 3) does no further increase cytotoxic effect of metformin alone [6].

This was further confirmed by Seahorse experiment, where it can be clearly seen that already after 24h metformin at 5 mM concentration almost completely blocks oxidative phosphorylation both in basal and stressed conditions as determined by decrease in OCR (oxygen consumption rate) consistently with its inhibition of mitochondrial complex I (Figure 4) one of the intracellular metformin targets [4]. Therefore, access of metformin to cell interior in non-electroporated cells is not a limiting factor in cell line MDA-MB-231.

To conclude, electroporation does not increase metformin action on MDA-MB-231 cells and probably its synergistic effects could be obtained in cell lines with low level of OCT1 transporters.

Acknowledgements

The study was supported by the Slovenian Research Agency (research projects No. J3-6794, J7-7424, research core funding P1-0055, young researchers program and MRIC UL IP-0510 Infrastructure program. The authors wish to express their gratitude to Agilent Technologies, Inc. and Chemass d.o.o. for providing access to Seahorse XFp analyser and Dr. Svetoslav Kalaydjiev for training and technical assistance.

Bibliography

[1] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[2] M. G. V. Heiden, L. C. Cantley, and C. B. Thompson, “Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation,” *Science*, vol. 324, no. 5930, pp. 1029–1033, May 2009.

[3] M. Bodmer, C. Meier, S. Krähenbühl, S. S. Jick, and C. R. Meier, “Long-term metformin use is associated with decreased risk of breast cancer,” *Diabetes Care*, vol. 33, no. 6, pp. 1304–1308, Jun. 2010.

[4] W. W. Wheaton *et al.*, “Metformin inhibits mitochondrial complex I of cancer cells to reduce tumorigenesis,” *eLife*, vol. 3, p. e02242, May 2014.

[5] S. Meng *et al.*, “Metformin Activates AMP-activated Protein Kinase by Promoting Formation of the αβγ Heterotrimeric Complex,” *J. Biol. Chem.*, vol. 290, no. 6, pp. 3793–3802, Feb. 2015.

[6] M. Rajh, K. Dolinar, K. Miš, M. Pavlin, and S. Pirkmajer, “Medium Renewal Blocks Anti-Proliferative Effects of Metformin in Cultured MDA-MB-231 Breast Cancer Cells,” *PLoS ONE*, vol. 11, no. 5, May 2016.

[7] M. Pavlin and M. Kanduser, “New Insights into the Mechanisms of Gene Electrotransfer - Experimental and Theoretical Analysis,” *Sci. Rep.*, vol. 5, p. 9132, Mar. 2015.

[8] A. Gothelf and J. Gehl, “What you always needed to know about electroporation based DNA vaccines,” *Hum. Vaccines Immunother.*, vol. 8, no. 11, pp. 1694–1702, Nov. 2012.

[9] M. Cemazar, G. Sersa, W. Frey, D. Miklavcic, and J. Teissié, “Recommendations and requirements for reporting on applications of electric pulse delivery for electroporation of biological samples,” *Bioelectrochemistry*, vol. 122, pp. 69–76, Aug. 2018.

[10] D. Miklavčič, B. Mali, B. Kos, R. Heller, and G. Serša, “Electrochemotherapy: from the drawing board into medical practice,” *Biomed. Eng. OnLine*, vol. 13, no. 1, p. 29, 2014.

[11] L. G. Campana *et al.*, “Treatment efficacy with electrochemotherapy: A multi-institutional prospective observational study on 376 patients with superficial tumors,” *Eur. J. Surg. Oncol. EJSO*, vol. 42, no. 12, pp. 1914–1923, Dec. 2016.

Vloga kognitivnih funkcij pri vožnji v simulatorju

Marina Horvat[1], Sara Tement[1], Bojan Musil[1], Grega Jakus[2], Jaka Sodnik[2]

[1]Oddelek za psihologijo, Filozofska fakulteta, Univerza v Mariboru, Koroška cesta 160, 2000 Maribor, Slovenija
[2] Laboratorij za informacijske tehnologije, Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška c. 25, 1000 Ljubljana, Slovenija
E-pošta: marina.horvat2@um.si

The role of cognitive functioning in a driving simulator

Abstract. *Driving is a complex task that requires quick adaptation to the changes in traffic and demands on the road. This study aimed to assess the changes in driving performance in different contexts of driving (city, highway) and at difficulty (low, high) levels. In addition, the role of cognitive functioning in these situations was examined. Participants were tested on a cognitive functioning test and drove different scenarios in a simulated driving environment. The results showed that driving performance was dependent on the context and difficulty level. This relationship was moderated by cognitive functioning. The obtained results suggest that besides accuracy and reaction time, the efficacy of attentional orienting is the most consistent cognitive predictor of driving performance. Overall, the study provides new insights of cognitive functioning in relation to driving performance.*

1 Uvod

Na tvegano vedenje v prometu vpliva več dejavnikov. Ob (ne)izkušenosti in starosti [1], na vožnjo vplivajo tudi psihološki dejavniki, kot so stališča do prometne varnosti [2], osebnost [3] in kognitivne funkcije [4]–[6]. Mednje uvrščamo tudi funkcije pozornosti, ki so med najpogosteje proučevanimi kognitivnimi funkcijami v povezavi z vožnjo tako v realnem okolju kot v simulatorju vožnje. Ugotovitve različnih raziskav kažejo, da je prav voznikova nepozornost pogost razlog tveganja za nesreče ali skorajšnje nesreče, kar je še posebej očitno v zahtevnejših okoljih (npr. pri gostejšem prometu) [7].

V preteklih letih se je zaradi dobro definiranih in preverljivih hipotez [8] pri proučevanju funkcij pozornosti na indikatorje vožnje uveljavil t.i. ANT model [9] (ang. *attentional network test*). Model predvideva obstoj treh ločenih, a koreliranih funkcij pozornosti [9]–[11]: izvršilni nadzor (ang. *executive control*) predstavlja mehanizem razreševanja kognitivnega konflikta. Orientacija pozornosti (ang. *attentional orienting*) je sposobnost prednostne selekcije informacij iz senzoričnega vhoda z izbiro modalnosti ali lokacije. Pripravljenost (ang. *alertness*) pa je doseganje in vzdrževanje stanja visoke občutljivosti na prihajajoč dražljaj in se deli na (a) stanje pripravljenosti, ki se nanaša na pripravljenost odziva v kratkem časovnem obdobju po opozorilnem znaku ter (b) tonično pripravljenost, ki se nanaša na vzdrževanje pozornosti v daljšem časovnem obdobju. Dosedanje študije ugotavljajo, da je pri napovedovanju parametrov vožnje pomemben celoten reakcijski čas na kognitivni nalogi [8], medtem ko rezultati novejših študij kažejo na pomembno vlogo orientacijske učinkovitosti v situacijah ob približujoči se nevarnosti [5]. V preteklih študijah so udeleženci običajno vozili le v enem scenariju vožnje, pri čemer so avtorji proučevali vedenje udeležencev neposredno pred izpostavljenostjo nevarni situaciji [5], primanjkuje pa študij, ki bi proučevale diferencialno vlogo kognitivnih funkcij v različno kompleksnih scenarijih.

Namen pričujoče študije je bil ugotoviti, ali lahko na podlagi rezultatov testa pozornosti napovedujemo posameznikovo vožnjo v simulatorju. Zanimalo nas je še, kako se različni indikatorji vožnje spreminjajo v odvisnosti od konteksta in zahtevnosti vožnje ter kakšna je vloga različnih funkcij pozornosti v tem odnosu.

2 Metoda

2.1 Udeleženci

V eksperimentu je sodelovalo 50 udeležencev, od tega 3 udeleženci zaradi občutka slabosti med vožnjo v simulatorju [12] niso dokončali eksperimenta. Delež udeležencev s slabostjo je primerljiv z deležem v drugih študijah [13]. Končni vzorec je tako sestavljalo 47 zdravih odraslih udeležencev (55,3 % žensk). Razpon starosti udeležencev je bil med 20 in 51 let (*M* (aritmetična sredina) = 30,34; *SD* (standardna deviacija) = 10,15). Vsi udeleženci so imeli veljavni vozniški izpit od 2 do 33 let (*M* = 11,87; *SD* = 9,78). Večina udeležencev (92%) je poročala, da avtomobil vozi vsaj nekajkrat na teden, približno polovica vozi avtomobil tedensko na avtocesti.

2.2 Pripomočki

Test pozornosti. Za merjenje pozornosti udeležencev smo uporabili test omrežja pozornosti za interakcije in pazljivost reševanja (ANTI-V) [3, 4]. Dražljaji so bili predstavljeni v programu E-prime v2.0 [15] na 24" zaslonu računalnika. Na začetku je bilo udeležencem predstavljeno sivo ozadje v obliki ceste s točko fiksacije (za ohranjanje pogleda udeleženca na eno točko), ki je ostalo vidno do konca naloge. Vsak dražljaj je bil sestavljen iz petih avtomobilov v vrsti (en za drugim) nad ali pod točko fiksacije. Naloga udeleženca je bila, da določi smer avtomobila na sredini vrste s pomočjo

tipkovnice računalnika. Vsak dražljaj je bil prikazan 200 ms, udeleženci pa so imeli do 2000 ms časa za odziv. V nekaterih primerih je bil avtomobil v sredini obrnjen v enako smer kot ostali avtomobili (kongruenten dražljaj), v drugih pa v obratno smer (inkongruenten). S tem smo ocenili izvršilni nadzor pozornosti. Za oceno omrežja orientacije je bil 100 ms pred prikazom vrste avtomobilov prikazan vizualni namig pod ali nad točko fiksacije (zvezdica * za 50 ms). Namig je lahko bil veljaven (zvezdica na enaki lokaciji kot prihajajoč dražljaj), neveljaven (na nasprotni lokaciji) ali pa namiga pred dražljajem ni bilo. Delovanje omrežja pripravljenosti smo ocenili preko zvočnih signalov v obliki zvočnega opozorila, ki je nastopil 50 ms pred dražljajem (pogoj z opozorilom) ali pa zvočnega opozorila ni bilo.

Testna naloga je bila sestavljena iz 4 nizov (v vsak nizu je bilo predstavljenih 64 dražljajev). Pred testnimi dražljaji je bilo z namenom privajanja na nalogo udeležencu predstavljenih 32 poskusnih dražljajev s povratno informacijo o uspešnosti reševanja. Pri testnih serijah povratne informacije o uspešnosti ni bilo.

Simulator vožnje. V eksperimentu smo uporabili simulator vožnje podjetja NERVteh [16]. Pri tem je udeleženec sedel v simulatorju, opremljenim z vozniškim sedežem in je imel nadzor nad vozilom kot v običajnem avtomobilu (volan, smerniki, stopalke ipd.). Pred seboj je imel tri ukrivljene 48" zaslone, ki omogočajo večji zorni kot in boljši pregled nad okolico vozila. Na začetku je vsak udeleženec imel na voljo nekaj minut za privajanje na simulator. Nato so udeleženci sodelovali v štirih različnih scenarijih vožnje. Scenariji so se razlikovali v kontekstu vožnje (mesto, avtocesta) ter stopnji zahtevnosti (lažji, zahtevnejši scenarij). Lažji pogoj je v obeh kontekstih pomenil manj avtomobilov na cesti, manj nepredvidljivih situacij, prehitevanj ipd., težji pa več avtomobilov, večjo izpostavljenost nevarnim situacijam, več sprememb v prometni signalizaciji ipd. Vsak scenarij je trajal 5 minut, med različnimi scenariji so udeleženci imeli možnost kratkega počitka.

2.3 Postopek

Poskus je potekal v dveh prostorih Filozofske fakultete Univerze v Mariboru. V prvem prostoru so udeleženci individualno reševali kognitivno nalogo, ki je trajala pribl. 30 minut, zatem so šli v prostor, kjer je bil nameščen simulator vožnje. Vsi udeleženci so vozili v enakih zgoraj omenjenih scenarijih, pri tem so začeli z lažjim pogojem in nadaljevali s težkim, vrstni red konteksta (avtocesta, mesto) je bil naključen. Udeleženci so dobili navodila, da vozijo kot bi vozili v naravnem okolju. Vožnja v simulatorju je v povprečju trajala med 30 in 40 minut.

2.4 Analiza podatkov

Analizo podatkov smo izvedeli s statističnim programom SPSS 24. Za vsak pogoj kognitivne naloge smo izračunali povprečni reakcijski čas (v ms) ter točnost (delež napak). Za vsakega udeleženca smo izračunali tri zgoraj funkcije pozornosti: izvršilni nadzor, orientacijo pozornosti in stanje pripravljenosti v obliki reakcijskih časov in pravilnosti reševanja. Izvršilni nadzor smo izračunali kot razliko med inkongruentnimi in kongruentnimi dražljaji (višja vrednost kaže na večji vpliv inkongruentnih dražljajev). Orientacijo pozornosti smo izračunali kot razliko med neveljavnimi in veljavnimi namigi; stanje pripravljenosti smo izračunali samo za dražljaje brez vizualnega namiga in sicer kot razliko med zvočnim signalom in dražljajev brez njega. Poleg tega smo izračunali še reakcijski čas in pravilnost reševanja (delež pravilnih odgovorov) za vse dražljaje skupaj.

Indikatorje vožnje v simulatorju smo izračunali za vsak scenarij posebej in so vključevali:

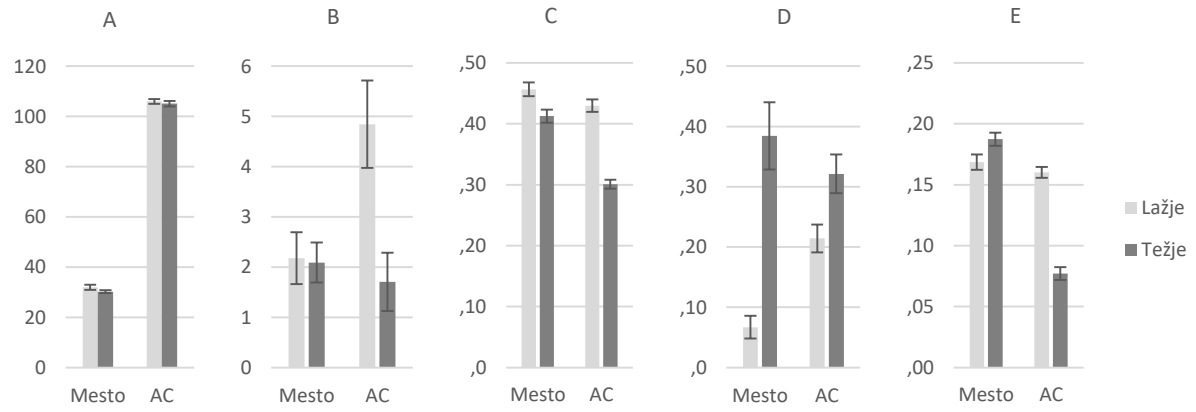
- povprečno hitrost vožnje ter povprečno prekoračitev omejitve hitrosti,
- standardno deviacijo odmika od sredine voznega pasu (mera vijuganja),
- kršitev varnostne razdalje, izračunano kot razmerje med časom kršenja varnostne razdalje in skupnim časom vožnje ter
- delež prekomernih trzljajev (trzljaj je odvod pospeška), ki je mera za sunkovitost vožnje.

Za testiranje razlik v posameznih indikatorjih vožnje glede na scenarij smo uporabili večsmerno analizo variance (ANOVA) za ponovljene meritve (RM ANOVA) s kontekstom (mesto, avtocesta) in zahtevnostjo (lažji, zahtevnejši) kot ponovljenima faktorjema. Učinek funkcij pozornosti glede na kontekst in zahtevnost smo definirali kot faktor med skupinami. Pri parnih primerjavah smo upoštevali Bonferronijev popravek. Rezultate statističnih testov smo dopolnili z velikostmi učinka (delna eta^2). Korelacije med funkcijami pozornosti in indikatorji vožnje smo izračunali s Pearsonovim koeficientom korelacije.

3 Rezultati

3.1 Rezultati na testu pozornosti

Glavni učinki 2 (opozorilo) x 3 (veljavnost) x 2 (kongruentnost) RM ANOVE za reakcijske čase so bili signifikantni. Posamezniki so bili hitrejši, kadar so (1) pred dražljajem slišali zvočni signal, $F(1,46) = 16,64$; $p < 0,001$; $eta^2 = 0,27$; (2) se je prikazal veljavni namig v primerjavi z neveljavnim namigom in brez namiga, $F(2,92) = 14,51$; $p < 0,001$; $eta^2 = 0,24$. Parne primerjave so potrdile razlike med vsemi pogoji ($p < 0,001$). Ter (3) ko je bil prikazan kongruentni v primerjavi z inkongruentnim dražljajem, $F(1,46) = 107,9$; $p < 0,001$; $eta^2 = 0,70$. Od interakcij se je potrdila samo interakcija med veljavnostjo in kongruentnostjo, $F(2,92) = 3,25$; $p = 0,04$; $eta^2 = 0,07$. Nadaljnje analize so pokazale, da je bil učinek kongruentnosti višji pri neveljavnem namigu v primerjavi z veljavnim, $F(1,46) = 5,5$; $p = 0,02$; $eta^2 = 0,11$, medtem ko je bila razlika pri veljavnem namigu in pri pogoju brez namiga manjša in ni bila statistično pomembna, $F(1,46) = 1,51$; $p = 0,23$; $eta^2 = 0,03$. Rezultati natančnosti reševanja so pokazali signifikanten



Slika 1. Posamezni indikatorji vožnje glede na kontekst vožnje (mesto, avtocesta - AC) ter težavnost (lažji, težji pogoj) s standardnimi napakami (SE); A = povprečna hitrost, B = prekoračitev hitrosti, C = vijuganje, D = kršenje varnostne razdalje, E = prekomerni trzljaji med vožnjo.

glavni učinek samo za kongruentnost, $F(1,46) = 54,13$; $p < 0,001$; $eta^2 = 0,54$; ostala glavna učinka nista bila statistično pomembna, prav tako ni bila statistično pomembna nobena interakcija.

3.2 Rezultati simulatorja vožnje

Različni indikatorji vožnje v simulatorju glede na kontekst in težavnost scenarija so prikazani na sliki 1. Rezultati RM ANOVE so pokazali, da so udeleženci v kontekstu avtoceste vozili hitreje kot v mestnem kontekstu, $F(1,46) = 5378,42$; $p < 0,001$; $eta^2 = 0,99$; udeleženci se niso razlikovali glede na težavnost pogoja (Slika 1A). Udeleženci so v povprečju večkrat prekoračili hitrost na avtocesti, $F(1,46) = 6,03$; $p = 0,02$; $eta^2 = 0,12$ in v lažjem pogoju, $F(1,46) = 11,55$; $p = 0,001$; $eta^2 = 0,20$. Statistično pomembna interakcija kontekst x težavnost kaže, da je težavnost imela večji učinek v kontekstu avtoceste, v primerjavi z mestom, $F(1,46) = 11,56$; $p = 0,001$; $eta^2 = 0,20$ (Slika 1B). Indikator vijuganja (Slika 1C) se je statistično pomembno razlikoval glede na kontekst, $F(1,46) = 73,52$; $p < 0,001$; $eta^2 = 0,62$; težavnost, $F(1,46) = 125,48$; $p < 0,001$; $eta^2 = 0,73$; statistično značilna je bila tudi interakcija kontekst x težavnost ($F(1,46) = 32,42$; $p < 0,001$; $eta^2 = 0,41$), ki kaže, da je bil učinek težavnosti bolj očiten v kontekstu avtoceste. Kršitev varnostne razdalje se v povprečju ni razlikovala glede na kontekst (Slika 1D), $F(1,42) = 1,32$; $p = 0,26$; $eta^2 = 0,03$; statistično pomemben je bil učinek težavnosti ($F(1,42) = 53,65$; $p < 0,001$; $eta^2 = 0,56$), ki je bil bolj izrazit pri vožnji v težjem pogoju v primerjavi z lažjim, $F(1,42) = 6,35$; $p = 0,02$; $eta^2 = 0,13$. V številu trzljajev je prišlo do pomembnih razlik glede na kontekst vožnje ($F(1,46) = 187,59$; $p < 0,001$; $eta^2 = 0,80$), kjer jih je bilo primerjalno več v mestni vožnji (Slika 1E), ter glede na težavnost, $F(1,46) = 117,93$; $p < 0,001$; $eta^2 = 0,88$), kjer je bilo v povprečju več trzljajev v lažjem pogoju. Učinek težavnosti je bil bolj izrazit pri vožnji na avtocesti, $F(1,46) = 332,91$; $p < 0,001$; $eta^2 = 0,88$).

3.3 Povezava med funkcijami pozornosti in indikatorji vožnje

Korelacijska analiza je pokazala, da je delež pravilnih odgovorov povezan z manj vijuganja v vseh pogojih ($r = -0,23$ do $-0,39$; $p < 0,1$), ter z višjo povprečno hitrostjo v vseh pogojih ($r = 0,18$ do $0,41$; $p \approx 0,1$). Daljši reakcijski čas je bil povezan z več vijuganja v skoraj vseh pogojih ($r = 0,19$ do $0,38$; $p \approx 0,1$). Najbolj konsistentno je bila z indikatorji vožnje povezana funkcija orientacije. Boljša orientacijska učinkovitost je bila povezana z nižjo hitrostjo ($r = -0,29$ do $-0,49$; $p < 0,05$), razen v pogoju zahtevnejše avtoceste. Podobno se je pokazalo tudi za prekoračitev hitrosti. Boljša orientacijska učinkovitost je bila povezana z manj trzljaji ($r = -0,17$ do $-0,29$; $p \approx 0,2$); oboje je bilo bolj značilno za vožnjo v mestu, medtem ko na avtocesti, učinka praktično ni bilo več. Ostale funkcije pozornosti so bile nekonsistentno povezane z indikatorji vožnje.

Na koncu smo preverili še, ali imajo funkcije pozornosti diferencialni učinek glede na kontekst in težavnost scenarija. Orientacija pozornosti je statistično pomembno moderirala odnos med težavnostjo in povprečno hitrostjo ($F(1,45) = 4,57$; $p = 0,04$; $eta^2 = 0,09$). Udeleženci s slabšo orientacijsko učinkovitostjo so v težjih pogojih bolj upočasnili hitrost v primerjavi z udeleženci z boljšo orientacijsko učinkovitostjo. Udeleženci z boljšo izvršilno kontrolo so v lažjem pogoju sicer pogosteje prekoračili hitrost, vendar so v težjem pomembno zmanjšali prekoračitev hitrosti, $F(1,45) = 4,23$; $p = 0,04$; $eta^2 = 0,09$. Do interakcije med težavnostjo in izvršilno kontrolo je prišlo tudi pri kršitvi varnostne razdalje, $F(1,41) = 5,28$; $p = 0,03$; $eta^2 = 0,11$, pri čemer so udeleženci z boljšo izvršilno kontrolo pri večji težavnosti manjkrat kršili varnostno razdaljo. Do interakcije med kontekstom vožnje in orientacijo pozornosti je prišlo pri vijuganju, $F(1,45) = 6,75$; $p = 0,01$; $eta^2 = 0,13$ in pri sunkovitosti vožnje, $F(1,45) = 7,73$; $p = 0,01$; $eta^2 = 0,15$. Pri vožnji na avtocesti so udeleženci s slabšo orientacijsko učinkovitostjo manj vijugali v primerjavi z udeleženci z boljšo orientacijsko učinkovitostjo, medtem ko pri vožnji v mestu ni bilo razlik. Po drugi strani so udeleženci s slabšo orientacijsko učinkovitostjo vozili manj sunkovito v mestnem okolju, na avtocesti razlik med skupinama ni bilo.

4 Diskusija

Namen raziskave je bil proučiti povezanost funkcij pozornosti z indikatorji vožnje v simulatorju ter raziskati vlogo le-teh v različnih kontekstih (mesto, avtocesta) ter težavnostih (lažji, težji) scenarija.

Rezultati na nalogi pozornosti so skladni z rezultati drugih študij [5], [14] z izjemo pravilnosti odgovorov, kjer ni prišlo do značilnih glavnih učinkov veljavnosti in opozorila, zato je rezultate potrebno interpretirati s previdnostjo. Na našem vzorcu smo v povprečju dobili nekoliko več napak, kar je verjetno posledica večje starostne raznolikosti našega vzorca v primerjavi s podobnimi študijami.

Rezultati iz simulatorja vožnje kažejo, da so udeleženci spremenili način vožnje glede na kontekst in zahtevnost scenarija vožnje, kar kaže na veljavnost manipulacije različnih pogojev. V povezavi z rezultati pozornosti sta bila večja pravilnost reševanja in krajši reakcijski čas povezana z manj vijuganja na cesti, kar kaže na večjo previdnost in bolj enakomerno vožnjo teh voznikov. Sklepamo lahko, da posamezniki, ki naredijo več napak na testu pozornosti, vozijo nekoliko počasneje, zanimivo pa je, da pravilnost reševanja ni bila povezana z večjo prekoračitvijo omejitve hitrosti.

Konsistentno z drugimi ugotovitvami o učinkovitosti orientacije pozornosti [5], se je tudi v naši raziskavi učinkovitost orientacije pokazal kot najbolj konsistenten napovednik indikatorjev vožnje, tudi v povezavi s kompleksnostjo scenarijev. Glede na to, da je orientacijska učinkovitost povezana s selekcijo informacij iz okolja z usmerjanjem pozornosti na potencialno relevantne situacije [14], lahko sklepamo, da je ključnega pomena pri usmerjanju pozornosti na potencialne nevarnosti pri vožnji v realnem okolju s prilagajanjem vedenja. Po drugi strani lahko imajo ti posamezniki nekaj težav v situacijah, kjer je več potencialno nevarnih situacij in je manj verjetno, da bo katera dejansko nevarna (npr. več vozil se bliža križišču, pa samo eno pozabi prižgati smernik) [5].

Raziskava je bila izvedena v simulatorju vožnje, zato ugotovitev ne moremo direktno prenesti v realno okolje, še posebej, ker je 50% udeležencev odgovorilo, da so vozili drugače kot običajno. Prav tako vrstni red scenarijev ni bil popolnoma naključen in bi lahko prišlo do učinkov učenja. Naslednja omejitev je vezana na odvisne spremenljivke simulatorja. Čeprav smo dobili napovednike indikatorjev vožnje, pa nam to bolj malo pove o dejanski varni vožnji voznikov. V drugih študijah so običajno opazovali število prekrškov ali vedenje neposredno pred izpostavljenostjo konkretnim vedenjskim in okolijskim nevarnostim ali nevarnostim, ki zahtevajo deljeno pozornost, npr. približujoč se avtomobil z nasprotne smeri ob prehitevanju [5]. V prihodnjih študijah je potrebno vključiti večjo raznolikost kognitivnih testov ter po vzoru drugih študij proučevati vožnjo povezano s specifičnimi nevarnimi situacijami v prometu [5] ali z odbitki točk za nevarno vožnjo [4], da bi dobili boljši vpogled v kognitivne funkcije varnih oz. nevarnih voznikov.

Zahvala

Raziskava je bila podprta s strani ARRS v okviru projekta L2-8178. Avtorji se zahvaljujejo podjetju NERVteh za zagotovitev simulatorja vožnje, ter Kristini Stojmenovi in Klemnu Novaku za tehnično podporo pri zbiranju podatkov.

Literatura

[1] J. M. Wood, "Age and Visual Impairment Decrease Driving Performance as Measured on a Closed-Road Circuit," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 44, no. 3, pp. 482–494, 2002.

[2] H. Iversen and T. Rundmo, "Attitudes towards traffic safety, driving behaviour and accident involvement among the Norwegian public," *Ergonomics*, vol. 47, no. 5, pp. 555–572, 2004.

[3] M. Šucha and D. Černochová, "Driver Personality as a Valid Predictor of Risky Driving," in *Transportation Research Procedia*, 2016, vol. 14, pp. 4286–4295.

[4] U. León-Domínguez, I. Solís-Marcos, E. Barrio-Álvarez, J. M. Barroso y Martín, and J. León-Carrión, "Safe driving and executive functions in healthy middle-aged drivers," *Appl. Neuropsychol.*, vol. 24, no. 5, pp. 395–403, 2017.

[5] J. Roca, D. Crundall, S. Moreno-Ríos, C. Castro, and J. Lupiáñez, "The influence of differences in the functioning of the neurocognitive attentional networks on drivers' performance," *Accid. Anal. Prev.*, vol. 50, pp. 1193–1206, 2013.

[6] T. Mäntylä, M. J. Karlsson, and M. Marklund, "Executive control functions in simulated driving," *Appl. Neuropsychol.*, vol. 16, no. 1, pp. 11–18, 2009.

[7] S. G. Klauer *et al.*, "The Impact of Driver Inattention On Near Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data," *Analysis*, no. April, p. 226, 2006.

[8] B. Weaver, M. Bédard, J. McAuliffe, and M. Parkkari, "Using the Attention Network Test to predict driving test scores," *Accid. Anal. Prev.*, vol. 41, no. 1, pp. 76–83, 2009.

[9] J. Fan, B. D. McCandliss, T. Sommer, A. Raz, and M. I. Posner, "Testing the efficiency and independence of attentional networks," *J. Cogn. Neurosci.*, vol. 14, no. 3, pp. 340–347, 2002.

[10] M. I. Posner and S. E. Petersen, "The attention system of the human brain.," *Annu. Rev. Neurosci.*, vol. 13, pp. 25–42, 1990.

[11] S. E. Petersen and M. I. Posner, "The Attention System of the Human Brain: 20 Years After," *Annu. Rev. Neurosci.*, vol. 35, no. 1, pp. 73–89, 2012.

[12] A. Helland, S. Lydersen, L. E. Lervåg, G. D. Jenssen, J. Mørland, and L. Slørdal, "Driving simulator sickness: Impact on driving performance, influence of blood alcohol concentration, and effect of repeated simulator exposures," *Accid. Anal. Prev.*, vol. 94, pp. 180–187, 2016.

[13] J. O. Brooks *et al.*, "Simulator sickness during driving simulation studies," *Accid. Anal. Prev.*, vol. 42, no. 3, pp. 788–796, 2010.

[14] J. Roca, C. Castro, M. F. López-Ramón, and J. Lupiáñez, "Measuring vigilance while assessing the functioning of the three attentional networks: The ANTI-Vigilance task," *J. Neurosci. Methods*, vol. 198, no. 2, pp. 312–324, 2011.

[15] "Psychology Software Tools, Inc. [E-Prime 2.0]," 2017. [Online]. Available: http://www.pstnet.com.

[16] "Nervteh Simulation Technologies," 2018. [Online]. Available: https://www.nerv-teh.com/.

MONAURAL AND BINAURAL LISTENING TO INFRASOUND; CAN PHYSIOLOGY HELP?

Matej Tomc, Aleš Magajne, Gregor Geršak, Samo Beguš

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: mt8907@student.uni-lj.si

V prispevku predstavljamo študijo ugotavljanja vpliva načina poslušanja zvokov nizkih frekvenc z enim ali z obema ušesoma. V gluhi sobi so zdravi prostovoljci poslušali zvoke frekvenc okoli spodnjega slišnega praga na oba načina. Ob tem smo prostovoljcem dodatno merili fiziološke parametre (prevodnost kože in srčni utrip) in ugotavljali, ali bi lahko s preprosto fiziologijo sklepali o nadležnosti zvokov določenih frekvenc. Naši rezultati ne kažejo pomembne povezave med fiziologijo in načinom poslušanja.

Keywords—infrasound, monaural, binaural, EARS2, perception

1 Introduction

1.1 Infrasound

The term "infrasound" represents sounds in the very low frequency range. Normally 20 Hz is considered the lower human hearing threshold, with some variance between individuals, with sounds below that frequency being considered inaudible, hence the word infrasound ("infra"= under). This is however not true, as sounds down to a few hertz can be audible, but the amplitudes of those sounds must be much higher than in normal hearing range (20 Hz < frequency < 20 kHz). Within infrasound frequency range, the dynamic range of auditory system is known to diminish rapidly with decreasing frequency. This results in compressed equal-loudness-level contours in a certain frequency range, and it implies that a slight increase in sound pressure level can change the perceived loudness from barely audible to loud [1]. In the infrasound range, the perception of sound also changes. Sound loses its tonal properties and begins to feel discontinuous [2].
In our everyday life the infrasound is commonly generated as a side product of activities like ventilation, air conditioning, traffic, aviation, wind turbines, etc as well as from natural sources such as thunder, wind, earthquakes, volcanic eruptions and sea waves [1]. The recent trends of industrialisation and mobilisation have resulted in increase in sound pollution in all frequency ranges, which is affecting quality of our everyday lives and is also a potential health hazard [4-10]. Low-frequency noise, including infrasound, specifically poses a significant threat, as it carries more energy compared to other frequencies while also being able to traverse larger distances and loosing less amplitude when passing through buildings. Normal hearing protection was also proven to be less effective in low-frequency range [1].

1.2 Ears II project

Project Ears II (Metrology for modern hearing assessment and protecting public health from emerging noise sources) is a part of EMPIR (European Metrology Programme for Innovation and Research) initiative, which is a programme co-founded by the European Union's Horizon 2020 and EMPIR participating states. Ears II is to be run from 2016 to 2019 as a follow-up project to the successful Ears I project.
One of the aspects of Ears II project is tied to hearing conservation and explores the response of humans to infrasound and ultrasound as well as potential health hazards of aforementioned sounds [3]. In order to properly limit potential health risks emerging from infrasound and ultrasound, a deeper understanding of human hearing and sound perception is needed. With this sounds only being audible at extremely high amplitudes, an interdisciplinary approach is mandatory, combining audiology with neuroscience and psychology. New methods for determining the effects of these sounds had to be developed. The other studies of this project are already making use of technologies such as EEG and fMRI in combination with traditional audiological methods.

1.3 Monaural versus binaural hearing

Recent studies in this field [11-14] administered the sound to test persons monaurally, via an insert earphone. While there is definitely a difference between the whole body and only ears being exposed to the sound pressure, studies suggest that ears are still the primary sensory organ for infrasound [2]. There is however no consensus when it comes to "binaural advantage". The term refers to a common idea that binaural thresholds are slightly lower, typically around 3 dB. While it is widely accepted in terms of frequencies above 200 Hz, some studies have observed it in low-frequency and infrasound frequency range as well [15, 16], while others found no binaural advantage in this frequency range [13, 17].

1.4 The aim of this study

The goal of this study was to research the possibility of using objective physiological parameters to assess the psychological effects of infrasound on human subject. The level of psychological arousal, measured in form of activity of subject's electrodermal activity (EDA) and heart rate (HR), was evaluated within two tasks: i) listening to infrasound by means of monaural and ii) by means of binaural hearing.

2 Measuring the effects of infrasound in monaural and binaural measuring systems

2.1 Measuring set-up

For the purpose of this experiment, an anechoic chamber (within Faculty of Electrical Engineering University of Ljubljana) was used [18]. The loudspeaker used for the experiment was custom made and constructed specifically for this project by Physikalisch-technische Bundesanstalt, Braunschweig, Germany [13]. The sound source produces very low harmonic distortion in 2-250 Hz range.

Atop the loudspeaker, a polyethylene hose is mounted to direct the flow of air pressure. After a few decimetres, the hose is split in two by a Y-splitter. The two tubes go through a valve each and are connected to the headset of a common stethoscope. The stethoscope had its tubing removed and is only used as a means to effectively connect the silicon hoses with the ear canal. The valves are manual and are controlled by an Arduino controlled servo motor, so they can be closed slowly. This prevents the problem of audible pops, that would otherwise occur when opening and shutting solenoid valves.

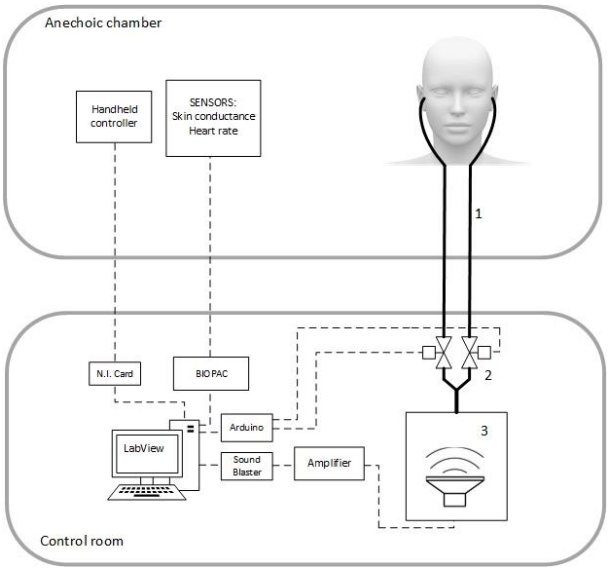


Figure 1. Measuring set-up: anechoic chamber, loudspeaker, stethoscope, valves, Biopac, computer, handheld controller

Prior the study hoses of different diameters, lengths and stiffness were tested. In the end a pair of harder hoses were chosen in order to conduct the least amount of noise from outside the anechoic chamber and overall best frequency response compared to the much softer silicone hose or the much wider polyvinyl chloride hose. The chosen hoses were about eight meters long with inner diameter of 4 mm and outer diameter of 6 mm.

To measure the physiological response of the research participants, Biopac MP150 module was used as a data acquisition device. Modules EDA100C and PPG100C were used to monitor skin conductance and heart-rate, respectively. Wet reusable Ag-AgCl electrodes for EDA and photoplethysmografic transducer were used as sensors. To ensure time synchronisation one of MP150 analogue channels was used as a triggering channel and was connected to the National Instruments DAQ 2344XX module, which was operated by the main computer.

The test person was given a handheld controller with a pushbutton and a potentiometer, to give them control over the amplitude in the first part of the experiment where the perceived threshold for each and both ears is found.

2.2 The test procedure

The experiment consists of three parts. The first part is finding the amplitude threshold at 125 Hz. This frequency was chosen, because it is a standard test frequency and accords to ISO 389-2 standard. The test subject is prompted to readjust the amplitude of the signal until they are certain that they have found their threshold. Test person is instructed to find their threshold for both ears once and then for each one ear once. Then the whole cycle is repeated two more times, resulting in three threshold measurements for each ear and both ears. Which ear the test person is having tested at the given time is predetermined in the program that shuts the correct valve, preventing the airflow from the loudspeaker and consequentially cutting of the stimuli for the respective ear. At the end of the first part of the test, a better ear is chosen, based on the criteria of having a lower average threshold at the given frequency. The third part is the same as the first. This allows us to observe any threshold shifts between the start and end of the experiment to ensure no hearing loss occurred during experiment.

The second part of the test consists of listening to a series of groups sounds. For the purpose of this paper, the term a “group of sounds” is used to refer to a specific composition of sound and silence that will be discussed further bellow.

In between the two parts, 90 seconds of inactivity was included to allow the test person’s physiology to stabilise and baseline physiology values to be acquired. 30 seconds before every group of test is played, a valve is turned, which causes a quiet humming noise. The humming noise only lasts a second, so the 30 second period is enough for physiology to stabilise again. This 30 seconds time period also serves as a divider between two groups of sounds, minimising the effect of the previous groups of sounds on the physiology when the next group of sounds is played.

Every test person is subjected to hearing sounds at the following frequencies: 8 Hz (infrasound, far from normal frequency range), 16 Hz (infrasound near normal frequency range) and 125 Hz (in normal

frequency range, clearly audible) (Figure 2). Groups of sounds of all frequencies are played twice: once both ears receive the stimuli and once only the better ear does. The sequence of frequencies and usage of both/better ears is not randomised; rather a balanced Latin square design is used. The same frequency is always played in pairs of monaural and binaural, but the order of frequencies and the order position within pairs changes with each test person.

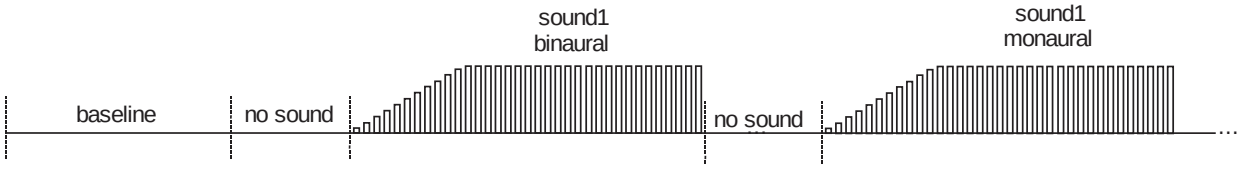


Figure 2. The measuring protocol (only first two played sounds are presented, in total every test subject listened to 6 sounds, i.e. 3 sounds (125 Hz, 16 Hz and 8 Hz), each twice (bi- and monaural)). Baseline period was 90 seconds, periods of silence 30 seconds and periods of played sounds 30 seconds long. In total, after 90 seconds of baseline relaxation period every subject was exposed to 6 randomized 30 seconds sessions of sounds (frequencies of 125 Hz, 16 Hz and 8 Hz).

The aforementioned groups of sounds consist of pure sinusoidal tones, which vary in amplitude following the envelope shown in Figure 2. We noted that raising the amplitude from zero to the desired value too fast could cause audible pops due to the rapid change in air pressure level in the tubes. Therefore, linear ramps were added to the begging and the end of the envelope of each pulse (Figure 2). The first few pulses in the series are also rising in amplitude. The middle part with the highest amplitude is the part that should be audible, while the other pulses provide smooth transition over the hearing threshold. This method was chosen because it allows us to monitor physiological parameters when the hearing threshold is crossed, while also minimizing the possibility of occurrence of physiology changes because of a shock due to suddenly hearing of a sound at a reasonably high amplitude.

After the experiment, the subjects were asked to describe their experiences, including possible discomfort caused by the anechoic chamber, boredom and any of the stimuli. This was done in a non-formal free interview.

2.3 Software implementation

The main program runs the National Instruments LabVIEW software. The program was designed in three parts, following the experiment protocol. At the start of the program, a trigger signal is sent to BIOPAC MP 150 to synchronise the two computers. LabVIEW is also used to control the Arduino, responsible for opening and closing of the valves. The first part of the program is designed to pick the test subject's better ear through comparing the thresholds that they set for themselves. The second part consists of playing groups of sounds as explained in the section 3.2. The start time of each group is recorded and written into a text file. All the start times, amplitudes and frequencies are set in advance, so there is no need to record this data.

To gather data from BIOPAC MP 150, AcqKnowledge 4.1 was used, where the bulk of data analysis also took place. Using the built-in functions the skin conductance responses (SCR) were identified, counted and their amplitudes measured. In addition, skin conductance level (SCL) within the baseline period and each interval was calculated. Average heart rate and its standard deviation was calculated.

2.4 Test persons

Fifteen healthy volunteers (students, 5 women and 10 men) participated in the study. Before the experiment, the protocol was briefly explained to them and they signed the written consent to participate in the study allowing them to leave the experiment if chosen.

2.5 Data analysis

Acquired physiological parameters were pre-processed by removing the outliers and filtering for the moving artefacts, which were manually identified. Descriptive statistics (mean value and standard deviation) od skin conductance level (SCL) and heart rate (HR) and number and amplitude of skin conductance responses (SCR) for the 90 seconds baseline and duration of each sound exposure were calculated. Number of SCR pulses (pulses in skin conductance signal, with amplitudes exceeding 0.02 uS and representing a measure of subject's momentary arousal) and sum of the mean SCR pulse amplitudes were calculated.

Normality was checked using Shapiro-Wilk test and the independent samples equal variances T test was used to compare effects of the monaural and binaural hearing to the subject's physiology. Statistical significance was assumed at $p \leq 0.05$.

3 Results

The physiological data of the subject was collected. In 30 % of the signals the heart rate values and especially heart rate variability parameters, such as standard deviation of heart rate, was corrupted and discarded from the following processing due to high content of moving artefacts caused by moving the hands when operating the handheld controller.

In general, the following was expected. An increase in SCL values and standard deviation of SCL is associated with increase in psychological arousal of the subject. I.e. larger SCL values with monaural hearing would indicate higher annoyance/discomfort level as

compared to binaural hearing of the same frequency sound. Similarly, the number and amplitude of SCR pulses increases with arousal. An increase in value of HR and standard deviation of HR (i.e. heart rate variability) would indicate the arousal and hence higher annoyance level of the subject.

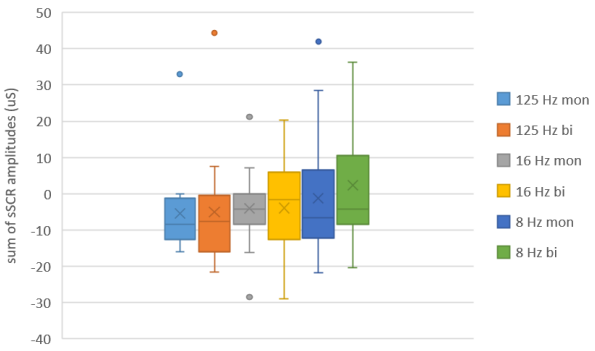


Figure 3. Comparison of sums of SCR amplitudes for different frequencies of sounds and different types of hearning (monaural, binaural).

Table 1. Comparison of physiology during binaural and monaural hearing

Physiology	Sound frequency	t value	df	p
SCL	125 Hz	1,186038	28	0,245574
	16 Hz	0,187703	28	0,852463
	8 Hz	0,991527	28	0,32992
stdSCl	125 Hz	0,830926	28	0,413045
	16 Hz	1,189663	28	0,244169
	8 Hz	1,237406	28	0,226213
SCR	125 Hz	0,213857	28	0,832208
	16 Hz	1,187777	28	0,444169
	8 Hz	0,855287	28	0,399652
SCRampl	125 Hz	0,080528	28	0,93639
	16 Hz	0,023826	28	0,98116
	8 Hz	0,586761	28	0,562066
avgHR	125 Hz	0,479043	26	0,635913
	16 Hz	0,159985	26	0,874129
	8 Hz	0,4146	26	0,681835
stdHR	125 Hz	0,084415	28	0,933327
	16 Hz	0,530837	28	0,599718
	8 Hz	0,133207	28	0,894983

In table 1 the comparison of physiology during binaural and monaural hearing of 30 seconds sounds of different frequencies (below infrasound threshold (8 Hz), threshold (16 Hz), audio range (125 Hz)). No statistically significant differences were found for any of the physiology parameters.

4 Discussion

This study investigated whether simple physiology, electrodermal activity (SCL, SCR) and heart activity (heart rate), could be suitable measures for assessing the physiological effects of infrasound on human subjects. We compared binaural and monaural hearing of sounds near infrasound hearing thresholds (well above (125 Hz), approximately at (16 Hz) and below (8 Hz) threshold). Our results showed no difference between monaural and binaural hearing of these sounds.

5 References

[1] Moller H, Pedersen C S. Hearing at low and infrasonic frequencies. Noise Health 2004;6:37-57

[2] Yeowart N.S., Bryan M.E., Tempest W. (1967), The monaural M.A.P. threshold of hearing at frequencies from 1.5 to 100 c/s, J. Sound Vib., Vol. 6(3), 335-342

[3] B. Berglund, P. Hassmén, in R. F. S. Job, „Sources and effects of low-frequency noise“, J. Acoust. Soc. Am., let. 99, št. 5, str. 2985–3002, maj 1996.

[4] Leventhall H G. Low frequency noise and annoyance. Noise Health 2004;6:59-72

[5] EEA European Environment Agency, Noise in Europe 2014, Copenhagen, 2014, ISBN 978-92-9213-505-8.

[6] 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.

[7] HPA 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/

[8] Salt, A. N. & Lichtenhan, J. T., How Does Wind Turbine Noise Affect People? Acoustics Today, 2014, 10, 20 – 27.

[9] Kaczmarska A., Łuczak A. »A Study of Annoyance Caused by LowFrequency Noise During Mental Work«, International Journal of Occupational Safety and Ergonomics (JOSE) 2007, Vol. 13, No. 2, 117–125

[10] Ears ll: The Project. [Na spletu]. Dostopno na: https://www.ears-project.eu/ears2-project.html. [Pridobljeno: 23-maj-2018].

[11] Weichenberger M, Bauer M, Kühler R, Hensel J, Forlim CG, Ihlenfeld A, et al. (2017) Altered cortical and subcortical connectivity due to infrasound administered near the hearing threshold – Evidence from fMRI. PLoS ONE 12(4): e0174420.

[12] Weichenberger M, Kühler R, Bauer M, Hensel J, Brühl R, Ihlenfeld A, Ittermann B, Gallinat J, Koch C, Sander T, Kühn S. Brief bursts of infrasound may improve cognitive function--An fMRI study Hearing research 328, 87-93 (2015)

[13] Robert Kühler, Thomas Fedtke and Johannes Hensel, Infrasonic and low-frequency insert earphone hearing threshold, J Acoust Soc Am 137(4), 2015

[14] Martin Bauer, Tilmann Sander-Thömmes, Albrecht Ihlenfeld, Simone Kühn, Robert Kühler and Christian Koch, Investigation of perception at infrasound frequencies by functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG), Congress on Sound and Vibration, ICSV 22 (2015)

[15] Yeowart N.S., Evans M.J. (1974), Thresholds of audibility for very low-frequency pure tones, J. Acoust. Soc. Am., Vol. 55(4), 814-818

[16] Anderson C.M.B., Whittle L.S. (1971), Physiological noise and the missing 6 dB, Acustica, Vol. 24, 261-272

[17] Sivian L.J., White S.D. (1933), On minimum audible sound fields, J. Acoust. Soc. Am., Vol. 4, 288

[18] Rosana Černelič, Merjenje akustičnih karakteristik gluhe sobe in snemalnega studia, diplomsko delo, Univerza v Ljubljani, Fakulteta za elektrotehniko, Ljubljana: 2015.

Izpostavitev ravninskih lipidnih dvoslojev električnemu polju nanosekundnih pulzev

Peter Kramar[1,2], Matej Reberšek[1], Damijan Miklavčič[1], P. Thomas Vernier[2]

[1]University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
[2]Frank Reidy Research Center for Bioelectrics, Old Dominion University, Norfolk (VA), USA
E-pošta: peter.kramar@fe.uni-lj.si

Planar lipid bilayers exposed to nanosecond pulsed electric field

Abstract. *Planar lipid bilayers are often used to study electroporation phenomena. In this study, preliminary experiments were carried out to measure changes in planar lipid bilayer resistance and capacitance resulting from exposure to nanosecond pulsed electric fields. We also looked at the effects of calcium ions on bilayer electrical properties. The results of the measured planar lipid bilayer resistance and capacitance obtained after exposure to nanosecond pulse rises several new questions about basic understanding of membrane electroporation mechanism under nanosecond electric field, which requires additional work to be performed.*

1 Uvod

Izpostavitev bioloških celic električnemu polju se je izkazala kot uporabna tehnologija za manipulacijo s prepustnostjo celične membrane ter s tem potencialno široko uporabo v številnih biomedicinskih in biotehnoloških aplikacijah ter tudi v predelavi hrane [1,2]. Že kratkotrajna izpostavitev električnemu polju povzroči strukturne spremembe na nivoju lipidnih molekul v celični membrani. V odvisnosti od parametrov električnega polja (amplituda, trajanje, frekvenca) postane membrana začasno prepustna, tudi po zaključeni izpostavitvi električnemu polju, za molekule, ki običajno ne morejo prehajati celične membrane. Pojav imenujemo elektroporacija ali elektropermeabilizacija [3-5].

Kljub temu, da je dandanes elektroporacija široko uporabljena metoda v različnih biomedicinskih in biotehnoloških aplikacijah [1,2], vključno s predelavo hrane [6,7], so nekateri osnovni fizikalni in biološki ter molekularni mehanizmi prenosa snovi pri elektroporaciji še vedno niso v celoti razjasnjeni.

Elektroporacijo delimo na dve osnovni veji: mikrosekundno elektroporacijo ter nanosekundno elektroporacijo, pri čemer predpostavljamo, da so osnovni mehanizmi enaki. Pri mikrosekundni elektroporaciji uporabljamo električne pulze z električnim poljem manjšim od 1 MV/m ter mikrosekundnim trajanjem. Nanosekunda elektroporacija uporablja električno polje večje od 1 MV/m ter trajanjem manjše od 1 μs [8]. Večina poizkusov v *in vitro* ter *in vivo* pogojih in aplikacij elektroporacije je bila narejena v mikrosekundnem področju. V nanosekundnem področju pa so bili narejeni osnovni modeli elektroporacije z molekularno dinamiko poleg tega pa tudi poizkusi *in vitro* ter *in vivo* pogojih. Za premostitev razlik med nano in mikrosekundnim območjem in razlago dinamike dogodkov na celični membrani ob elektroporaciji, smo zasnovali poizkuse na ravninskih lipidnih dvoslojih tako, da smo jih izpostavili nanosekundnim pulzom.

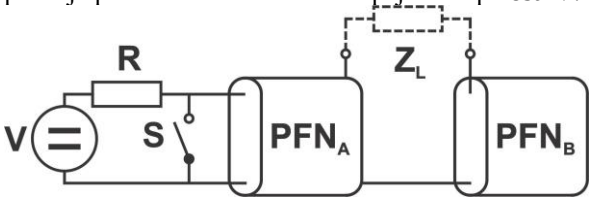
Ravninski lipidni dvosloj je umetna struktura lipidnih membran, ki po svoji zgradbi ponazarja infinitezimalno majhen košček celične membrane. Pogosto ga uporabljamo za preučevanje pojava elektroporacije tako na eksperimentalnem nivoju z merjenjem kapacitivnosti ter porušitvene napetosti membrane kot tudi na nivoju molekularne dinamike. Odlika ravninskega lipidnega dvosloja je njegova električna dostopnost z obeh strani in preprosta predstavitev z električnim vezjem vzporedno vezanega upora in kondenzatorja [9]. Na ravninskih lipidnih dvoslojih so v preteklosti že bili narejeni poizkusi za različne lipidne molekule, pri različnih temperaturah v odvisnosti od faznega prehoda lipidnih molekul in z uporabo različnih stimulacijskih signalov. Z meritvami kapacitivnosti posameznega ravninskega lipidnega dvosloja preverimo kvaliteto zgrajega ravninskega lipidnega dvosloja. Za preučevanje pojava elektroporacije je pomemben parameter porušitvena napetost, ki nam podaja stabilnost ravninskega lipidnega dvosloja v električnem polju.

Namen študije je bil izpostavitev ravninskega lipidnega dvosloja nanosekundim pulzem in pri tem preučevaje strukturnih spememb, kot so reorentacija lipidnih molekul, pore ter porušitvene napetosti. Ker poizkusov z nanosekundnimi pulzi v preteklosti na ravninskih lipidnih dvoslojih še ni bilo, je bil del študije razvoj unikatnega eksperimentalnega sistema za elektroporacijo ravninskih lipidnih dvoslojev z električnimi pulzi v območju nekaj deset voltov ter trajanjem pulzov nekaj nanosekund. Optično opazovanje ravninskega lipidnega dvosloja je zaradi narave komore, kjer zgradimo ravninski lipidni dvosloj nemogoče. Zato smo morali spremembe na ravninskem lipidnem dvosloju meriti električno s primernim napetostnim vzbujanjem ter merjenjem odziva toka. Merjenje toka v času nanosekundnega napetostnega vzbujanja je izredno zahtevno. Zato smo pred in po izpostavitvijo ravninskega lipidnega dvosloja nanosekundim pulzem merili kapacitivnost in upornost ravninskega lipidnega dvosloja z LCR metrom.

2 Materiali in metode

2.1 Nanosekundni pulzni generator

Tehnike generiranja nanosekundih pulzev se zelo razlikujejo od tistih za generiranje mikro- ali milisekundnih pulzov [10-14]. Vzrok so predvsem stikala, ki ne omogočajo izključitve v nanosekundnem časovnem območju. Za generiranje nanosekundih pulzev uporabljamo mrežo za oblikovanje pulza (ang. *pulse-forming networks* – PFN). Mreža za oblikovanje pulza je zgrajena iz večih podsekcij, kjer so posamezne sekcije sestavljene iz tuljav (*L*) in kondenzatorjev (*C*). Generator Blumlein (slika 1) tvori nanosekundni pulz s pomočjo dveh mrež za oblikovanje pulza. Blumlein generator deluje v dveh fazah, naelektritvi linij in razelektritvi. V času faze naelektritve mrež za oblikovanje pulza je stikalo (S) izključeno. Napetostni vir z nastavljivo amplitudo enosmerno napetostjo (V) nabije mrežo za oblikovanje pulza preko upora (R). V fazi razelektritve vključimo stikalo (S), tako razelektrimo naboj, shranjen v mreži za oblikovanje pulza preko impedance bremena (Z_L) oziroma ravninskega lipidnega dvosloja. Ker želimo generirati napetostni pulz brez odbojev, se mora impedanca Z_L čim bolj ujemati z impedanco mreže za oblikovanje pulza, Impedanco mreže za oblikovanje pulza določimo z enačbo $(L/C)^{0.5}$, kjer L in C predatavljata induktivnost tuljave in kapacitivnost kondenzatorja v posamezni podsekciji. Trajanje nanosekundnega pulza je $2N(LC)^{0.5}$, kjer je N število podsekcij mreže za oblikovanje pulza. Izhodna napetost nanosekundnega pulza je približno enaka izbrani napajalni napetosti V.



Slika 1. Shema generatorja Blumlein z dvema mrežama za oblkovanje pulza (PFN_A and PFN_B), nastavljivim napetosnim napajalnikom V, visoko impedančnim uporom R za nabijanje mreže za oblikovanje pulza, stikalom S in impedanco bremena – ravninskega lipidnega dvosloja Z_L.

Pri nanosekundni elektroporaciji je Blumline generator uporabljen praktično v vseh poizkusih za generiranje nanosekundnih pulzih. Njegova prednost je nizka cena, enostavnost izdelave, dobro kontrolo nad ponovljivostjo električnih parametrov. Kljub temu pa je pomembno, da je impedanca mreže za oblikovanje pulza prilagojena ravninskemu lipidnemu dvosloju, da preprečimo atenuacije amplitude pulza ter odboje nanosekundnega pulza. Pulzni generator Blumline, ki smo ga razvili za elektroporacijo ravninskih lipidnih dvoslojev, tvori nanosekundne pulze amplitud od 0 do 60 V in trajanja 7 ns.

2.2 Merjenje električnih lastnosti ravninskega lipidnega dvosloja

Merjenje strukturnih sprememb ravninskih lipidnih dvoslojev ob izpostavitvi nanosekundnim pulzem smo izvedli z LCR metrom Agilent E4980A, 20 Hz – 300 kHz, (ZDA). LCR meter deluje na principu mostične metode. Izbrana parametra merjenja sta bila paralelna upornost in kapacitivnos R_P-C_P. Amplituda merilne sinusne napetosti je bila izbrana 20 mV, frekvenca pa 2 kHz.

Merjenje impedance z LCR metrom in generiranje nanosekundnega pulza sočasno ni bilo mogoče. Zato smo razvili posebno analogno stikalo, ki je preklapljalo elektrode med nanosekundnim Blumeline generatorjem in LCR metrom. Prožilnik preklaplanja smo izvedli z digitalnim zakasnitvenim generatorjem DG645 (Stanford Research Systems, ZDA)

Na analogno stikalo smo priključili štiri Ag-AgCl elektrode, ki smo jih potopili v prevodno raztopino; dve na vsako stran ravninskega lipidnega dvosloja. Teflonska komora je sestavljena iz dveh kockastih rezervoarjev volumnov 5,3 cm^3. V povezovalno luknjico med rezervoarjema smo vstavili 25 μm tanko teflonsko folijo na sredi katere je bila luknjica s premerom 106 μm. Lipidne dvosloje smo tvorili z metodo dvigovanja gladine [15].

Celoten sistem smo nadzirali z uporabo programskega orodja Matlab 2017a ter dodatnim modulom Instrumentation Control Toolbox za nadzor merilnih instumentov preko protokola VISA.

2.3 Priprava ravninskih lipidnih dvoslojev

Prevodno raztopino smo pripravili iz 0,1 molarne KCl (Merck, Nemčija) in 0,01 molarne raztopine Hepes (Merck, Nemčija). Zmešali smo ju v enakem razmerju in dodali toliko kapljic 1 M NaOH (Merck, Nemčija), da smo dosegli pH raztopine 7,4.

Lipidne dvosloje smo pripravljali iz 1-palmitil 2-oleil 3-fosfatidilholin-a (POPC), ki smo jih kupili v prahu (Avanti Polar-Lipids, inc., ZDA). Lipid v prahu smo raztopili v mešanici heksana (Riedel-de Haën, Nemčija) in etanola (Riedel-de Haën, Nemčija) v razmerju 9:1. Za podporni torus smo uporabili mešanico heksadekana (MERCK, Nemčija) in etanola v razmerju 3:7.

Za preizkus zmožnosti detekcije strukturnih sprememb ravninskega lipidnega dvosloja smo pripravili dodatne poizkuse z dodajanjem dvovalentnih ionskih raztopin Ca^{2+} v okoliško prevodno raztopino. V komoro smo na eno stran ravninskega lipidnega dvosloja postopoma dodajali po 0.5 μL 1M $CaCl_2$ raztopine, kar je pomenilo, da smo molarnost $CaCl_2$ z vsakim vnosom povečali za 0,5 mM.

2.4 Merilni protokol

Potek poizkusov je temljil na cikličnem merilnem protokolu. Najprej smo v komori tvorili ravninski lipidni dvosloj, nato smo z LCR metrom merili kapacitivnost in upornost ravninskega lipidnega

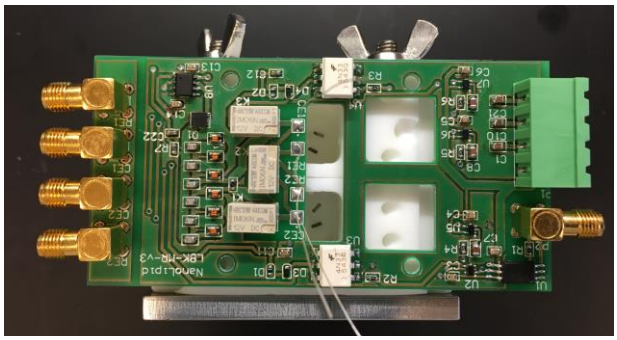
dvosloja. V primeru, da vrednosti le teh nista ustrezali ravninskemu lipidnemu dvosloju, smo postopek ponovili. Ob primernih vrednostih kapacitivnosti in upornosti ravninskega lipidnega dvosloja smo pričeli s postopkom merjenja in generiranja nanosekundnega pulza. Posamezna meritev na LCR metru je trajala vsaj 400 sekund. Med tem smo enkrat vbrizgali dvovalentno raztopino $CaCl_2$ in opazovali spremebo upornosti in kapacitivnosti ravninskega lipidnega dvosloja. Med posameznimi meritvami smo tudi večkrat prožili nanosekundni pulz različnih amplitud.

3 Rezultati

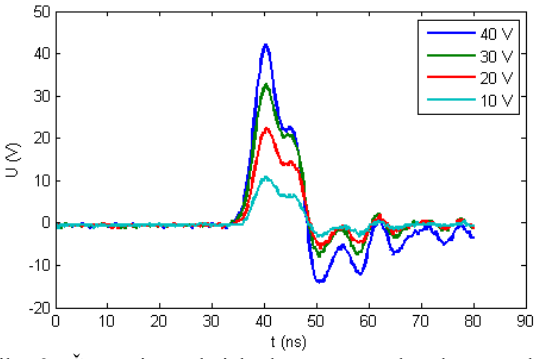
Študija je obsegala razvoj nanosekundnega pulznega generatorja ter izpostavitev ravninskih lipidnega dvosloja nanosekundim pulzom in pri tem preučevaje strukturnih spememb ter porušitvene napetosti z merjenjem upornosti in kapacitivnosti.

3.1 Generator nanosekundnih pulzov Blumeline

Komora za ravninski lipidni dvosloj z nizko prevodnostjo in visoko kapacitivnostjo predstavlja električno breme z visoko časovno konstanto. Za generiranje nanosekundnega pulza na bremenu z visoko časovno konstanto mora zato imeti pulzni generator nizko impedanco in mora biti električno blizu bremena. Blumlein generator z mrežo za oblikovanje pulza je eden izmed najprimernejših konceptov za generiranje pulzov na nizki impedanci, saj je njegovo vezje in električne komponente najlažje optimizirati, da dosežemo zahteve za izhodni pulz. Generator smo najprej optimizirali s programom PSPICE (Cadence, ZDA) [16] in ga nato še izdelali (slika 2). Obe mreži za oblikovanje pulza v Blumlein generatorju sta sestavljeni iz treh pod sklopov z induktivnostjo 6 nH in kapacitivnostjo 240 pF. Izhodni pulzi iz nanosekundnega Blumlein generatorja se dobro ujemajo s PSPICE simulacijo. Polna dolžina pulza pri polovični amplitudi (FWHM) je 7 ns, maksimalna amplituda pulzov pa je enaka napajalni napetosti (slika 3). Nepravokotnost pulzov in 20% negativni odboji so bili predvideni že pri PSPICE simulaciji in so posledica majhnega število pod sklopov mreže za oblikovanje pulza. To pa zato, da smo lahko realizirali kompaktno vezje na majhni komori za ravninski lipidni dvosloj.



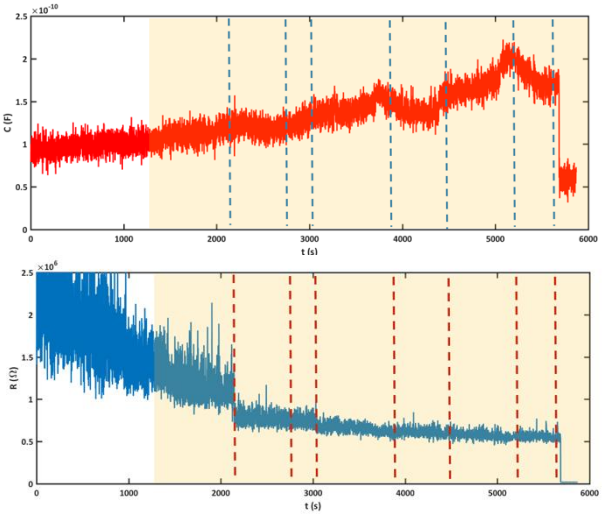
Slika 2. Vezje nanosekundnega Blumlein generatorja na komori za ravninski lipini dvosloj.



Slika 3. Časovni potek izhodnega nanosekundnega pulza iz Blumlein generatorja pri različnih napajalnih napetostih (10 - 40 V).

3.2 Ravninski lipidni dvosloji izpostavljeni nanosekundnim pulzom

Rezultati merjenja kapacitivnosti in upornosti ravninskega lipidnega dvosloja z uporabo LCR meta so pokazali, da nanosekundni pulzi na ravninski lipidni dvosloj nimajo vpliva. Pri dodajanju dvovalentne raztopine $CaCl_2$, opazimo povečanje kapacitivnosti in zmanjšanje upornosti ravninskega lipidnega dvosloja. V kombinaciji vzbujanja ravninskega lipidnega dvosloja in dodajanja $CaCl_2$ opazimo, da se upornost in kapacitivnost ravninskega lipidnega dvosloja spremeni le ob dodajanju $CaCl_2$ (slika 4). Ravninski lipidni dvosloj se zaradi vzbujanja z nanosekundnimi pulzi ni porušil, kot smo pričakovali. Ko je prevodna raztopina vsebovala 4 mM $CaCl_2$, se je ravninski lipidni dvosloj običajno porušil, zaradi dodajanja $CaCl_2$ v prevodno raztopino. Pri ponovnem tvorjenju ravninskega lipidnega dvosloja se lipidni dvosloj ni več vzpostavil. Poizkuse smo ponovili večkrat, na sliki 4 je prikazan eden od rezultatov.



Slika 4. Kapacitivnost in upornost ravninskega lipidnega dvosloja v odvisnosti od časa. Črtkane črte na obeh grafih predstavljajo dodajanje 0,5 mM $CaCl_2$ v solno raztopino na eno stran ravninskega lipidnega dvosloja. Sivo polje predstavlja vzujanje ravninskega lipidnega dvosloja z nanosekundnim pulzom amplitude 50 V in trajanja 7 ns. Nanosekundni pulzi so bili proženi na vsakih 50 s.

4 Diskusija

Ravninski lipidni dvosloj, kot preprost model celične membrane, do sedaj ni bil izpostavljen nanosekundnim elektroporacijskim pulzom. Cilj študije je bil razvoj nanosekundnega generatorja za izpostavitev ravninskih lipidnih dvoslojev nanosekundnim pulzom, da bi razložili razliko eksperimentov *in vitro* ter *in vivo* z uporabo nanosekundnih in mikrosekundnih puziov. V skladu z študijami v *in vitro* in *in vivo* pogojih ter simulacijami molekularne dinamike smo ob izpostavitvi nanosekundnim pulzom pričakovali strukturne spremembe oziroma porušitev ravninskega lipidnega dvosloja.

Opazovanje v času trajanja naosekundnega pulza je težko izvedljiva. Optično opazovanje z mikroskopom ni mogoče zaradi komore, kjer so pripravljeni ravninski lipidni dvosloji. Zato moramo spremembe ravninskega lipidnega dvosloja opazovati električno z merjenjem spremembe kapacitivnosti in upornosti. Pred in po vzpostavitvi ravninskega lipidnega dvosloja nanosekundim pulzom smo merili kapacitivnost in upornost ravninskega lipidnega dvosloja z LCR metrom.

Rezultati merjenja kapacitivnosti in upornosti so pokazali, da izpostavitev ravninskega lipidnega dvosloja nanosekundnim pulzom ne vpliva, opazne pa so spremembe ob dodajanju dvo-valentnih ionov v okoliško raztopino. To odpira nekatera nova vprašanja na nivoju osnovnega razumevanja mehanizmov elektroporacije na nivoju nanosekundih pulzev, ki zahtevajo dodatne raziskave.

Z dinamičnim merjenjem toka v času trajanja pulza bi lahko izmerili perturbacije lipidnih molekul v membrani. Vendar so lastnosti merilnega sistema, kot je pasovna širina v gigaherčnem frekvenčnem območju, hitra dinamika merilnika, merjenje toka v pikoamperih, prezahtevne. Izdelati ali kupiti bi bilo potrebno tudi novo mikro komoro za izgradnjo ravninskih lipidnih dvoslojev ter prilagoditi merilno vezje za visoko frekvenčno merjenje. Za primerjavo poizkusov med nanosekundnim območjem ter mikrosekundnim območjem bo v prihodnosti potrebno načrtovati poizkuse tako z napetostnim kot tudi tokovnim vzbujanjem s pulzi.

Zahvala

Raziskava je podprta s strani Javne agencije za raziskovalno dejavnost Republike Slovenije. Raziskave so bile opravljene v sklopu laboratorijev EBAM skupine evropskih laboratorijev (LEA). Del poizkusov je bil opravljen v sklopu programa Mreže infrastrukturnih centrov Univerze v Ljubljani (MRIC UL IP-0510). Del poizkusov je bil opravljen v sklopu štirimesečnega obiska v Frank Reidy Research Center for Bioelectrics, Old Dominion University, Norfolk, Virginia, ZDA, za katerega je doc. dr. Peter Kramar prejel Fulbrightovo štipendijo.

Literatura

[1] M.L. Yarmush *et al.*: Electroporation-based technologies for medicine: principles, applications, and challenges. Annu. Rev. Biomed. Eng. 16: 295-320, 2014.

[2] T. Kotnik *et al.*: Electroporation-based applications in biotechnology. Trends Biotechnol. 33: 480-488, 2015.

[3] E. Neumann *et al.*: Permeability changes induced by electric impulses in vesicular membranes. J. Membr. Biol. 10: 279–290, 1972.

[4] D. Miklavčič: Handbook of Electroporation. Springer 2017.

[5] T. Kotnik *et al.*: Cell membrane electroporation – Part 1: The phenomenon. IEEE Electr. Insul. M. 28(5): 14-23, 2012.

[6] S. Mahnič-Kalamiza *et al.*: Electroporation in food processing and biorefinery. J. Membrane Biol. 247: 1279-1304, 2014.

[7] E. Puértolas *et al.*: Improving mass transfer to soften tissues by pulsed electric fields: fundamentals and applications. Annu Rev Food Sci Technol 3: 263–82, 2012.

[8] T. Batista Napotnik *et al.*: Effects of high voltage nanosecond electric pulses on eukaryotic cells (in vitro): A systematic review. Bioelectrochemistry 110: 1-12, 2016.

[9] P. Kramar *et al.*: Merjenje lastnosti ravninskih lipidnih dvoslojev. Elektrotehniški vestnik 76: 293-298, 2009.

[10] M. Reberšek *et al.*: Cell membrane electroporation – Part 3: The equipment. IEEE Electr. Insul. M. 30(3): 8-18, 2014.

[11] J. F. Kolb *et al.*: Nanosecond pulsed electric field generators for the study of subcellular effects. Bioelectromagnetics, 27: 172–187, 2006.

[12] M. Reberšek *et al.*: Blumlein configuration for high-repetitionrate pulse generation of variable duration and polarity using synchronized switch control. IEEE Trans. Biomed. Eng., 56(11): 2642–2648, 2009.

[13] C. Merla *et al.*: A 10-Ω high-voltage nanosecond pulse generator. IEEE Trans. Microw. Theory Tech., 58(12): 4079–4085, 2010.

[14] J. F. Kolb: Generation of ultrashort pulse. In Advanced Electroporation Techniques in Biology in Medicine, A. G. Pakhomov, D. Miklavčič, and M. S. Markov, Ed. Boca Raton, FL: CRC Press, 2010, pp. 341–352.

[15] M. Montal *et al.*: Formation of Bimolecular Membranes from Lipid Monolayers and a Study of Their Electrical Properties. Proceedings of the National Academy of Sciences 69: 3561–3566, 1972.

[16] E. Puértolas *et al.*: Improving mass transfer to soften tissues by pulsed electric fields: fundamentals and applications. Annu Rev Food Sci Technol 3: 263–82, 2012.

Vpliv merjenja krvnega tlaka na prevodnost kože in anksioznost zaradi merjenja

Valentina Stanić, Gregor Geršak, Nina Gržinič Frelih
Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: gregor.gersak@fe.uni-lj.si

Effect of measuring blood pressure on skin conductivity and anxiety because of measurement

The paper describes the process of recognizing the anxiety of a person by measuring the change in the skin conductance due to measurement of blood pressure using a home non-invasive blood pressure (NIBP) monitor. Skin conductance is measured by means of electrodes attached to the left hand's index and middle finger of the measured person. The measurement also includes continuous blood pressure monitor and a home NIBP device, where the latter might be a potential cause for the anxiety of the measured person. Signals are captured and processed by software packet Acqknowledge. We obtained data on the actual events that caused the anxiety of the measured person through the statistical processing of captured signals. In addition to physiological measures, the subjects also filled-in psychological questionnaires for self-evaluation of the level of effect of causes of anxiety as well as their personality and tendency to anxiety. The paper presents the results, focusing on the analysis of the skin conductance signals, continuous blood pressure signals and psychological measures.

Key words: anxiety, skin conductance, blood pressure, psychophysiology

1 Uvod

Anksioznost označuje čustveni odziv posameznika, ki aktivira posameznika, da se izogne situaciji, ki je zanj ogrožajoča [1]. Na telesni ravni se odraža kot povečana aktivnost avtonomnega, natančneje simpatičnega, živčnega sistema, kar vodi do mišične napetosti, odziva dihalnega aparata ter spremembe srčnega utripa in aktivnosti kože. Oseba jo doživlja kot neprijeten občutek napetosti, ki je običajno neopredeljen in nespecifičen [1].

Anksioznost je pogosto prisotna v našem vsakdanjiku; služba, obveznosti, obisk zdravnika, opravljanje izpita, nepričakovani dogodki in podobno so potencialni povzročitelji anksioznosti. Pogosta in dolgotrajna obdobja anksioznosti močno vplivajo na psihološko in telesno zdravje osebe ter lahko vodijo v resna bolezenska stanja, ki jih je potrebno zdraviti. Ravno zaradi tega je potrebno anksioznost opazovati, meriti in ovrednotiti njen vpliv na človeka, s ciljem zmanjšanja ali celo preprečenja tovrstnih bolezenskih stanj.

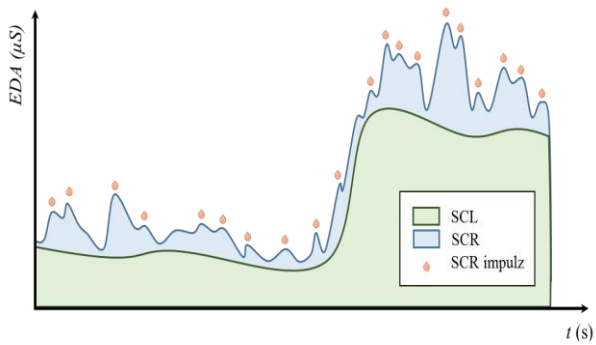
Prevodnost, temperatura kože in krvni tlak so ene izmed fizioloških funkcij, ki jih uravnava avtonomni živčni sistem. Njegovo delovanje je odvisno od psihofiziološkega stanja osebe. Bistvena lastnost avtonomnega živčevja je, da ga oseba ne more zavestno nadzirati, torej deluje neodvisno od volje osebe [2].

Psihofiziološko stanje osebe je močno spremenljivo, zato lahko ob različnih časovnih trenutkih in v različnih psiholoških stanjih merjeni fiziološki parametri močno variirajo. Običajno se zaradi tega v psihofizioloških meritvah uporabljajo spremembe fizioloških parametrov proti stabilnemu začetnemu stanju [3].

Anksioznost je korelirana s spremembo električne aktivnosti kože (angl. Electrodermal activity - EDA), ki je edina psihofiziološka veličina, na katero ne vpliva delovanje parasimpatičnega živčevja [2].

2 Metode

Električna aktivnost kože označuje vse električne lastnosti kože, v največji meri električni odziv na izločanje znoja [4]. Ena izmed lastnosti je prevodnost kože (angl. skin conductance – SC), katere vrednost se spreminja z aktivnostjo žlez znojnic, kar hkrati odraža tudi psihološko stanje osebe. EDA sestavljata dve komponenti, raven prevodnosti kože (angl. skin conductance level - SCL) in odziv prevodnosti kože (angl. skin conductance response - SCR). Vrednost SCL opisuje stopnjo psihološkega vznemirjenja merjenje osebe, SCR pa opisuje trenutno spremembo prevodnosti kože [5]. SCR impulzi ali specifični SCR odzivi (angl. specific skin conductance response – sSCR) so vrednosti signala prevodnosti kože, kjer amplituda presega določeno raven v nekem časovnem okvirju po stimulusu. Psihološka vznemirjenost je določena s številom SCR impulzov na minuto. Vse naštete vrednosti prevodnosti kože so prikazane na sliki 1.

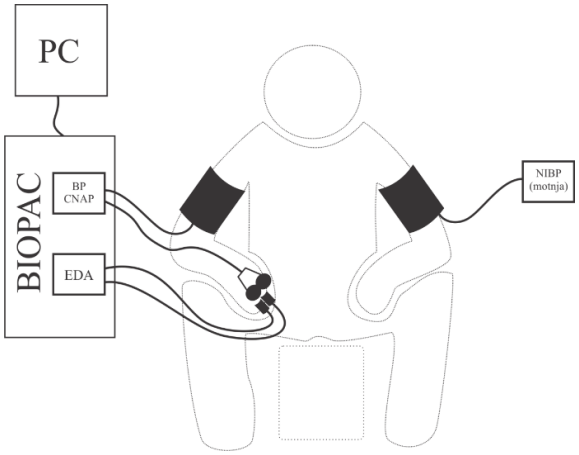


Slika 1: EDA se spreminja s časom in psihofiziološkim stanjem osebe. EDA je sestavljena iz SCL (označen z zeleno) in SCR (označen z modro) ter je izražena v μS. Z oranžnimi kapljicami so označeni SCR impulzi.

Prevodnost kože se običajno meri na roki, na zadnjih členkih dveh prstov, kjer je največja gostota žlez znojnic [6] [7] [5].

Krvni tlak je pritisk, ki ga vrši krožeča kri na stene žil. Uporabljata se dve metodi merjenja krvnega tlaka, invazivna in neinvazivna metoda (NIBP) [8]. Neinvazivna metoda temelji na merjenju fizikalnih pojavov, ki jih lahko zaznamo na površini telesa in so korelirani s krvnim tlakom, npr. zvok, premik stene žil ali kože. Invazivna metoda neposredno meri arterijski tlak, zato je neprimerna za široko uporabo (uporablja se izključno v kliničnih okoljih, predvsem na oddelkih intenzivne nege) [9].

Meritve smo izvajali v prostoru, v katerem je bilo zagotovljenih konstantnih 21 °C. Na merjeno osebo smo namestili tri merilne inštrumente: sistem za merjenje prevodnosti kože EDA100C (Biopac, ZDA), naprstni manšeti merilnika za zvezno merjenje krvnega tlaka NIBP100D (Biopac CNAP, ZDA) in validiran merilnik za domačo uporabo Omron M5, s katerim smo ob vnaprej predpisanih časovnih trenutkih sprožili merjenje krvnega tlaka in s tem povzročali motnje. Poleg tega je oseba rešila tudi določene psihološke teste. Na sliki 3 je prikazana shema namestitve merilne instrumentacije. CNAP je naprava, ki omogoča nenivazivno in zvezno merjenje sistoličnega in diastoličnega krvnega tlaka s podobno natančnostjo kot invazivne metode.



Slika 2: Merjena oseba z nameščenimi merilnimi instrumenti. CNAP in EDA pošiljata podatke na računalnik.

Motnjo oziroma vzrok za anksioznost merjene osebe generiramo z neinvazivnim merjenjem krvnega tlaka s pomočjo merilnika za domačo uporabo Omron M5.

Eden izmed uporabljenih psiholoških testov je lestvica merilne anksioznosti – LMA, ki je prirejena lestvica testne anksioznosti (angl. Test Anxiety Inventory – TAI) [11]. TAI je bila razvita na Oddelku za psihologijo Filozofske fakultete Univerze v Ljubljani za namene raziskav, vezanih na pregled pri zdravniku. LMA je prirejena tako, da jo lahko uporabimo pri merjenju psihofizioloških spremenljivk in preko tega ugotovimo doživljanje udeleženca v času meritev.

Za merjenje komponent osebnosti smo uporabili prostodostopne vprašalnike, ki merijo socialno anksioznost, umirjenost in samozavest. Za namen te raziskave smo jih prevedli v slovenski jezik [12].

V raziskavi je sodelovalo 25 oseb, 13 moških in 12 žensk, brez kardiovaskularnih zapletov. Osebe so bile stare med 23 in 63 let. Etičnost raziskave je bila preverjena s strani Komisije RS za medicinsko etiko (2012).

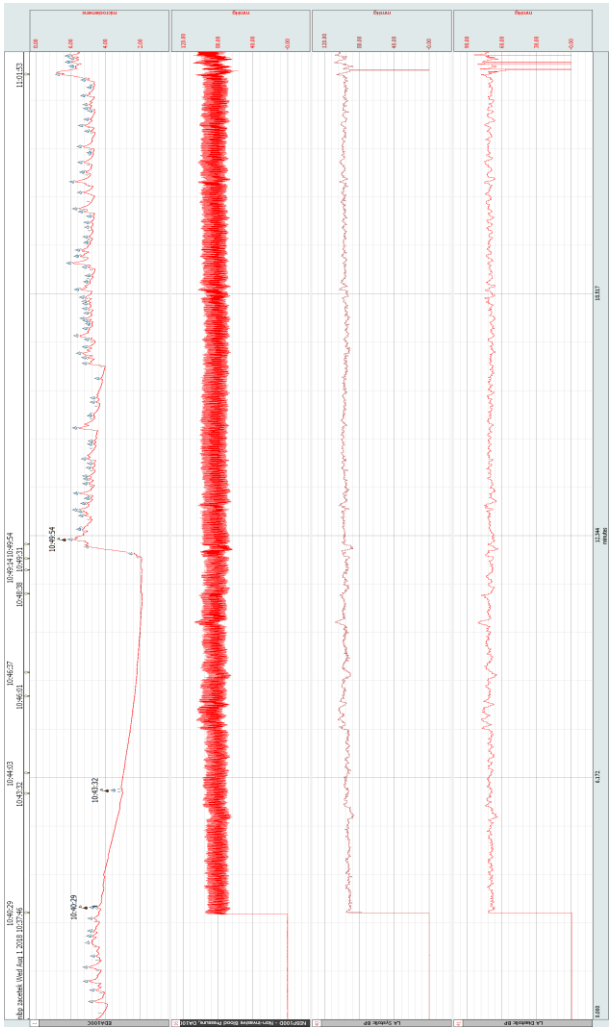
Pred izvajanjem meritve izvajalec meritve zapiše temperaturo prostora, pripravi izjavo in psihološke vprašalnike in kalibrira elektrode za merjenje prevodnosti kože. Naloga merjene osebe je, da se sprosti in umirjeno sedi na vnaprej pripravljenem stolu. Prva faza raziskave je faza sproščanja in umirjanja signalov. Oseba sedi in se sprošča tri minute, vmes se ne premika in ne govori. Po fazi sproščanja sledijo tri meritve krvnega tlaka s pomočjo NIBP (ki pravzaprav predstavljajo tri motnje, oziroma tri stimuluse za povečanje anksioznosti merjene osebe). Po vsaki meritvi NIBP sledita dve minuti premora. Po premoru, ki sledi tretji meritvi, izvajalec pove merjeni osebi, da je meritev zaključena ter odstrani merilnik krvnega tlaka NIBP. Na koncu merjena oseba izpolni pripravljene psihološke vprašalnike. Ko oseba zaključi z reševanjem vprašalnikov, izvajalec odstrani še preostale merilnike.

Izvajalec meritve na zajetih grafih sproti označuje dogodke, ki potencialno lahko povzročajo psihološko vznemirjenost merjene osebe, in sicer: začetek meritve, začetek 1. meritve z NIBP, konec 1. meritve z NIBP, začetek 2. meritve z NIBP, konec 2. meritve z NIBP, začetek 3. meritve z NIBP, konec 3. meritve z NIBP, trenutek, ko izvajalec pove, da je zaključen prvi del celotnega merilnega postopka, začetek reševanja psihološkega vprašalnika, konec reševanja psihološkega vprašalnika.

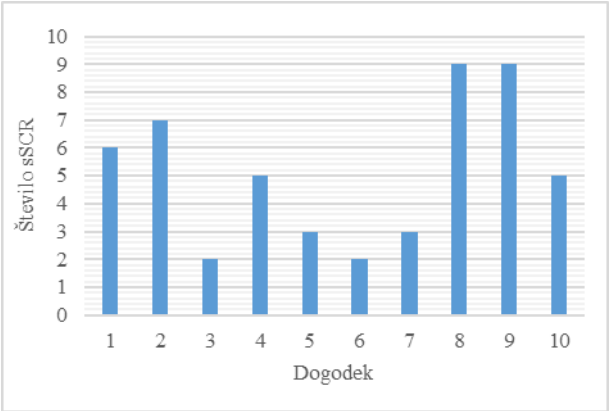
3 Rezultati

Rezultati kažejo, da sta bila najbolj obremenjujoča 8. in 9. dogodek, to je zaključek zadnje meritve in začetek reševanja psiholoških testov. Ta dva dogodka sta pri devetih osebah (36 %) povzročila odziv vrednosti prevodnosti kože. Začetek prve meritve z NIBP (2. dogodek) je povzročil odziv pri sedmih osebah (28 %). Osebe so bile najmanj reaktivne pri 3. in 6. dogodku, ki predstavljata konec prve meritve in začetek tretje meritve z NIBP. Rezultati za posamezne dogodke so prikazani na sliki 4.

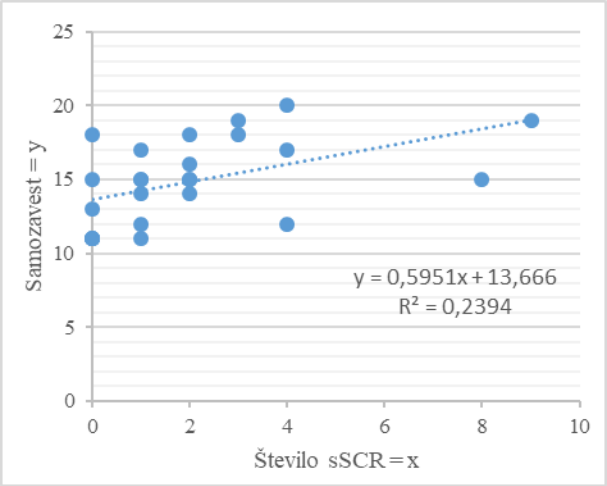
Rezultati reševanja psiholoških vprašalnikov kažejo, da so bile meritve z NIBP vsakič manj moteče, vendar so še vedno predstavljale motnjo. Rezultati kažejo tudi, da je bil najbolj moteč zvezni merilnik CNAP, kar je pričakovano, saj naprstne manšete stalno vršijo pritisk na enega izmed prstov, kar je zelo moteče. Ostala dva merilnika, EDA in NIBP, sta bila nekoliko manj moteča. Sama nastavitev merilnikov oseb ni pretirano motila.



Slika 3: Primer zajetih signalov ene merjene osebe. Prvi graf prikazuje EDA, drugi krvni tlak, tretji vrednost sistoličnega krvnega tlaka in četrti graf vrednost diastoličnega tlaka. Dogodki so označeni z žarnicami nad grafom.



Slika 4: Graf prikazuje število sSCR pri posameznem dogodku za vse merjene osebe. Največje število sSCR je pri 8. in 9. dogodku, torej gre za najbolj vznemirljive dogodke, najmanjše število sSCR pa je pri 3. in 6. dogodku, torej sta to najmanj vznemirljiva dogodka.



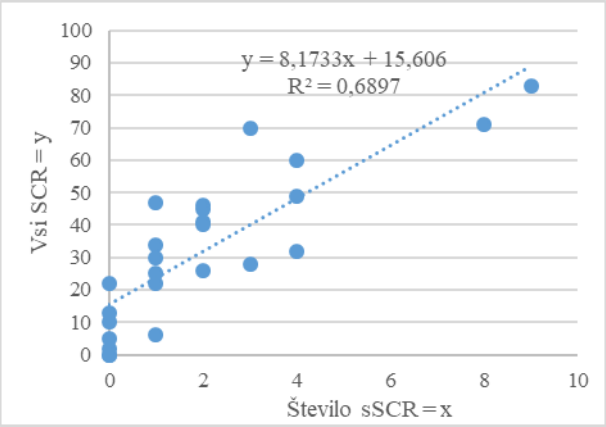
Slika 5: Na sliki je prikazan graf povezave samozavesti in števila sSCR. Višja je vrednost na ordinatni osi, manj je oseba samozavestna. Pri manj samozavestnih osebah je v povprečju več sSCR, kar pomeni, da so bolj psihološko reaktivne, bolj jih situacija vznemirja.

LMA je pokazala, da so merjene osebe v splošnem med meritvami bolj anksiozne, kot pri trenutno izvedeni meritvi, saj so druge meritve pomembnejše in imajo lahko negativne rezultate, kar povzroča največ skrbi. Osebe so lahko tudi značajsko bolj nagnjene k anksioznosti. Te osebe doživljajo anksioznost zelo pogosto in hkrati tudi bolj intenzivno [1]. Večina oseb je bila tudi bolj umirjene narave, to je večina jih je ostala mirna tudi pod pritiskom, hkrati pa so vsi označili, da jih v srednji ali večji meri prevzamejo čustva. Tretje področje testa osebnosti je pokazalo, da večina merjenih oseb ocenjuje, da so samozavestne. Večina jih je potrdila, da se cenijo in znajo v trenutku oceniti, kako se morajo odzvati. Na sliki 5 so predstavljeni rezultati samoocenjevanja samozavesti v korelaciji s številom sSCR dogodkov.

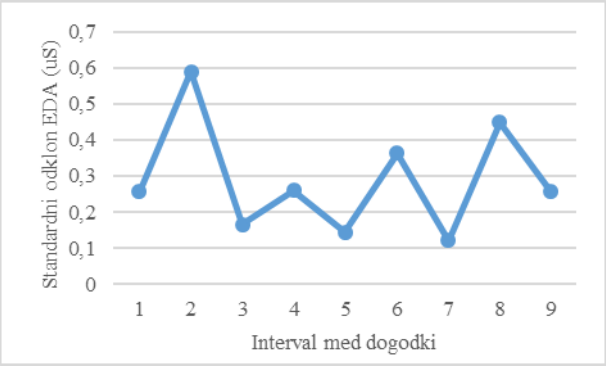
Pri vseh merjenih osebah smo med izvajanjem meritve zabeležili spremembo prevodnosti kože. Vsaka oseba je imela določeno vrednost SCR, pri čemer ni nujno, da se je pojavil tudi kakšen specifičen odziv sSCR. Graf vseh vrednosti SCR vseh merjenih oseb pri vseh dogodkih v odvisnosti od vseh specifičnih odzivov je prikazan na sliki 6. Višja je vrednost SCR, bolj je bila oseba anksiozna. Iz grafa je razvidno, da je osebe, ki so bile med meritvijo bolj anksiozne, bolj vznemiril tudi posamezen dogodek meritve.

Standardni odklon nosi informacijo o reaktivnosti osebe ob predvidenih dogodkih med meritvijo, zato so v nadaljevanju prikazani grafi standardnih odklonov v odvisnosti od intervalov med dogodki. Skupaj torej devet intervalov.

Na sliki 7 je prikazan primer, kjer je bila oseba močno reaktivna med prvo meritvijo NIBP ter izrazito reaktivna med drugo in tretjo meritvijo NIBP in med reševanjem psihološkega testa.



Slika 6: Graf vseh vrednosti SCR v odvisnosti od števila vseh sSCR. Višja je vrednost SCR in več je sSCR, bolj je bila oseba anksiozna med izvajanjem meritve.



Slika 7: Graf odvisnosti standardnega odklona EDA od intervalov med dogodki. Podatki veljajo za osebo 1.

Večina merjenih oseb (92%) je bila v prvem delu meritev nereaktivna ali zelo malo reaktivna (povprečno število sSCR je manj kot 5 na fazo), po koncu prvega dela meritev in v času reševanja psiholoških vprašalnikov pa so bili močno reaktivni.

4 Zaključek

Prispevek obravnava vpliv običajnega neinvazivnega merjenja krvnega tlaka na opazovano osebo, konkretno na njeno električno aktivnost kože in njeno anksioznost. Opisan je merilni protokol ter predstavljena uporabljena merilna instrumentacija (NIBP, CNAP, EDA) in psihološki testi (LMA, test osebnosti, lestvice vpliva motenj meritev). Pri študiji smo upoštevali različne dejavnike, ki lahko vplivajo na počutje merjene osebe: temperatura in hrup v prostoru, v katerem se izvaja meritev, primerno nameščena (ustrezno razporejena) merilna instrumentacija, čas umirjanja osebe in spol izvajalca meritev.

Pri vseh osebah je bila opazna sprememba aktivnosti prevodnosti kože. Pri vseh osebah je opazen dvig ravni prevodnosti kože v času reševanja psiholoških vprašalnikov. Bolj anksiozne osebe so imele večjo vrednost SCR, prav tako več SCR impulzov. Osebe, ki so bile bolj reaktivne tekom celotne meritve, so bile bolj reaktivne tudi ob posameznih dogodkih oz. fazah meritve.

Dinamika krvnega tlaka se je tekom meritve stalno spreminjala. Pri nekaterih osebah se je krvni tlak bolj spreminjal v času merjenja NIBP, pri drugih pa v času reševanja psiholoških testov. Pri večini oseb je bil diastolični tlak bolj reaktiven (večji standardni odklon) od sistoličnega.

Psihološki vprašalniki so razkrili, da je merjene osebe merjenje z merilnikom za domačo rabo (NIBP) zmotil, vendar ne toliko kot CNAP, ki je stalno meril krvni tlak. Razkrili so tudi osebnostne značilnice oseb, ki vplivajo na doživljanje anksioznosti (manj samozavestne osebe so bolj pogosto in bolj intenzivno doživljale anksioznost).

Zaključujemo, da neinvazivno merjenje krvnega tlaka z merilnikom za domačo rabo predstavlja motnjo, ki zviša aktivnost prevodnosti kože in hkrati tudi povzroči ali zviša anksioznost merjene osebe.

Literatura

[1] J. Norcross, G. VandenBos, and D. Freedheim, “APA handbook of clinical psychology: Theory and research (Vol. 4).,” APA Handb. Clin. Psychol. Theory Res., p. 572, 2016.

[2] J. J. Braithwaite, D. Derrick, G. Watson, R. Jones, and M. Rowe, “A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments.”

[3] G. Gersak, “Enostavni nizkocenovni merilniki prevodnosti koze,” Elektroteh. Vestnik/Electrotechnical Rev., vol. 80, no. 1–2, pp. 64–72, 2013.

[4] M. Benedek and C. Kaernbach, “A continuous measure of phasic electrodermal activity.,” J. Neurosci. Methods, vol. 190, no. 1, pp. 80–91, Jun. 2010.

[5] M. E. Dawson, A. M. Schell, and C. G. Courtney, “The Skin Conductance Response, Anticipation, and Decision-Making,” J. Neurosci. Psychol. Econ., vol. 4, no. 2, pp. 111–116, 2011.

[6] W. Boucsein et al., “Publication recommendations for electrodermal measurements,” Psychophysiology, vol. 49, no. 8, pp. 1017–1034, Aug. 2012.

[7] W. Boucsein, Electrodermal activity. Springer Science+Business Media, LLC, 2012.

[8] N. Gržinič, “Personalizirana medicinska instrumentacija - merilnik krvnega tlaka,” 2009.

[9] J. Drnovšek, J. Bojkovski, V. Batagelj, and S. Beguš, Avtomatizirani in virtualni merilni sistemi : laboratorijske vaje. Univerza v Ljubljani, Fakulteta za elektrotehniko, Laboratorij za metrologijo in kakovost, 2016, 2016.

[10] [“CNAP Monitor 500.” [Online]. Available: https://www.cnsystems.com/products/cnap-monitor-500. [Accessed: 05-Sep-2018].

[11] M. S. Ali and M. N. Mohsin, “Test Anxiety Inventory (TAI) : Factor Analysis and Psychometric Properties.” 2013.

[12] “International Personality Item Pool.” [Online]. Available: https://ipip.ori.org/.

Didaktika

Didactics

Uporaba krmilnika Siemens S7-314C-2PN/DP v programu elektrotehnik v srednjem strokovnem izobraževanju (SSI)

Borut Pogačnik

Srednja tehniška šola Šolski center Kranj, Kidričeva cesta 55, Kranj
E-pošta: borut.pogacnik@guest.arnes.si,

Use of the controller Siemens S7-314C in the program electrical technician of secondary vocational school system

Abstract. *The article describes the use of the controller Siemens S7-314C-2PN/DP in the program electrical technician of secondary vocational school system and presents a didactical approach to teaching programming of controllers within the curricular module Management with programmable devices. In the introduction, the controller Siemens S7-314C-2PN/DP is presented in the following steps: main features of a controlling system and an introduction of hardware and software. The first part of the article introduces some elementary steps of programming the controller Siemens S7-314C-2PN/DP based on the example of a simple combined control system. Two practical applications are presented in the second part of the article: controlling a traffic light with the controller Siemens S7-314C-2PN/DP and controlling a Festo distribution station with the controller Siemens S7-314C-2PN/DP.*

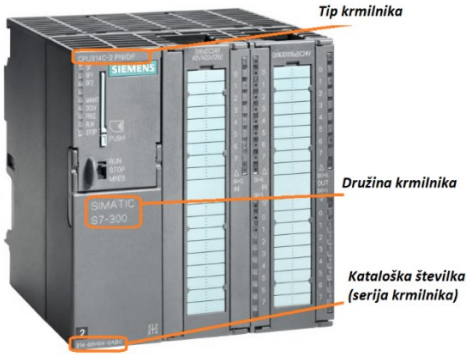
1 Uvod

V prispevku predstavimo sistematično strukturiran didaktični pristop k poučevanju programiranja programirljivih logičnih krmilnikov (PLK) v okviru modula Upravljanje s programirljivimi napravami (UPN) v programu elektrotehnik SSI. Učni cilji, katere morajo dijaki usvojiti pri posameznih modulih, so določeni v katalogih znanj, izbira didaktičnih metod pa je v največji meri prepuščena presoji učitelja, ki mora biti strokovno in didaktično usposobljen. Pri poučevanju v tehniških poklicih je za doseganje učnih ciljev zelo pomembno, da dijaki znajo posamezna teoretična in praktična znanja povezovati v uporabno celoto.

Motivacija dijakov za čim večjo učno aktivnost je odvisna od več faktorjev. Učitelj z načrtovanjem in uporabo ustreznih učnih oblik in metod postavi osnovne temelje. Pomembno pa je tudi, da zna strokovno učno vsebino posredovati zanimivo in aktualno. Dijaki naj bi bili čim bolj aktivni, pri tem pa je pomembna tudi uporaba moderne učne tehnologije in IKT. Učenje programiranja je zelo zahtevno, ker moramo pri dijakih razvijati nov način razmišljanja, ki pa je za precej dijakov v začetnih korakih težko. Pri vzpodbujanju učne aktivnosti dijakov je zelo pomembno, da začetne korake natančno načrtujemo in da spoštujemo didaktična načela (nazornost, postopnost, sistematičnost).

2 Predstavitev krmilnika Siemens S7-314C in TIA portala

Z metodo razlage dijakom predstavimo krmilnik Siemens S7-314C-PN/DP, in sicer v naslednjih korakih: osnovne lastnosti krmilnika, predstavitev strojne opreme in osnovni koraki priprave krmilnika za uporabo. V nadaljevanju dijakom predstavimo pregled programske opreme SIMATIC za delo s pomnilniško programirljivimi krmilniki, ki je zelo obsežna. Na šolskih računalnikih imamo nameščen TIA portal V13, zato smo ga tudi bolj natančno spoznali. TIA portal nam nudi pomoč pri kreiranju rešitev na področju avtomatizacije. Osnovni koraki pri konfiguraciji avtomatizacije so: kreiranje projekta, konfiguracija strojne opreme (hardware), povezovanje naprav (PC-PLK), programiranje PLK, konfiguracija vizualizacije, prenos konfiguracijskih podatkov in diagnostika delovanja. Portal TIA lahko uporabimo za konfiguracijo PLK-ja in za vizualizacijo sistema. Vsi podatki so shranjeni v enem projektu. Komponente za programiranje (STEP7) in vizualizacijo (EinCC) niso ločeni programi, temveč urejevalnik sistema dostopa do skupne podatkovne baze. Vsi podatki so shranjeni v skupni projektni datoteki.



Slika 1: Krmilnik Siemens S7-314C-2PN/DP

Družina SIMATIC S7-300 so krmilniki, ki se uporabljajo v srednje zahtevnih procesih in predstavljajo srednji cenovni razred krmilnikov za avtomatizacijo krmilnih in regulacijskih nalog strojev in procesov.

Sistem, ki ga uporabljamo pri praktičnem delu modula UPN, je sestavljen iz naslednjih

komponent: krmilnika Siemens SIMATIC S7-314C-2PN/DP, digitalnega vhodno/izhodnega modula (DI16/DI16x24VDC/0,5A), mehatronske enote Festo FSM205 razdeljevalna postaja in osebni računalnik z nameščenim programskim paketom TIA portal.

3 Spoznavanje programskih jezikov za programiranje PLK-jev

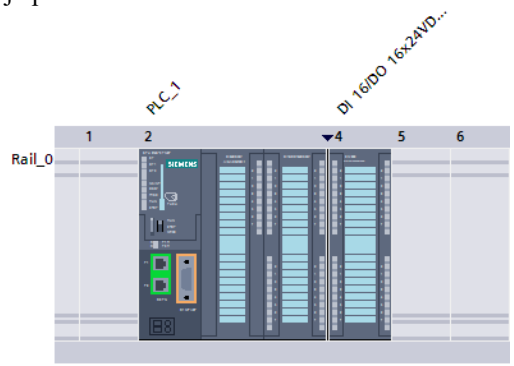
V okviru modula UPN je glavni cilj uporabe krmilnika, da dijaki spoznajo programske jezike, ki se uporabljajo za programiranje PLK. Standard IEC 61131-3 definira tri grafične programske jezike (lestvični diagram LD, funkcijski blokovni diagram FBD in sekvenčni funkcijski diagram SFC) in dva tekstovna programska jezika (inštrukcijska lista IL, strukturiran tekst ST) za kodiranje uporabniških programov PLK. Pri spoznavanju programskih jezikov smo uporabili laboratorijsko-eksperimentalno učno metodo, ki temelji na praktični aktivnosti dijakov, ki je različna. Na najnižji stopnji samostojnosti dijaki postopno spremljajo učiteljeva navodila. Nato sledi dijakovo samostojno delo. Učitelj vnaprej pojasni vse postopke. Na najvišji stopnji samostojno določijo postopke in samostojno delajo.

Modul UPN je razdeljen v dva vsebinska sklopa, in sicer v osnove programiranja in programirljiva krmilja. Da dijaki dosežejo cilje vsebinskega sklopa Programirljiva krmilja, morajo opraviti tri učne situacije, in sicer: 1. učna situacija je spoznavanje osnovnih korakov programiranja krmilnika (programirajo enostavno kombinacijsko krmilje v LD in FBD), 2. učna situacija je programiranje koračnega krmilja – semaforja in 3. učna situacija programiranje Festo-ve mehatronske postaje – razdeljevalna postaja. Modul UPN ima določeno število ur praktičnega pouka, ki je namenjen, da teoretična znanja uporabijo v praksi in vse skupaj povežejo v uporabno in aktualno celoto. Praktični del modula UPN poteka v specializirani učilnici za elektrotehniko, kje ima vsako delovno mesto osebni računalnik, krmilnik SIMATIC S7-314C in osem različnih Festo-vih mehatronskih postaj. Med teoretičnim poukom dijaki spoznajo zgradbo PLK, osnove programskih jezikov za programiranje PLK (LD, FBD, SFC) ter binarne senzorje in to znanje v okviru praktičnega pouka tudi ustrezno uporabijo in utrdijo.

3.1 Spoznavanje osnovnih korakov programiranja krmilnika

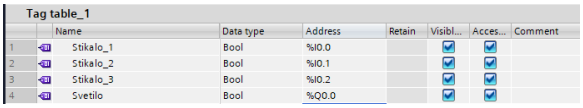
V okviru prve vaje krmiljenje svetila s tremi enopolnimi stikali smo spoznali osnovne korake programiranja krmilnika. V prostoru imamo tri enopolna stikala. Z vsakim stikalom lahko prižgemo ali ugasnemo svetilo. Za zahtevano kombinacijsko krmilje smo zapisali pravilnostno tabelo in logično enačbo.

V naslednjem koraku smo v TIA portalu kreirali nov projekt in ustrezno konfigurirali sistem (hardware). Po fizični povezavi krmilnika in računalnika, je potrebno v TIA portalu določiti, katere naprave bomo uporabljali v sklopu krmilnega sistema. Konfiguracija našega sistema je zajemala centralno procesno enoto (CPU 314C-2PN/DP) in digitalni vhodno/izhodni modul (DI16/DI16x24VDC/0,5A), kar je prikazano na sliki 2.



Slika 2: Konfiguriran krmilni sistem

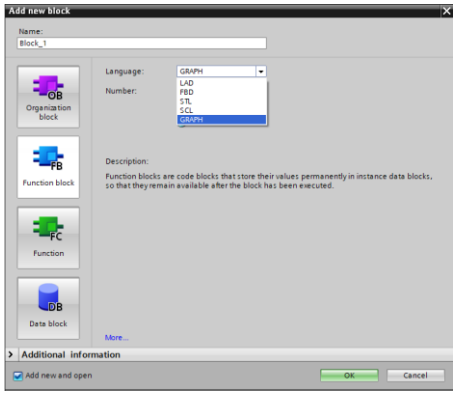
Ko ustvarimo nov projekt in izvedemo postopek konfiguracije naprav, lahko začnemo izdelovati uporabniški program. Pred začetkom izdelave uporabniškega programa moramo določiti prireditveno tabelo (PLK tags), v katero vpišemo imena vhodov, izhodov in pomnilniških celic, ki jih bomo uporabljali v uporabniškem programu, in jim dodelimo absolutne naslove, kar je prikazano na sliki 3.


Tag table_1

	Name	Data type	Address	Retain	Visibl...	Acces...	Comment
1	Stikalo_1	Bool	%I0.0		☑	☑	
2	Stikalo_2	Bool	%I0.1		☑	☑	
3	Stikalo_3	Bool	%I0.2		☑	☑	
4	Svetilo	Bool	%Q0.0		☑	☑	

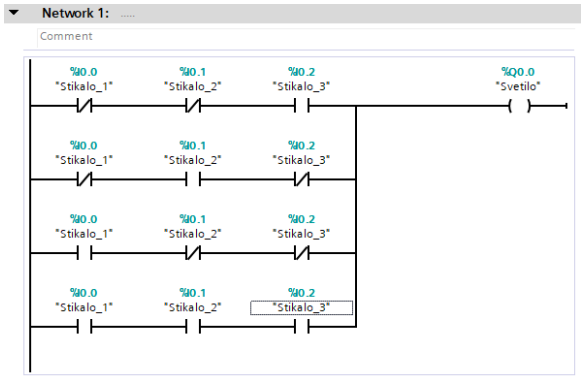
Slika 3: Prireditvena tabela (PLK tags)

Naslednji korak je ustvarjanje potrebnih programskih blokov. Osnovno teorijo o programskih blokih smo predstavili med teoretičnim poukom. Program se piše v blokih. V osnovi je v novem projektu dodan organizacijski blok Main{OB1}. Iz tega bloka so klicani vsi bloki, ki jih kasneje dodamo sami. Pri dodajanju novega funkcijskega programskega bloka, kar prikazuje slika 4, moramo izbrati tudi programski jezik, v katerem želimo izdelati uporabniški program.



Slika 4: Dodajanje blokov

Ko ustvarimo funkcijske bloke, se nam odpre okno, v katerem napišemo program, ki ga prikazuje slika 5.



Slika 5: Program za krmiljenje svetila v LD

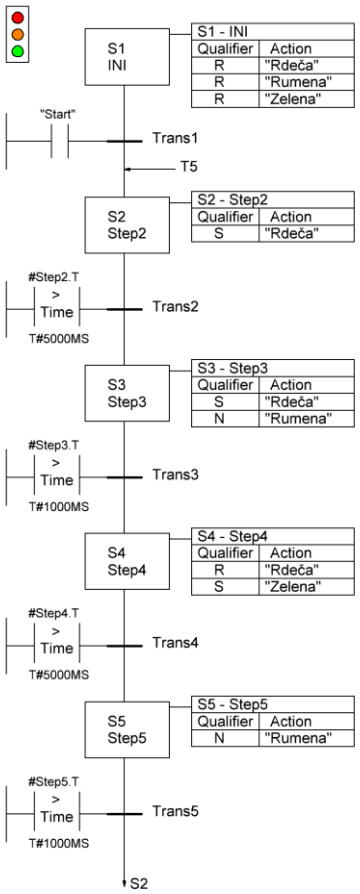
V zadnjem koraku uporabniški program prevedemo v obliko, ji jo lahko krmilnik prebere in prevedeni program prenesemo na krmilnik ter opravimo funkcijski preizkus.

V uvodnem delu dijakom demonstracijsko predstavimo posamezne korake, ki so potrebni pri programiranju krmilnikov. V nadaljevanju dijaki predstavljeno vajo samostojno opravijo brez večjih težav. Dijaki so prvo vajo krmiljenje svetila sprogramirali v dveh programskih jezikih , in sicer v lestvičnem diagramu in funkcijskem blokovnem diagramu. V analizi učne situacije smo ugotovili, da so vsi dijaki samostojno opravili vse zahtevane naloge.

3.2 Programiranje koračnega krmilja – semaforja

Dijaki so se pri teoretičnem pouku že seznanili z lastnostmi koračnih krmilij in osnovnimi koraki pri delovanju semaforja. V okviru druge učne situacije smo dijakom demonstracijsko predstavili programiranje semaforja v programskem jeziku GRAPH in jim ob tem podrobno razložili posamezne gradnike GRAPH-a. Pri tej učni situaciji dijaki postopno sledijo učiteljevim navodilom in zelo pomembno je, da posamezne postopke pri programiranju preverjamo sproti in s tem ugotavljamo, če so usvojili zahtevano učno vsebino. V primeru, da ne usvojijo temeljnih znaj programiranja, ima to lahko več negativnih posledic, kot so neuspešno spremljanje sledečih učnih vsebin, odpor in strah pred programiranjem, kar lahko povzroči tudi izostajanje od pouka in neuspešnost dijakov. Po končani predstavitvi, dijaki samostojno še enkrat sprogramirajo semafor in postavljajo vprašanja o morebitnih problemih pri programiranju.

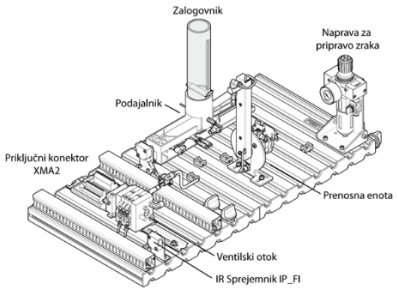
V zaključni analizi smo ugotovili, da je polovica dijakov samostojno opravila vse zahtevane naloge, ostali dijaki so potrebovali dodatno razlago in pomoč, na koncu pa so bili uspešni vsi. V razgovoru z dijaki smo ugotovili, da so bile učne metode in oblike dela ustrezno izbrane, da je večine dijakov obravnavano učno tematiko z zanimanjem spremljala.



Slika 6: Program za krmiljenje semaforja v GRAPH-u

3.3 Programiranje razdeljevalna postaje

V tretji učni situaciji dijaki že usvojena znanja pri prejšnjih vajah povežejo in uporabijo pri reševanju konkretne naloge. V uvodnem delu dijakom predstavimo osnovne gradnike razdeljevalne postaje, in sicer: zalogovnik, podajalnik, prenosno enoto in senzorje, ki so uporabljeni v razdeljevalni postaji (induktivni senzor, vakumski senzor, končno stikalo).



Slika 7: Tehnološka shema razdeljevalne postaje

Razdelilna postaja je primer oskrbovalne naprave. Oskrbovalne naprave so naprave, ki opravljajo funkcijo prenosa, razvrščanja in oskrbovanja drugih enot z materialom in obdelovanci. Dodatno lahko oskrbovalna naprava sortira komponente glede na različne lastnosti, kot so oblika, teža, itd.. Naloga razdelilne postaje je podati

obdelovanec iz zalogovnika ter ga prenesti do naslednje enote.

V nadaljevanju dijakom natančno opišemo in razložimo delovanje razdeljevalne postaje, od postpka inicializacije le-te, delovanje osnovnega cikla in dodatne zahteve v primeru, če pride do nekaterih napak, kot so neuspešen prijem obdelovanca, izguba obdelovanca med prenosom in neustreno obrnjen obdelovanec v zalogovniku. Dijaki imajo vso zahtevano dokumentacijo razdeljevalne postaje dostopno v spletni učilnici modula UPN.

Naloga je razdeljena na več korakov, in sicer: spoznavanje s tehnologijo krmiljenega procesa, konfiguriranje novega projekta in strojne opreme, določitev prireditvene tabele na osnovi načrta priključitve senzorjev in izvršnih členov, ustvarjane ustreznih programskih blokov, pisanje in prevajanje programa v krmilniku razumljivo obliko in prenos programa v krmilnik. Posamezni koraki so bolj natančno opisani v poglavju 3.1. Po končani razlagi smo dijakom najprej demonstracijsko prikazali celoten postopek programiranja razdeljevalne postaje. Pri tem je zelo pomembno, da spoštujemo didaktična načela, kot so nazornost, postopnost in sistematičnost. Po demonstraciji skupaj z dijaki na učne liste v obliki povzetka zapišemo posamezne korake programiranja in v nadaljevanju dijaki samostojno ponovijo posamezne korake, ki so potrebni za programiranje razdeljevalne postaje.

4 Ovrednotenje dela in preverjanje znanja

Ovrednotenje rezultatov opravimo na koncu učne situacije ali na koncu vsebinskega sklopa. Cilji ovrednotenja dela so: predstavitev rezultatov, ugotavljanje dijakovega napredka, ugotavljanje ustreznosti izbranih metod in oblik dela. Pomembno je, da za vrednotenje dela predvidimo dovolj časa. V razgovoru z dijaki smo ugotovili, da so bile učne metode in oblike dela ustrezno izbrane, da je večina dijakov obravnavano učno tematiko z zanimanjem spremljala.

S preverjanjem znanja načrtno in sistematično zbiramo podatke o tem, koliko in katere učne cilje so dijaki usvojili. Ocenjevanje je postopek, ko učnim dosežkom dodelimo številčno ali opisno oceno. Namen preverjanja je, da dobimo ustrezne informacije o dijakovi aktivnosti oz. katere vsebinske sklope je več ali manj usvojil. Dijake seznanimo o tem, ali delajo dovolj in ali uporabljajo pravilne metode in oblike učenja. Učitelj pa iz preverjanja pridobi informacijo o uspešnosti posameznih dijakov in celotnega razreda, o tem ali uporablja ustrezne metode in oblike dela, če je hitrost obravnavanja učne snovi primerna in če so dijaki ustrezno motivirani. Preverjanje je pomembno tudi za izboljšanje kakovosti učenja in poučevanja.

Dijaki so dobili učne liste (iz spletne učilnice ali na učnem mestu), ki so jih izpolnjevali ob izvajanju posameznih vsebinskih enot. Ob koncu vsake vsebinske enote so le-to predstavili učitelju in jo na prvi strani učnega lista označili kot realizirano. Opisani način preverjanja nam lahko v veliki meri pomaga tudi pri ocenjevanju. Oceno dijakovega znanja določimo po opredeljenih kriterijih za ocenjevanje znanja (odvisna je tudi od števila opravljenih nalog). Z opisanim načinom dela lahko odpravimo tudi negativne učinke ocenjevanja.

5 Zaključek

V prispevku smo prikazali didaktični pristop k poučevanju programiranja krmilnikov (PLK) v okviru modula UPN v programu elektrotehnik. Za dijake programa elektrotehnike je znanje programiranja krmilnikov primarna vsebina, ki se ocenjuje na poklicni maturi, zato je zelo pomembno, da jo čim bolje usvojijo.

Klasične metode in oblike dela so za poučevanje strokovnih vsebin vse manj primerne, kajti aktivnost dijakov je pri pouku minimalna. Moderne učne tehnologije nam omogočajo, da pouk naredimo bolj zanimiv in s tem tudi povečamo aktivnost dijakov. Priprava učne situacije, kjer dijaki čim več delajo samostojno, v začetni fazi zahteva zelo veliko priprav, v nadaljevanju pa omogoča lažje delo pri vnašanju sprememb in dopolnitev.

V predstavljenem načinu dela sta dosežena aktivno sodelovanje dijakov in dober učni uspeh. Opravljena evalvacija kaže, da so zadovoljni tudi dijaki. S tem, ko dijaka stalno spremljamo in preverjamo njegovo znanje med njegovim praktičnim delom v specializiranih učilnicah, lahko tudi ocenjevanje poteka v drugačnih oblikah. Dijake, ki uspešno opravijo zahtevane naloge, lahko že na osnovi spremljanja in preverjanja ustrezno ocenimo po vnaprej določenih kriterijih za ocenjevanje znanja.

Viri in literatura

[1] SIMENS Programming Guideline for S7-1200/1500 STEP7 (TIA portal) V1.2 03/2014

[2] SIMENS Ladder Logic (LAD) for S7-300 and S7-400 Programming Reference Manual 5/2010

[3] Klemen Hleb (2015) Uporaba krmilnega sistema Siemens S7-1500. Maribor: FERI, diplomsko delo

[4] Karl-Heinz John, Michael Tiegelkamp: IEC 61131-3 Programming Industrial Automation System Springer 2nd Edition 2010

Metodika poučevanja povezovanja LPC sistema krmilnikov v CAN mrežo in programiranje

Boris Pregelj

ŠC Nova Gorica, Cankarjeva 10, 5000 NOVA GORICA
E-pošta: borisp@scng.si

The method of teaching how to connect LPC system controllers in CAN network and programming

LPC-2, LPC-3 and GOT controllers are examples of compound and simple controllers, which can easily be connected into a CAN net. They contain all computer hardware for the connection integrated in the basic unit. Due to simple design they are an appropriate teaching tool at school as students can simulate various situations from fieldwork, e.g. they can be connected among each other either in a hotel, luxury cruiser, residential tower or dwelling-house

Povzetek

Krmilniki LPC-2, LPC-3 in GOT so primeri sestavljenih in enostavnih krmilnikov, ki se jih da enostavno povezati v CAN (Controler Area Network) mrežo. Omenjeni krmilniki imajo že vso strojno opremo za povezavo integrirano v osnovnem modulu. Zaradi enostavne zasnove so omenjeni krmilniki primerni za poučevanje na šoli. Dijaki lahko simulirajo različne situacije iz prakse. Tako lahko med seboj povežemo krmilnike v hotelu, luksuzni križarki in v stolpnici ali stanovanjski hiši.

1 Uvod

V industrijski proizvodnji se v današnjem času veliko uporablja krmilnike takega ali drugačnega tipa. Enostavne probleme lahko rešimo z uporabo samo enega enostavnega krmilnika ali ene krmilniške konfiguracije. Na primer regulacija temperature v domači hiši, krmiljenje žaluzij, vklop luči, idr. Ko je potrebno avtomatizirati zahtevnejše projekte (avtomatizacija v hotelih, potniških ladjah, naftnih ploščadih, proizvodnih linijah, ipd.), je potrebno povezovati konfiguracije v mrežo. Samo tako lahko zagotovimo nemoten pretok informacij znotraj sistema. Zaradi velikih potreb po taki avtomatizaciji se je na trgu pojavilo veliko število PLC naprav, ki omogočajo medsebojno povezovanje v različne tipe mrež. LPC-2, LPC-3 in GOT krmilniki domače proizvodnje so eni izmed mnogih. Imajo tako možnost neposrednega povezovanja v CAN mrežo brez nakupa dodatnih modulov. CAN open je 7-nivojski protokol, ki definira komunikacijo in funkcijo naprav v CAN sistemih. Uporablja se v različnih panogah industrije in domačo uporabo, - pametne hiše, medicina, železniška, avtomobilska, pomorska, letalska industrija idr.

2 Predstavitev učila

Na šoli sem na učni tabli sestavil 5 različnih krmilniških konfiguracij in jih ustrezno povezal. Dijaki morajo napisati program za pravilno delovanje posamezne konfiguracije in nato še programsko povezati vse konfiguracije v mrežo. Pripravil sem tudi priročnik z opisom delovanja programov, ki nam olajšajo povezovanje krmilnikov v CAN mrežo. Ta priročnik vsebuje tudi opis delovanja osnovnih in naprednih elementov, ki jih uporabljamo pri programiranju v programskem jeziku Ladder in FBD. Na koncu sem dodal nekaj vaj, s katerimi dijaki rešujejo probleme iz prakse.

Na učilu najdemo naslednje komponente (slika 1).

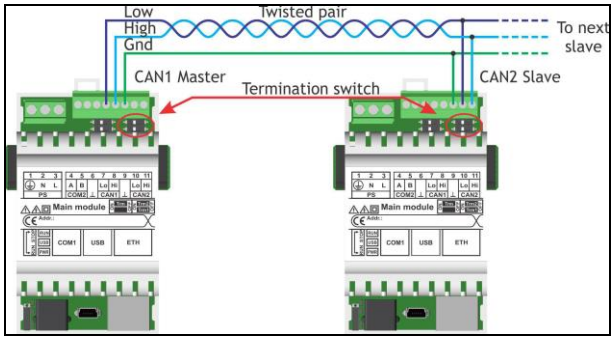
- MC8, IOU, IOR IOT in GOT osnovni krmilni moduli s pripadajočimi napajalniki
- R0-1 modul
- DD-2 digitalni dimmer
- letev za montažo konfiguracije
- napajalni kabel
- kabel za programiranje
- vtičnice
- CH-1, CA-1 in DP-1 moduli
- tipke
- kontaktor
- dve svetilki
- ventilator
- elektromagnetno ključavnico
- stikala
- temperaturno NTK tipalo
- dodatno opremo



Slika 1: Učilo - elementi, montirani na pleksi steklu

3 CAN MREŽA

CAN mreža Controler Area Network se široko uporablja zaradi dobrih lastnosti, med katerimi bi izpostavil predvsem robustnost delovanja (eno vozlišče se pokvari, vendar mreža deluje naprej), odpornost na motnje, možnost kratkotrajnih preobremenitev brez škodljivih posledic na prenos podatkov, možnost povezovanja enot v mrežo z navadnimi energetskimi žicami in kabli kakor tudi s kabli za prenos podatkov in optiko [1].



Slika 2: prikazuje shemo povezave v CAN mrežo.

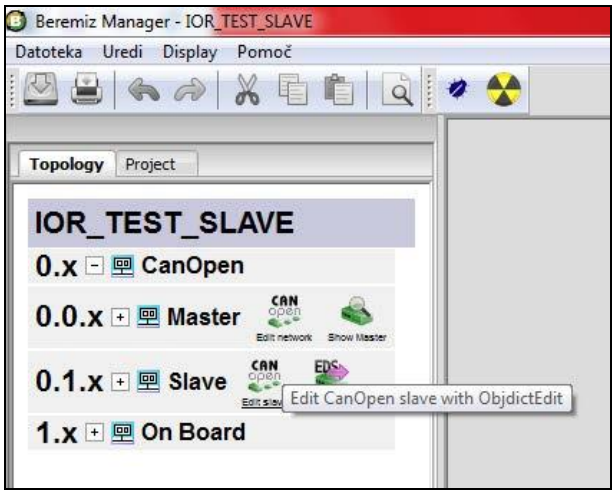
CAN mrežo lahko, odvisno od potreb, zgradimo v različnih topologijah. V svetu se največ uporablja mešana topologija med zvezdo in vodilom.
Signale preko CAN mreže prenašamo s pomočjo NRZ (Non Return to Zero) kode. Dobra lastnost te kode je, da ne vsebuje enosmerne komponente, potrebno pa je poskrbeti, da izločimo sinhronizacijski signal iz samega podatkovnega signala.

4 POSTOPEK DELA

Preden pričnemo z izdelovanjem mreže, je potrebno ožičiti krmilnike, ki bodo sestavljali mrežna vozlišča. To pomeni, da na posamezni krmilnik priključimo vse vhodno-izhodne naprave. Po možnosti testiramo posamezne konfiguracije, tako da naložimo delujoč program in preizkusimo delovanje posameznih konfiguracij brez mrežnih spremenljivk. To je priporočljivo, saj je odkrivanje napak v programu, ko so vnešeni mrežni parametri, mnogo težje. Krmilnike nato povežemo med seboj v mrežo z ustreznimi žicami (ponavadi prepletena parica) ter dopolnimo program.

5 UPORABA LPC SMARTEH IDE

LPC Smarteh IDE je programsko orodje, ki v enem programu združuje orodje za programiranje, orodje za povezavo krmilnikov v mrežo in orodje za izdelavo vizualnega grafičnega okolja za GOT krmilnik. Program je brezplačen in dokaj enostaven za uporabo in omogoča, da z njegovo pomočjo hitro izdelamo mrežo, napišemo program in narišemo grafično okolje za GOT touch screeen krmilnik. (Slika 3).



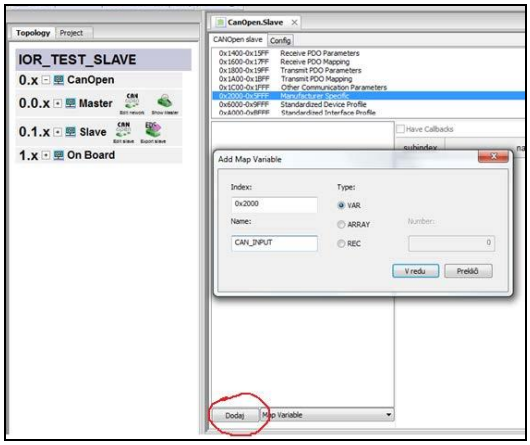
Slika 3: Okno programa LPC Smarteh IDE

6 Izdelava projekta v nekaj korakih

Projekt lahko z uporabo programskega orodja zgradimo v nekaj korakih. Poznati moramo postopke dela in biti pazljivi pri povezovanju mrežnih spremenljivk z drugimi programskimi spremenljivkami.

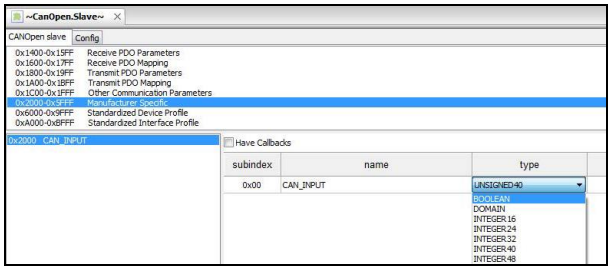
6.1 Določitev vloge krmilniku

Ko začnemo z delom moramo najprej izbrati kateri krmilnik v mreži bo opravljal vlogo nadrejenega krmilnika. Vsi ostali bodo v vlogi podrejenega. Tako moramo tudi fizično povezati krmilnike v mrežo na pripadajoče mrežne vhode. V našem primeru ima GOT krmilnik vlogo nadrejenega krmilnika, ostali so v vlogi podrejenih. Pri delu moramo najprej nastaviti vse potrebno v krmilnikih, ki imajo vlogo podrejenega. Najprej določimo številko vozlišča, ki jo predstavlja posamezni krmilnik, nato pa dodamo še spisek vseh potrebnih spremenljivk, ki jih krmilnik omogoča za delovanje v mreži.



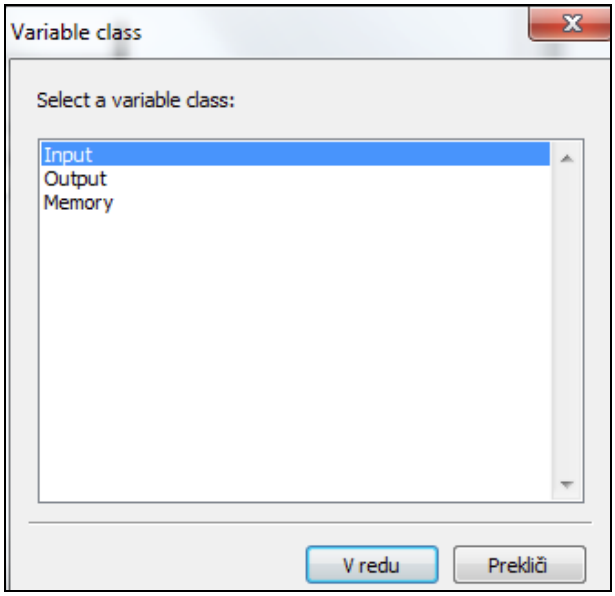
Slika 4: Primer okna za izbiro spremenljivk za delo krmilnika v mreži.

V naslednjem oknu, ki ga prikazuje slika 5, moramo izbrati spremenljivko in ji določiti tip.



Slika 5: Izbira tipa mrežne spremenljivke

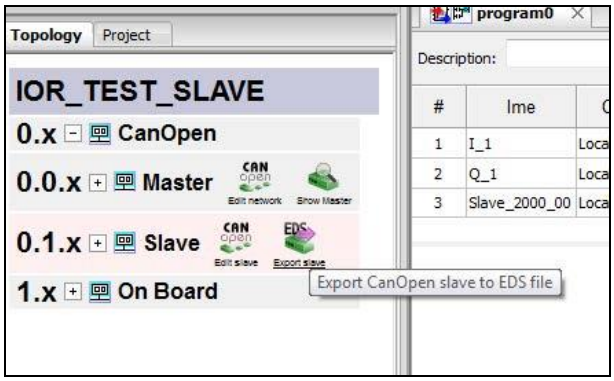
V naslednjem oknu moramo določiti ali bo izbrana spremenljivka vhodna ali izhodna in jo ustrezno poimenovati.



Slika 6: Določitev spremenljivke

6.2 Povezovanje spremenljivk v programu

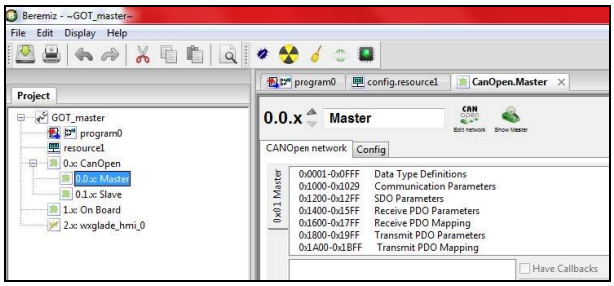
V 2. koraku moramo spremenljivko povezati s programom, ki opravlja določeno funkcijo. To naredimo s preprostim vlečnjem spremenljivke na delovno površino in hkratnim povezovanjem. Delo na podrejenem krmilniku končamo ko izvozimo pripadajočo datoteko v predpisano mapo.



Slika 7: Izbira za izvoz mrežnih parametrov na Slave krmilniku.

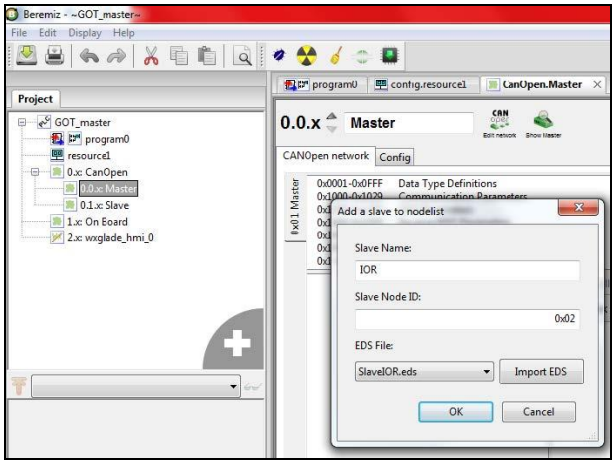
6.3 Povezava med podrejenimi in nadrejenimi spremenljivkami

Ostane nam še, da opravimo konfiguracijo krmilnika, ki ima vlogo nadrejenega. V začetnem oknu najprej izberemo ukaz Import EDS in izberemo datoteko, ki smo jo pred tem shranili.



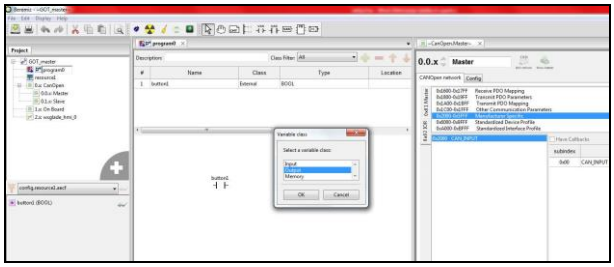
Slika 8: Prikaz izbire Master krmilnika v projektnem oknu.

Datoteko uvozimo in nastavimo enako številko vozlišča, kot jo ima krmilnik, ki opravlja vlogo podrejenega.



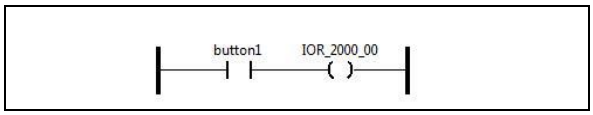
Slika 9: Prikaz konfiguracije nadrejenega krmilnika.

V zavihku CAN open Master označimo nadrejeni krmilnik in spremenljivko prenesemo v program. Tako je krmilnik povezan v mrežo. Pri delu moramo paziti, da je spremenljivka, ki je v podrejeni napravi določena kot izhod, v nadrejeni napravi določena kot vhod in obratno.



Slika 10: Konfiguracija spremenljivk v Master napravi

Na koncu moramo mrežne spremenljivke povezati med seboj v ustreznem programu. To naredimo grafično, ko povežemo med sabo ustrezno vhodno in izhodno spremenljivko. S tem je naše delo končano.

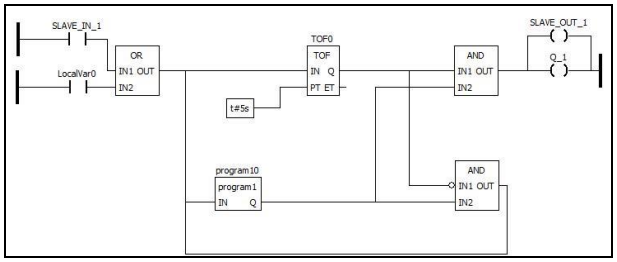


Slika 11: Prikaz povezave vhodne in mrežne spremenljivke.

Tako na dokaj preprost način zgradimo mrežo dveh ali več krmilnikov. Programsko okolje nam omogoča tudi simulacijo delovanja mreže in testiranje delovanja mreže v realnem času.

7 PISANJE PROGRAMOV

Ko je mreža zgrajena, moramo le še dopolniti ustrezne programe z mrežnimi spremenljivkami. To naredimo s programom LPC Beremiz, ki je del programskega paketa. Na koncu sledi testiranje pravilnosti delovanja mreže ali njenih posameznih delov. Ponavadi dodajamo mrežne spremenljivke vzporedno k vhodnim in izhodnim spremenljivkam in tako zagotovimo upravljanje naprave iz enega mesta.

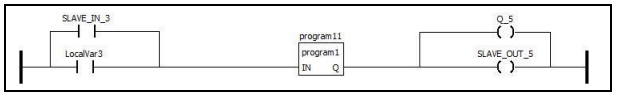


Slika 12: Del programa z uporabljenimi mrežnimi spremenljivkami

8 PRIMERI VAJ

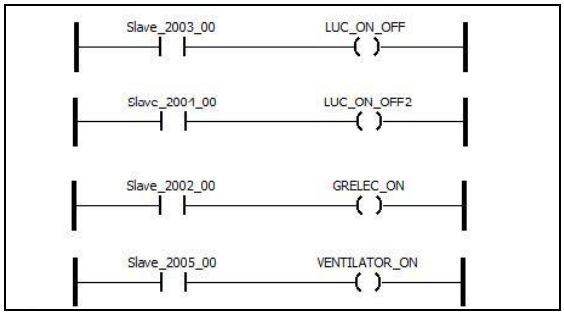
Vaje, ki sem jih pripravil, se delijo na zelo enostavne, na zahtevnejše in na najzahtevnejše primere, njihove rešitve pa lahko uporabimo v praksi. Učenje tako sledi načelu od enostavnega h kompleksnemu.

Primer enostavne vaje je vklop luči na krmilniku preko stikala ali tipke na drugem krmilniku.



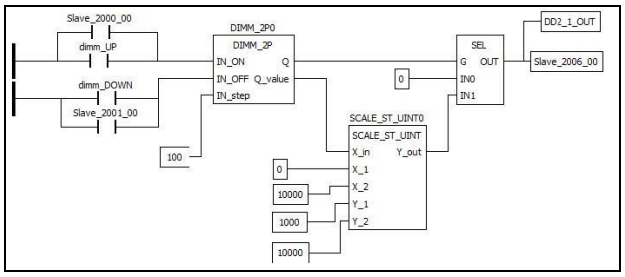
Slika 11: Primer povezave vhodne in mrežne spremenljivke.

Naslednji primer je vklapljanje aktuatorjev preko mreže.

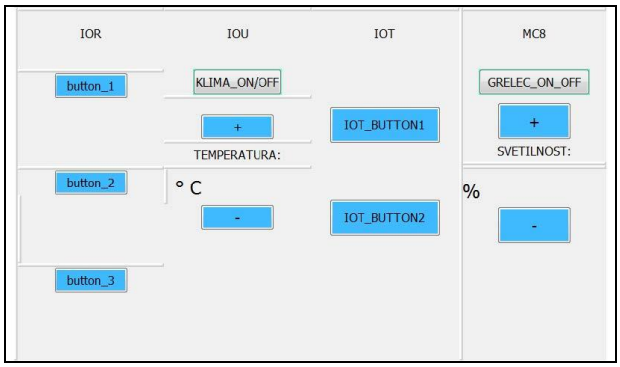


Slika 12: Prikaz krmiljenja več naprav preko mreže.

Zahtevnejši primer je pisanje programa za regulacijo ogrevanja in hlajenja v stanovanjski hiši, ali vklop in izklop luči preko mreže v več prostorih hkrati in nastavljanje svetlosti luči. (slika 13)



Slika 13: Prikaz rešitve zahtevnejšega primera s področja krmiljenja in regulacij.



Slika 11: Slika prametrov, ki jih upravljamo preko GOT master krmilnika.

9 ZAKLJUČEK

Dijaki se z uporabo učila na enostaven način naučijo najprej oživiti krmilnike in jih nato povezati v mrežo. V naslednjem koraku napišejo enostavne programe za povezovanje krmilnikov v mrežo. Tretji korak pa je že pisanje kompleksnih programov in iskanje samostojnih poti za rešitev določenega problema.
Prednost sistema je vsekakor ta, da je to plod domačega znanja. Zato je izdelek dostopen, podpora je v primeru težav v slovenskem jeziku. Izdelek je tudi cenovno ugoden v primerjavi s tujimi PLK in primerljivimi zmogljivostmi.
Učni pripomoček sem izdelal v letošnjem letu in ga nameravam uporabiti kot demonstracijsko učilo v 4. letniku SSI pri izbranih modulih. Z njim želim, da dijaki pridobijo temeljno znanje na področju povezave krmilniških konfiguracij v mrežo. Programsko okolje omogoča tudi tiskanje shem in shranjevanje mrežnih konfiguracij na disk. S pomočjo tega lahko dijaki pripravijo ustrezno tehnično dokumentacijo.

10 VIRI IN LITERATURA

[1] B. Pregelj: Sistem krmilnika LPC-3 in GOT 4 del: (priročnik za interno uporabo)

[2] SMARTEH: http://www.smarteh.si/podpora/uporabniski-prirocniki

Celostna prenova priprave nalog in ocenjevanja pri uvodnem programerskem predmetu

Luka Fürst

Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
E-pošta: luka.fuerst@fri.uni-lj.si

Holistic Reform of Task Preparation and Grading in an Introductory Programming Course

In the introductory programming course taught at the University of Ljubljana, Faculty of Computer and Information Science, we introduced a new task preparation and grading system. The main novelty of the system is the automatic evaluation of the students' programs using a set of test cases. The other, no less important part of the reform was the restructuring of the entire body of programming assignments. All assignments are now consistently formatted and equipped with a set of test cases divided into difficulty groups. In the paper, we present the rationale for the introduction of the new system, the system itself, and our experiences four years since its adoption.

1 Uvod

Računalništvo in informatika prodirata v vse pore našega življenja. Poučevanje ni nikakršna izjema. S »pametnimi« tablami je opremljena domala vsaka osnovna šola, da o tehnološki usposobljenosti, ki jo izkazujejo današnji učenci, niti ne govorimo. Računalniki pa niso v pomoč le pri podajanju snovi, ampak tudi pri ocenjevanju znanja. Obstaja čedalje več sistemov za ocenjevanje nalog, celo takih, pri katerih ne bi pričakovali, da jih bo mogoče kdaj avtomatizirati [1]. Prav nič nas ne preseneča, da so bile programerske naloge predmet enega od prvih avtomatskih ocenjevalcev [2], danes pa je na voljo kar lepo število sistemov [3, 4, 5].

Pri predmetu Programiranje 1 v prvem letniku univerzitetnega študija na Fakulteti za računalništvo in informatiko Univerze v Ljubljani (v nadaljevanju: FRI) imamo pestro zgodovino različnih ocenjevalnih režimov. V pričujočem prispevku predstavljamo prehod od načina ocenjevanja, ki smo ga v študijskem letu 2009/10 uvedli kot odziv na smernico, da naj bi študent polovico svoje ocene pri predmetu prejel s sprotnim delom tekom semestra, do avtomatiziranega sistema, ki smo ga po naraščajočem nezadovoljstvu z obstoječim stanjem vpeljali pet let kasneje. Poleg sistema ocenjevanja smo prenovili tudi sistem priprave programerskih nalog. Pri tem smo kot glavno merilo upoštevali konsistentnost: vse naloge naj bodo strukturirane na enak način in naj se po enakem postopku tudi ocenjujejo.

V razdelku 2 bomo predstavili predmet Programiranje 1 na FRI. Razdelek 3 je namenjen opisu ocenjevalnega sistema pred prenovo, razdelek 4 pa opisu trenutnega ocenjevalnega sistema. V razdelku 5 bomo razpravljali o izkušnjah s prenovo, z razdelkom 6 pa bomo članek zaključili.

2 Predmet Programiranje 1 na FRI

Pri predmetu Programiranje 1 se študentje na primeru programskega jezika java seznanijo s temeljnimi koncepti proceduralnega in objektno usmerjenega programiranja. Snov predmeta zajema krmilne strukture, funkcije (metode), tabele, razrede in objekte, dedovanje in osnove računalniške grafike. Predmet se izvaja v obliki predavanj in vaj. Predavanja so namenjena vsem študentom hkrati, na vajah pa so slušatelji razdeljeni v skupine po 18.

Študent[1] lahko na podlagi svojega sprotnega dela pridobi največ 50 točk, na izpitu pa največ 60. Da predmet opravi, mora skupaj zbrati najmanj 50 točk, od tega najmanj 25 na izpitu. Z uspešnim sodelovanjem na vsakoletnem tekmovanju v programiranju namiznih iger ali natečaju za pripravo domačih nalog [6] se mu k točkam za sprotno delo prišteje še nekaj dodatnih točk.

Gradiva za predmet Programiranje 1, ki jih pripravljamo in vzdržujemo izvajalci, vključujejo okrog 150 programerskih nalog. Poleg nalog, ki se rešujejo na predavanjih in vajah, ima študent na voljo 10 obveznih domačih in 76 neobveznih dodatnih nalog.

3 Ocenjevanje pred prenovo

Način ocenjevanja programerskih nalog se je v zgodovini predmeta Programiranje 1 večkrat spremenil. V tem razdelku bomo opisali sistem, ki smo ga uporabljali v letih od 2009/10 do vključno 2013/14.

V navedenem obdobju so študentje lahko prejeli do 30 točk iz domačih nalog, 20 točk iz teoretičnih kolokvijev in 50 točk iz izpita, ki so ga opravljali na papirju. Vsaka domača naloga je bila vredna po 10 točk, vendar pa smo pravilni rešitvi namenili le 4 točke. Preostalih 6 točk je študent lahko prejel na vajah v tednu po izteku roka za oddajo, in sicer tako, da je sprogramiral tri ločene nadgradnje domače naloge, urejene po naraščajoči težavnosti. Vsaka nadgradnja je prinesla po dve točki.

[1] Moška oblika se v prispevku uporablja nevtralno za oba spola.

Plagiatorstvo je tako izgubilo precejšen del svoje privlačnosti; če je študent rešitev prepisal, ne da bi jo razumel, je imel pri programiranju nadgradenj težave.

Začetno navdušenje nad sistemom pa je kmalu začelo plahneti. Prvič, težavnost implementacije nadgradnje je pogosto odvisna od implementacije osnovne naloge. Drugič, ker so študentje nadgradnje programirali na vajah in ker letno izvajamo 16 ali 17 skupin vaj, smo morali vsakokrat pripraviti več različnih domačih nalog in še več kompletov nadgradenj. Ponavadi smo pripravili štiri ali pet domačih nalog (za vsak dan izvajanja vaj po eno) in po dve, tri ali celo štiri komplete nadgradenj za vsako nalogo. Pri takšni količini je bilo težko zagotoviti že enakovrednost domačih nalog, kaj šele nadgradenj.

Slabosti so se pokazale tudi pri samem ocenjevanju. Od 90 minut, kolikor traja termin vaj, so študentje za programiranje nadgradenj imeli na voljo 60 minut, preostalih 30 minut pa je bilo namenjenih ocenjevanju. Asistent je tako posameznemu študentu namenil v povprečju manj kot dve minuti, kar pomeni, da je vsako nadgradnjo le enkrat pognal in morda na hitro preletel programsko kodo. Ker nismo imeli konsistentnih kriterijev (kaj šele avtomatiziranih testov), je na ocenjevanje dostikrat vplival halo-učinek. Tistemu, ki si je pridobil status »dobrega« študenta, smo (morda nehote) pogledali skozi prste in se zadovoljili z nedokončano rešitvijo nadgradnje, saj smo predpostavljali, da bi jo rešil do konca, če bi imel na voljo kakšno minuto več.

Poleg naštetega nas je čedalje bolj motilo dejstvo, da so trije termini vaj letno namenjeni izključno ocenjevanju. Bolje bi bilo, da bi v tistem času reševali naloge in tako pripravljali študente na izpit in poklicno življenje. Naposled smo se odločili, da navedene slabosti odpravimo, in tako smo v študijskem letu 2014/15 uvedli ocenjevalni sistem, ki ga še vedno uporabljamo.

4 Ocenjevanje po prenovi

4.1 Pregled sprememb

Glavna sprememba, ki jo je prenesla celostna prenova ocenjevalnega sistema, je uvedba *avtomatskega ocenjevanja*. Rešitve se po novem ne ocenjujejo več z ročnim poganjanjem in pregledovanjem programov, ampak z orodjem, ki program avtomatsko oceni na podlagi vnaprej pripravljenih testnih primerov [7]. Na ta način smo izločili subjektivni faktor ocenjevanja, poleg tega pa smo omogočili bistveno temeljitejše preverjanje rešitev, saj asistent na vajah praviloma ni imel časa, da bi preveril vse mogoče robne primere.

Vse naloge na vajah, domače naloge, izpitne naloge in neobvezne dodatne naloge so sedaj konsistentno oblikovane in opremljene z naborom testnih primerov. Izpit se po novem rešuje na računalnik, ocenjuje pa se na popolnoma enak način kot domače naloge.

Da bi povečali konsistentnost težavnosti nalog, smo se nadgradnjam odpovedali, zato pa smo število domačih nalog povečali s 3 na 10. Ta sprememba je skladna z že omenjeno smernico, ki vzpodbuja sprotno delo študentov.

4.2 Tipi nalog

Vsaka naloga pripada enemu od sledečih tipov:

Vhod-izhod. Pri nalogah tega tipa študentov program prebere svoje vhodne podatke s standardnega vhoda, rezultate pa izpiše na standardni izhod. Vsak testni primer je sestavljen iz datoteke s podatki, ki tvorijo vhod, in datoteke s podatki, ki tvorijo pripadajoči pričakovani izhod. Ocenjevalno orodje posreduje vsebino vhodne datoteke na standardni vhod študentovega programa, standardni izhod programa pa preusmeri v posebno datoteko. Ta datoteka se nato primerja z datoteko s pričakovanim izhodom. Testni primer se šteje kot pravilno obravnavan natanko tedaj, ko sta datoteki enaki.

Razred-izhod. Pri nalogah tega tipa študentje ne napišejo samostojnega programa, ampak enega ali več (neizvršljivih) razredov.[2] Vsak testni primer je sestavljen iz datoteke z izvršljivim testnim razredom in datoteke s pripadajočim pričakovanim izhodom. Testni razred ustvari enega ali več objektov razredov, ki tvorijo rešitev naloge, nato pa nad njimi kliče posamezne metode in rezultate klicev izpisuje na standardni izhod. Ocenjevalno orodje požene vsak testni razred posebej in standardni izhod preusmeri v posebno datoteko, njena vsebina pa se primerja z vsebino datoteke s pričakovanim izhodom. Testni primer se šteje kot pravilno obravnavan natanko tedaj, ko se vsebini ujemata.

Grafika. Pri nalogah tega tipa študentje dopolnijo vnaprej pripravljeno ogrodje tako, da se bo ob zagonu na grafično podlago (to je bodisi okno bodisi slikovna datoteka) narisal zahtevani vzorec. Študentje napišejo metodo, ki sprejme širino in višino podlage ter objekt javanskega razreda `Graphics2D`, s katerim je na podlago mogoče risati. Vsak testni primer je sestavljen iz izvršljivega razreda in pripadajoče referenčne slikovne datoteke. Ocenjevalno orodje požene izvršljivi razred, ta pa ustvari slikovno datoteko in pokliče študentovo metodo, da vanjo nariše vzorec. Takó izdelana datoteka se nato primerja z referenčno slikovno datoteko.

Za razliko od nalog tipov vhod-izhod in razred-izhod se posamezni testni primeri ne ocenjujejo zgolj z enico (popolno ujemanje) ali ničlo. Namesto tega se referenčna slika R in slika S, ki jo izdela študentova metoda, primerjata po posameznih pikah (slikovnih elementih), in to tako, da imajo vse *barve* enako težo; vsako konceptualno komponento slike je pri teh nalogah namreč treba narisati v svoji barvi.

Oceno ujemanja slik R in S izračunamo po formuli $\min\{s_{RS}, s_{SR}\}$, pri čemer

[2] Razred je mogoče pognati, če vsebuje metodo `main`.

$$s_{RS} = \frac{1}{C_R} \sum_{p=1}^{w_R} \sum_{q=1}^{h_R} \frac{t_{RS}(p,q)}{N_R(R(p,q))},$$

$$s_{SR} = \frac{1}{C_R} \sum_{p=1}^{w_S} \sum_{q=1}^{h_S} \frac{t_{SR}(p,q)}{N_R(S(p,q))},$$

$$t_{IJ}(p,q) = \begin{cases} 1 & \text{če } \exists u, v \in \{-1,0,1\}: \\ & \qquad J(p+u,\, q+v) = I(p,\, q) \\ 0 & \text{sicer} \end{cases}$$

Oznaka $I(p,\, q)$ predstavlja barvo pike na koordinatah $(x,\, y) = (p,\, q)$, oznaka $N_I(b)$ število pik barve b na sliki I, oznaka C_I pa število različnih barv na sliki I. Vse barve na referenčni sliki imajo enako skupno težo: če je, denimo, referenčna slika sestavljena iz r rdečih in z zelenih pik, ima vsaka rdeča pika težo $1/(2r)$, vsaka zelena pa težo $1/(2z)$. Količina $t_{IJ}(p,q)$ je enaka 1 natanko tedaj, ko se na položaju $(p,\, q)$ na sliki J ali v neposredni okolici tega položaja nahaja pika, ki ima enako barvo kot pika na položaju $(p,\, q)$ na sliki I. Odstopanje za eno piko v vsako smer dopuščamo zaradi morebitnih zaokrožitvenih napak.

Pri vseh treh tipih nalog ima študentov program na voljo le omejeno količino časa, denimo 1 sekundo za vsak testni primer. Če program do izteka časovne omejitve ne proizvede pravilnega izhoda, se izvajanje prekine, testni primer pa se šteje kot napačno obravnavan.

4.3 Zgradba nalog

Vse naloge (naloge na vajah, domače, dodatne in izpitne) imajo konsistentno zgradbo. Pri nalogah tipa vhod-izhod pričnemo z opisom problema, sledi natančna specifikacija vhoda in izhoda, zaključimo pa s konkretnim primerom vhoda in izhoda. Podatkovnih struktur ne predpisujemo, le pri (zelo redkih) nalogah kaj prepovemo. Na primer, kadar se v nalogi števila obravnavajo po števkah (npr. izračunaj vsoto števk podanega števila), prepovemo rabo tabel, nizov in podobnih struktur, ki programerju omogočajo, da se izogne aritmetičnim operacijam.

Pri nalogah tipa razred-izhod podamo specifikacije javno dostopnih konstruktorjev in metod, ki jih morajo študentje implementirati v svojih razredih. Za vsak tak element podamo tipe parametrov in morebitne omejitve ter pričakovani rezultat oziroma učinek. Privatnih elementov razreda (npr. atributov) ne predpisujemo.

Pri grafičnih nalogah študentje dopolnijo metodo, ki sprejme širino in višino grafične podlage ter objekt za risanje na podlago. V besedilu naloge opišemo posamezne grafične elemente ter njihove barve in relativne položaje in velikosti (glede na širino in višino podlage).

4.4 Struktura testnih primerov

Vsaki nalogi pripada po 10 javnih in 50 skritih testnih primerov. S pomočjo javnih primerov lahko študentje preizkušajo svoje rešitve, skriti pa so namenjeni ocenjevanju, zato jih objavimo šele po izteku roka za oddajo domače naloge. Testni primeri so razdeljeni na skupine različnih težavnostnih stopenj. Razmerja med skupinami v množici javnih primerov so enaka kot v množici skritih. Na primer, če je množica javnih primerov razdeljena na skupino treh najlažjih primerov, skupino petih srednje težkih primerov in skupino dveh najtežjih primerov, bodo v množici skritih primerov pripadajoče skupine obsegale 15, 25 in 10 primerov.

Pri nekaterih nalogah je težavnost določena predvsem z obsegom dela (za več pravilno obravnavanih testnih primerov in s tem za boljšo oceno je treba več narediti), pri nekaterih pa višja številka testnega primera pomeni trši oreh. Težavnost lahko marsikdaj uravnavamo z omejitvami vhodnih podatkov: za lažje testne primere določimo ostrejše, za težje pa milejše omejitve. Na ta način za nižjo oceno zadošča rešiti poseben, lažji problem, za višjo pa splošnejši (in tako težji) problem.

Oglejmo si razdelitev testnih primerov pri sedmi domači nalogi v sezoni 2014/15. Pri tej nalogi so študentje napisali program, ki s standardnega vhoda prebere število vrstic (m), število stolpcev (n) in dvojiško matriko velikosti $m \times n$, na standardni izhod pa izpiše število 4-povezanih skupkov enic v dvojiški matriki. Testni primeri so bili razdeljeni v šest skupin:

1. Vsak skupek je sestavljen iz ene same enice.
2. Skupki imajo obliko vodoravnih in navpičnih daljic.
3. Skupki imajo obliko polnih pravokotnikov.
4. Vsaka vrstica skupka je sestavljena iz enega samega strnjenega zaporedja enic.
5. Skupki so lahko poljubne oblike.
6. Skupki so lahko poljubne oblike, poleg tega pa velja $m, n \in [1, 1000]$ (za vse ostale skupine velja $m, n \in [1, 100]$).

Kot vidimo, so lastnosti lažjih testnih primerov vsebovane v lastnostih težjih; lažji primeri so tako zgolj posebni primeri težjih. Skupek, ki je sestavljen iz ene same enice, je hkrati tudi daljica in pravokotnik. Njegova edina vrstica je sestavljena iz enega samega zaporedja enic, seveda pa sodi tudi v kategorijo poljubnih oblik.

Če želimo rešiti peto skupino, moramo implementirati vsaj rekurzivni algoritem poplavljanja (angl. flood fill), pri šesti pa niti to ni dovolj, saj nam zaradi velikosti vhodne matrike pri izvajanju zmanjka sklada. Zato popolna rešitev vključuje iterativni algoritem poplavljanja. Kdor se zadovolji z nižjim številom točk, pa lahko izbira med različnimi *ad hoc* pristopi. Za rešitev prve skupine testnih primerov, denimo, zadošča, da preštejemo enice v vhodni matriki.

5 Izkušnje

Vpliv celostne prenove ocenjevanja težko kvantificiramo, še težje pa sklepamo, kaj lahko pripišemo prenovi, kaj pa drugim faktorjem. Rezultati izpitov so v zadnjih letih

boljši kot v preteklosti, vendar pa današnje izpite, ki potekajo na računalniku in se avtomatsko ocenjujejo, težko primerjamo z nekdanjimi papirnimi. Druga kvantitativna mera, na katero bi se morda lahko oprli, so študentske ocene. Tudi te se v splošnem izboljšujejo (avtor prispevka je za leto 2016 prejel celo študentsko nagrado »naj asistent na univerzitetnem študiju«), žal pa ne vemo, v kolikšni meri lahko večje zadovoljstvo z učnim osebjem pripišemo prenovi, v kolikšni pa aktivnejšemu načinu izvajanja vaj, upoštevanju študentskih mnenj itd.

Če se omejimo na kvalitativna merila, lahko trdimo, da nas je prenova odrešila številnih problemov, s katerimi smo si belili glavo v preteklosti. Sedaj ni več pripomb glede različne težavnosti domačih nalog, saj je vsaka domača naloga za vse študente enaka. Avtomatsko ocenjevanje je po eni strani neusmiljeno, saj ne pozna »skoraj« pravilnih rešitev, po drugi strani pa študente navaja na natančno sledenje specifikacijam. (Javni testni primeri to krutost sicer precej ublažijo.) Glavna prednost avtomatskega ocenjevanja pa je njegova objektivnost. Če so ogledi v času papirnih izpitov lahko trajali tudi dve uri, saj se je marsikdo skušal za oceno še pogajati, je ta vidik ocenjevanja po uvedbi avtomatizacije povsem izginil. Skrite testne primere objavimo po izteku roka za oddajo rešitev izpitnih nalog, tako da si študentje lahko oceno nemudoma izračunajo sami.

Ker se domače naloge po novem ne preverjajo več na vajah, ampak na službenih računalnikih asistentov, imamo vseh 14 tednov v semestru na voljo za reševanje nalog in pripravo na izpit. Zaradi konsistentne zgradbe nalog so študentje do izpita vsaj po tehnični plati (delo z vhodom in izhodom, poganjanje orodja za testiranje ...) že zelo dobro usposobljeni, zato na izpitu nimamo tovrstnih težav. Redne domače naloge študente silijo v sprotno delo, kar se komu morda zdi omejujoče, vendar pa po drugi strani v zadnjih letih precej redkeje opažamo študente, ki bi se še dan pred izpitom mučili z zanko za izpis prvih desetih naravnih števil. Z ocenjevanjem domačih in izpitnih nalog ni več skoraj nobenega dela; treba je zgolj pognati program, ki vse postori sam. Sicer je res, da moramo več truda vložiti v pripravo *posameznih* nalog in testnih primerov, a glede na to, da moramo sedaj na leto pripraviti bistveno manj nalog kot pri prejšnjem režimu ocenjevanja, smo časovno še vedno na boljšem.

Za največjo pomanjkljivost novega sistema se je izkazalo plagiatorstvo. Čeprav je število študentov, ki so nalogo vsaj enkrat prepisali ali ponudili v prepisovanje, od začetka do danes nekoliko upadlo (od 72 v sezoni 2014/15 do 49 v sezoni 2017/18) in čeprav primere prepisovanja s pomočjo sistema Moss[3] dosledno odkrivamo in (morda preblago) kaznujemo, problem še zdaleč ni izkoreninjen. Poleg tega gotovo obstajajo tudi študentje, ki jih ne zazna noben sistem za odkrivanje plagiatorstva, saj jim domačo nalogo reši nekdo izven letnika. Take študente lahko odkrijemo kvečjemu na vajah, če naletimo na očitno neskladje med njihovim samostojnim delom v laboratoriju in ocenami domačih nalog. Problem bi lahko do neke mere rešili z dodatnim preverjanjem razumevanja rešitev, vendar pa bi za to potrebovali čas in pripravljenost na »spopadanje« s problematičnimi študenti.

[3] `http://theory.stanford.edu/~aiken/moss/`

6 Zaključek

V članku smo predstavili prenovo ocenjevalnega sistema pri uvodnem programerskem predmetu na ljubljanski Fakulteti za računalništvo in informatiko. Poleg avtomatizacije ocenjevanja smo popolnoma prenovili sistem izdelave programerskih nalog. Naloge so sedaj konsistentno strukturirane in opremljene s testnimi primeri, razdeljenimi v težavnostne skupine. Ocenjujemo, da je prenova prinesla veliko pozitivnih učinkov. Edini zares negativni je po našem mnenju plagiatorstvo, seveda pa je to izziv za vse sisteme, v katerih se domače naloge ocenjujejo.

Če odštejemo tiste redke naloge, v katerih prepovemo uporabo tabel in nizov, imajo študentje pri reševanju domačih in tudi izpitnih nalog popolno svobodo; edini pogoj je ta, da program za podani vhod ali testni razred znotraj predpisanih časovnih okvirov proizvede pričakovani izhod. Pri četrti nalogi v sezoni 2014/15 pa smo manjši del točk namenili tudi »eleganci« rešitev. Naloga je zahtevala implementacijo razredov `Tocka` in `Premica`, ki sta bila zasnovana tako, da je bilo večino metod mogoče sprogramirati s klici že napisanih metod. Na primer, v metodi za izračun pravokotne projekcije točke na premico smo si lahko pomagali z metodo za izračun parametrov pravokotnice skozi točko in metodo za izračun presečišča premic. Število točk za »eleganco« se je izračunalo na podlagi števila pričakovanih klicev, ki so jih vsebovale posamezne metode. Tovrstnih zamisli kasneje nismo več preizkušali, kljub temu pa predstavljajo zanimivo temo za prihodnje raziskovalno delo.

Literatura

[1] K. Zupanc, Z. Bosnić, "Automated essay evaluation with semantic analysis," *Knowledge-Based Systems*, vol. 120, str. 118–132, 2017.

[2] J. Hollingsworth, "Automatic graders for programming classes," *Communications of the ACM*, vol. 3, no. 10, str. 528–529, 1960.

[3] K. Ala-Mutka, "A survey of automated assessment approaches for programming assignments," *Computer Science Education*, vol. 15, no. 2, str. 83–102, 2005.

[4] R. Romli, S. Sulaiman, K. Zamli, "Automatic programming assessment and test data generation - a review on its approaches," in *Information Technology (ITSim) 2010, Kuala Lumpur, Malaysia*, 2010, str. 1186–1192.

[5] V. Pieterse, "Automated assessment of programming assignments," in *Computer Science Education Research Conference (CSERC) 2013*, 2013, str. 45–56.

[6] L. Fürst, V. Mahnič, "Introductory programming course: motivating students with prior knowledge," *World Transactions on Engineering and Technology Education*, vol. 11, no. 4, str. 400–405, 2013.

[7] M. Poženel, L. Fürst, V. Mahnič, "Introduction of the automated assessment of homework assignments in a university-level programming course," in *MIPRO 2015, Opatija, Croatia, 2015*, 2015, str. 761–766.

Learning model for developing critical thinking and encouraging innovativeness in engineering education by problem based learning

Veronika Šuligoj, Janez Jamšek

Univerza v Ljubljani, Pedagoška fakulteta Ljubljana, Kardeljeva ploščad 16, Ljubljana
E-pošta: veronika.suligoj@pef.uni-lj.si

Abstract. *Paper scopes on learning model whose final aim is to obtain pre-service engineering education teachers transformation from the university level to the primary school level affecting student's design thinking and innovativeness. For this purpose, the model is based on problem-based learning. The proposed approach was executed at the university program level for elementary teacher program at Technical studies with didactics course in 201/2018 studying year at the University of Ljubljana. A laboratory exercise where students make paper gliders by using the presented learning model is described. Learning model begins by instructional design and problem challenge task which is followed by a problem solving task. Students were divided in five groups where gliders relevant concept knowledge was varied before solving the problem was initiated. Students showed group cooperation and discussed possible problem solutions before production. All groups could improve paper glider model to gain better functionality when loaded. The groups with a higher knowledge given were more successful however we could not observe a higher degrees of innovations at any group. Students are predominantly oriented toward finding a trivial solutions and rather gain experiential knowledge before deducting known one to a given case.*

1 Introduction

The number of problems is growing rapidly whereas their complexity increases [1]. Technological knowledge alone will not be sufficient to solve rapidly changing problems [2]. Education can prepare students to successful adaptation to the changing world by teaching them to be critical thinkers. These skills enable students to effectively analyse multiple sources of information, draw logical conclusions, and create new innovativeness [3]. Critical thinking is deeply processing knowledge to identify connections across disciplines and find potential creative solutions to problems [2, 3]. Critical thinkers use decision-making and problem-solving skills to analyse situations, evaluate arguments and draw appropriate inferences [3].

Nowadays, students have not enough research-oriented, interactive and collaborative learning experiences. Learning content should be more motivational. We have to strive to discourage passive learning environments that discourage learning motivation, and do not promote lifelong skills. All this can be achieved by using active teaching methods, the inductive methods like problem based learning. Besides that, students have to collaborate and actively engage in the learning process [4].

Education has long be criticized for not promoting creativity, innovation and technology, and that there is a little student involvement of their authentic experiences in learning process [5]. Making products is being strongly emphasised in technology education. Student's attention is focused on material they are manipulating, instead on creativity, technology, and innovation [6]. Primary school education should follow society demands and produce people who are proficient in basic skills, are prepared to learn new things, collaborate in solving problems, and can produce innovations [7]. There are many definitions for innovation. It can be defined as a significant improvement, refinement or introduction of something new. Innovation and creativity are significantly different. Innovativeness cannot occur without creativity, however creative thinking is not necessarily innovative [6]. Creativity is the ability to produce ideas, processes or product that are novel, original, imaginative, unexpected, and useful [6], meanwhile innovation include changes. Result is a useful product or process that is commercialized and widely spread. Students have to have open design tasks and learn knowledge and skills to be able to create innovative solutions [6].

At the Faculty of Education in Ljubljana, we educate primary school trainee teachers. In the 4th year of the undergraduate study programme, technology content is included in Technical studies with didactics course (TDC). The course consists from lectures (15 h) and laboratory exercises (30 h). In-service primary school teachers cover technology content from the 1st to the 5th grade of the primary school education. Before the undergraduate study the students have last had technology content in the 8th grade of the primary school. Big time gap results in a very poor technological knowledge. Existing knowledge is inadequate, unstructured, unrelated and with many misunderstandings of the basic technological concepts. The TDC main goal is that undergraduate students acquire all the necessary technology knowledge and skills for technology teaching. In the academic year 2017/2018, we carried out part of the laboratory exercises in accordance with 21st century competencies [4]. In particular, the development of innovative

thinking was highlighted. In addition to acquiring technological knowledge, skills, and competencies, we aim to encourage primary school teachers to innovate and to develop innovative products. Our goal is the production of new, original and useful products, and moving away from reproduction of simple and trivial solutions products. To gain innovativeness we need to achieve critical thinking and problem solving. In the following a learning model based on problem based learning is presented for a hands on laboratory exercise.

2 Method

The proposed learning model consists of 5 phases, Fig. 1 and is based on different teaching methods and our previous findings, Tab. 1. Laboratory exercise starts with understanding the problem by giving students a challenge by instructional design (ID), 1st phase (challenge). Students produce a simple product by a given plan that is relevant for the field of contents. Phase output is critical thinking. Application, analysis and evaluation process are enhanced in providing cause related to consequence (CC) answers while working in pairs and test in advance given product functionality. To be able to gain innovativeness students must have a deeper contents knowledge. This is provided in the 2nd phase (knowledge). E-learning material can be used. If students already have a satisfactory knowledge level the phase can be skipped. The longest phase is the 3rd one (problem) where a challenge from the 1st phase is graduated to more realistic, actual and real life however at the same time limited the given time possibilities. Although this phase is based on problem based method it rather a mixture of inductive methods with the purpose to emphases technological functionality and usability essence. The functionality to gate the minimum solution level and the usability to gain optimization level. This enables determining the best product solution and stimulates towards innovative solution products. The method used in 3rd phase combines challenge-based learning (CBL) and discovery based learning (DBL). Students gain knowledge through resolving a problem based on their past experiences (phases 1 and 2). But, it is not CBL, as the problem is not open-ended. Further the used method can be seen as a mixture of problem-based learning (PRBL) and project-based learning (PJBL). It enables students to use critical thinking, available resources and technology. The role of the teacher is assistant or tutor. The starting point is a problem with unclear solution that requires original solutions from students. Students are active participation of the learning process. But, it is not PRBL as the problem is not opened and students do not have to study the literature and search for information to solve the problem. The solution to the problem is not theoretical, but a concrete product. The method is also not a PJBL as the problem is fixed in advanced and the solutions to the problem do not rely on the students' interests [9]. In 3rd phase each student

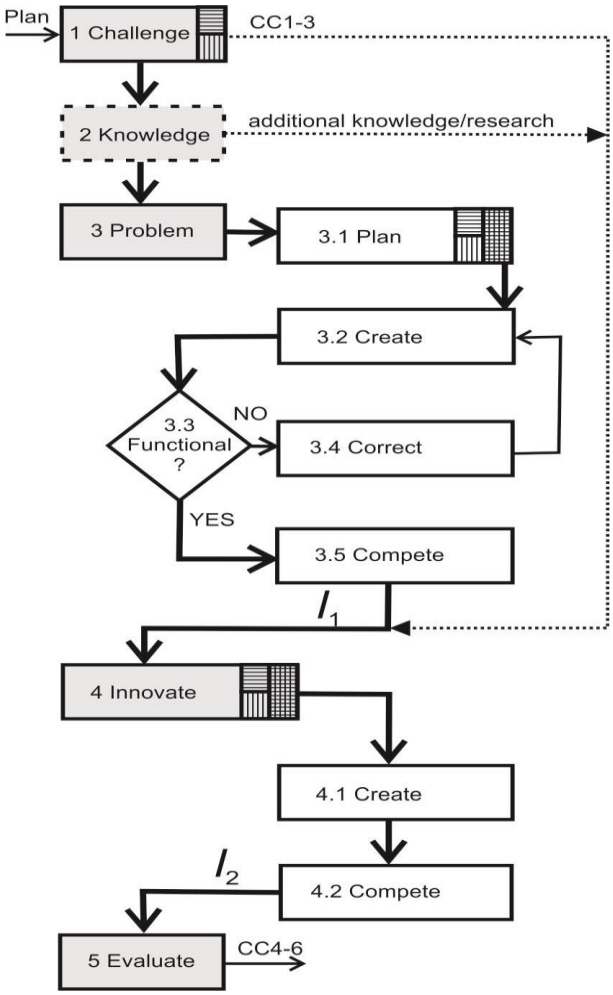


Figure 1. Proposed learning model, where CC- cause related to consequence answers, *I*- measure for functionality/usability.

In a pair produces their own draft product solution whereas the final solution is expected to be a mixture of both drafts with strong arguments behind.

Planning must start by setting the functionality criteria. Students gain the in advance product functionality set by correcting feedback loop until they do (sub phases 3.2-3.4). The phase ends with a competition to determine the best solution based on the usability measures. The most trivial usability is defined by graduate increasing load of the product solution. Once the 3rd phase is completed students gain new knowledge from all the different solutions produced. Due to simplicity of the problem product solution students are expected to obtain clear CC relations towards product functionality and usability. In the 4th phase (innovation) our learning model objectice is that students take into account best 3rd phase product solution, determined CC in the 1st phase and assimilation of the additional given knowledge into a new product solution being an upgrade to the 4th phase best solution thus gaining innovativness. While in the 3rd phase students can test functionality freely and use new material to make a product solution in the 4th phase

students use only the given material and do not test their functionality nor usability. This is done by a laboratory exercise conductor. The final, 5th phase (evaluate) is to present innovative solutions that are evaluated. Clearing the content concepts, and relations. Inductive reasoning to deduction stimulating critical thinking is enhanced. Students are asked to provide additional CC answers. Learning model ends with meaningful learning.

To measure critical thinking level a pre-test and a post-test are suggested encompassing multi-choice questions covering the higher three revised Bloom's Taxonomy levels (analyse, evaluate and create). Students' innovativeness is product related and group dependant assessed with a five-point scale, where 1 means no innovativeness and 5 means very innovative.

3 Results

Paper glider product is used in the proposed learning model. It is related to primary school curriculum (technology content). It is one of the most typical product where primary school teachers' lack of concept knowledge related to product functionality/ usability.

In the 1st phase students made an origami glider according to a given plan (longer side of A4) and material (A4 paper sheet). Functionality (gliding) was tested by glider hand throwing and obtaining the average gliding length. In pairs they provided answers regarding the causes for the difference in gliding distance (CC1- point of glider holding, CC2- throwing direction (up, straight, down), CC3- throwing velocity (push, slow, arm stretch)).

In the 2nd phase students were divided in five groups (G) where e-learning materials with information regarding gliders relevant concept knowledge was varied, Tab. 1. G1 skipped phase 2. G2 and G3 received e-learning material (aircraft definition, glider model components, gliding principle and glider examples), where G3 lacked of gliding principle. G4 and G5 both had additional insight into the problem understanding by expanding 1st phase challenge where glider plan changed (shorter side of A4 paper sheet) and possible usage up to 10 paper clips for glider balance. The purpose was to determine whether inductive active work results better than passive e-learning material reading. G5 additionally received the whole e-learning material.

In the 3st phase a problem was presented to students: build a model of paper glider that fly at least 4 m straight and carry a load of 30 g (functionality). Students could use two A4 paper sheets, scissors, adhesive tape or paper glue, and up to 20 paper clips.

In the 5st phase students provided answers regarding gliding distance difference (CC4-what increases gliding distance, CC5- what decreases gliding distance, CC6- what is difference between throwing a glider or a dart).

72 students took TDC whereas results are given only for students who solved both the pre-test and the post-test. The sample represents 45 students (6.66 % of male students and 93.33 % of female students).

Table 1. Variation of e-learning material, where *N* – number of students in the group, AD- aircraft definition, GMC- glider model components, GP - gliding principle, GE - glider examples.

Group (*N*)	Expan-ded phase 1	*Conceptual knowledge & examples*			
		AD	GMC	GP	GE
1 (6)	/	/	/	/	/
2 (9)	/	1	1	/	1
3 (5)	/	1	1	1	1
4 (14)	1	/	/	/	/
5 (11)	1	1	1	1	1

3.1 Critical thinking

Critical thinking was assessed by pre/post-test. The reliability of the test used was demonstrated by the coefficient Cronbach $\alpha = 0.63 > 0.60$. One-way variance analysis (ANOVA) did not show statistically significant differences between groups on the pre-test (F=0.73, df=44, α=0.58). Statistically significant differences were found between groups on the post-test (F=5.11, df=44, α=0.002). The Games-Howell post-hoc test showed statistically significant differences between the G2 and G4 (α=0.004) and between the G2 and G5 (α=0.000). On average, the G2 group reached the higher score ($\bar{x}$=5.67; s_x = 0.87), the G5 reached the lowest ($\bar{x}$=2.36; s_x =1.69), the G4 reached 3.07 points (s_x = 2.08).

Table 2. Gain (Δ) of knowledge in percentages by groups where the process dimension is tagged by 1-6 and the knowledge dimension is tagged by A-D.

Question (RBT)	Δ_{G1}/%	Δ_{G2}/%	Δ_{G3}/%	Δ_{G4}/%	Δ_{G5}/%
1.1 (C5)	0.0	33.3	0.0	7.1	-8.1
1.2 (C5)	-50.0	-1.1	-20.0	0.0	-81.8
2.1 (D4)	0.0	0.0	-80.0	7.1	-9.1
2.2 (D4)	0.0	33.3	0.00	7.2	-18.2
3.1 (D4)	-17.7	22.2	-80.0	0.0	-9.1
3.2 (D4)	0.00	0.0	-80.0	0.0	18.2
4.1 (D4)	-66.7	44.4	0.0	-21.4	-35.8
4.2 (D4)	-16.6	33.3	20.0	-28.6	-45.4
5.1 (D4)	16.7	44.5	0.0	-21.4	-36.3
5.2 (D4)	33.3	33.3	0.0	-28.5	-63.6

Questions are classified according to the revised cognitive Bloom's taxonomy (RBT), Tab. 2. It can be seen that G2 made the most progress in their knowledge while G5 made the least. For most questions, an anty-progress is evident. For other groups, very little progress in knowledge is noticeable.

3.2 Innovativeness

In assessing innovativeness, we have taken into account the characteristics of an innovative product [5], adjusted for our needs. We focused primarily on usability of paper gliders (gliding length according to load weight). We also considered number of modifications and the degree of modification: (1) without noticeable similarity to given examples, (2) in connection with theory, (3) without obvious connection with examples and theory, (4) exceeds the given framework, (5) completely

Table 3. Lengths of paper gliders by group and its subgroups in percentages and degree of innovativeness.

Group	l_1/ %	l_2 / %	Degree of innovativeness
1.1	75,0	112.5	2
1.2	100,0	112.5	2
1.3	100,0	150,0	2
1.4	62.5	75,0	1
2.1	0,0	112.5	2
2.2	80,0	125,0	2
2.3	37.5	50,0	1
2.4	112.5	112.5	2
2.5	150,0	87.5	2
3.1	112.5	100,0	2
3.2	37.5	75,0	2
3.3	87.5	80,0	1
3.4	25,0	75,0	1
3.5	37.5	62.5	1
4.1	87.5	112.5	2
4.2	100,0	125	2
4.3	100,0	112.5	2
5.1	107.5	37.5	2
5.2	37.5	100,0	2
5.3	100,0	62.5	3
5.4	147.5	152.5	3
5.5	125,0	27.5	2
5.6	125,0	25,0	2
5.7	102.5	102.5	2

different model shape flying the most loaded the longest distance. Tab. 3 shows gliding length (l_1) and improved glider model gliding length (l_2) for G1-G5 with its subgroups (pairs). Most groups made a functional paper glider in the 4th learning model phase. G5 was the most successful (average l_1 length) in 3rd phase - gray fields and G4 in the 4th phase - gray. Degree of innovativeness is below average ($\bar{x}$= 1.87; s_x = 0.53).

G1 and G4 exhibit understanding difficulties between glider and paper glider model (no additional knowledge in phase 2). Student's solution drafts were issuing from the origami plan. Most G1 and G4 students made glider's fuselage by using a triangular, circular or cone profile. Paper gliders were mostly of two pieces (fuselage and wings). Wings were similar to those of the origami glider. In phase 4, most students made only simple improvements: the position of the load, the shape of the wings, reinforcement of the fuselage etc.

For G2, G3 and G5 (additional knowledge in phase 2) a low level of innovativeness has been observed ($\bar{x}$= 1.88; s_x = 0.60). Paper gliders were similar to the examples. However, it was noticed that students had taken into account given additional knowledge regarding gliders. Improved glider models of these groups have horizontal and vertical stabilizers.

4 Conclusions

Although students exhibit low level of innovativeness a high level of application has been detected. Students are inclined towards finding a trivial solutions. Overview of existing products in the technology field is welcome, the analyse of these products can lead to a fixation of thinking and over reliance on already-produced products [10]. Knowledge gained from the learning material helped students to understand what a glider is, how it is working and what its components are. During the laboratory exercise, students made a little progress in their knowledge. The greatest advancement in knowledge has been unexpectedly made by G2 (additional knowledge without explaining the gliding principle), while G5 showed most decline in measured knowledge by pre/post-test though most additional knowledge was given in the 2nd phase of the proposed learning model. For other groups, progress is minimal, negative, or not at all. The presented model was a novelty for the students. They had the most difficulties in the 3rd (problem) phase of the learning model with the planning sub-phase and further in the 4th (innovate) phase. Students could hardly find more than one improvement that was predominantly a very simple one. Nevertheless concerning also a very short time given for the exercise execution (1,5 hours) obtained results do promote to use the proposed learning model to encouraging innovativeness.

References

[1] M. A. Runco. Creativity, *Annual Review of Psychology*, 55, p. 657-687, 2004.

[2] J. Ardies, S. De Maeyer, D. Gijbels in H. van Keulen. Students attitudes towards technology, *Int J Technol Des Educ*, 25, p. 43-65, 2015.

[3] R. Stobaugh. *Assessing Critical Thinking in Middle and High Schools.* New York, Taylor and Francis, 2013.

[4] L. Liebenberg in E. H. Mathews. Integrating innovation skills in an introductory engineering design-build course, *Int J Technol Des Educ*, 22, p. 93-113, 2012.

[5] D. Cropley in A. Cropley. Recognizing and fostering creativity in technological design education, *International Journal of Technology and Design Education*, 20, 3, p. 345-358, 2010.

[6] E. Linfors in A. Hilmola. Innovation learning in comprehensive education, Int J Technol Des Educ, 26, p. 373-389, 2016.

[7] J. Zang, H.-Y. Hong, M. Scardamalia, C. L, Teo in E. A. Morley. Sustaining knowledge building at a principle-based innovation in an elementary school, *The Journal of the Learning Scieneces*, 20, p. 262-307, 2011.

[8] A. M. Kalimullin in V. V. Utemov. Open Type Tasks as a Tool for Developing Creativity in Secondary School Students, *Interchange*, 48, p. 129-144, 2017.

[9] B. Aberšek. *Did. tehniškega izobraževanja med teorijo in prakso*, Ljubljana, Zavod Republike Slovenije za šolstvo, 2012.

[10] R. McLellan in B. Nicholl. "If I was going to design a chair, the last thing I woud look at is a chair": product analysis and the causes of fixation in students' design work 11-16 years, Int J Technol Des Educ, 21, p. 71-92, 2011.

Študentski članki

Student papers

Izdelava univerzalnega robotskega 4 osnega manipulatorja

Jan Slemenšek

[1]*FERI, Smetanova ulica 17, 2000 Maribor*
E-pošta: jan.slemensek@gmail.com

The making of 4 axis robotic manipulator

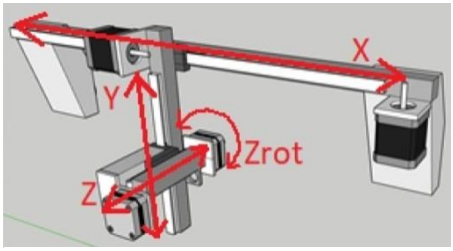
Abstract. *In the last 30 years, robots with it's revolution have shaped the world as we know it today. With every day, robots are getting smarter, faster and more accurate.*

1 Uvod

Roboti so v zadnjih 30-ih letih s svojim razvojem revolucionirali ter oblikovali svet kot ga poznamo danes. Z vsakim dnem postajajo roboti pametnejši, hitrejši ter bolj natančni.

2 Moje delo

Roboti so me od nekdaj fascinirali s svojimi zmogljivostmi.
Že nekaj časa sem si želel izdelati nekakšen robotski manipulator, ob izdelavi katerega bi se veliko naučil, hkrati pa bi ga lahko programiral, da bi manipulator izvajal neko ponovljivo nalogo.
Zamislil sem si 4 osni manipulator, kjer bi imel 3 translacijske osi *(X,Y,Z)* in eno rotacijsko os *(Zrot).*



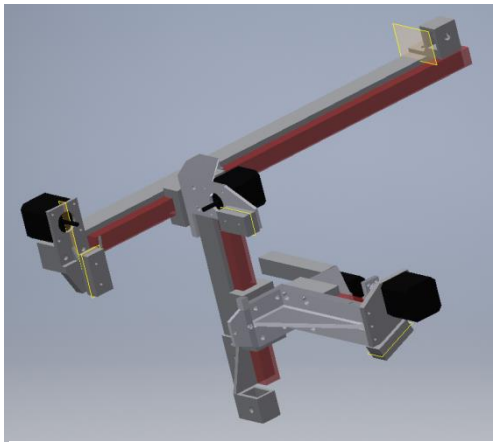
Slika 1: 3D skica manipulatorja

Na os četrtega motorja *(Zrot)* bi lahko nameščal različne "glave" manipulatorja. Zamislil sem si več različnih glav, ki bi omogočile manipulatorju opravljanje različnih nalog. Tako bi lahko izdelal nekakšno prijemalo, s katerim bi prijemal različne izdelke ter manipuliral z njimi.
Lahko bi izdelal glavo za 3D tisk, ter manipulator uporabil za 3D tiskanje, na glavo bi lahko namestil avtomatski spajkalnik ali pa mikroskop.

Najprej sem narisal skico, da sem si lahko lažje predstavljal celoten koncept ter kaj vse bo potrebno izdelati. Dobil sem tudi grob občutek glede potrebnih dimenzij linearnih vodil, jermenov in drugih komponent.

2.1 Izdelava

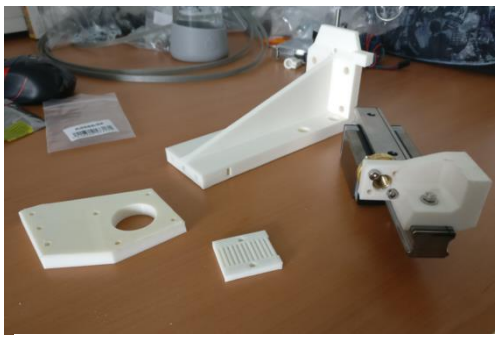
Narisal sem natančen 3D model *(assembly)* mehanizma, ter začel zbirati material.
Odločil sem se, da bom vse potrebne nosilce za mehanizem natisnil na navadnem *FDM* 3D tiskalniku.



Slika 2: Natančen 3D model mehanizma

Obdelovalni volumen bi pri taki geometriji znašal: 450 x 200 x 100 mm *(x,y,z).*
X in Y os uporabljata jermen in pripadajoče jermenice medtem ko Z os spremeni rotacijsko gibanje osi motorja v linearno gibanje drsnika s pomočjo vretena in pripadajoče matice. Korak navoja vretena je 1mm.

Nosilce sem natisnil iz bele *ABS* plastike.



Slika 3: nekaj 3D natisnjenih nosilcev

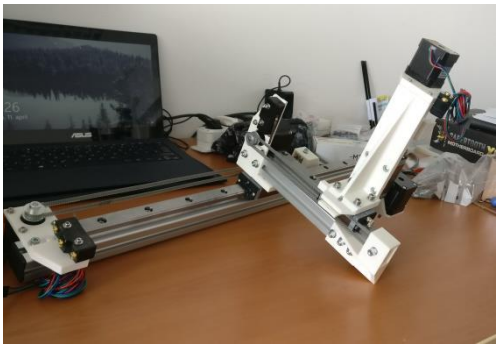
2.2 Sestavljanje

Ko sem prejel ves potreben material, sem lahko pričel s sestavljanjem manipulatorja.
Zaradi potreb po dobri natančnosti in ponovljivosti, sem se odločil za uporabo linearnih tirnih vodil.



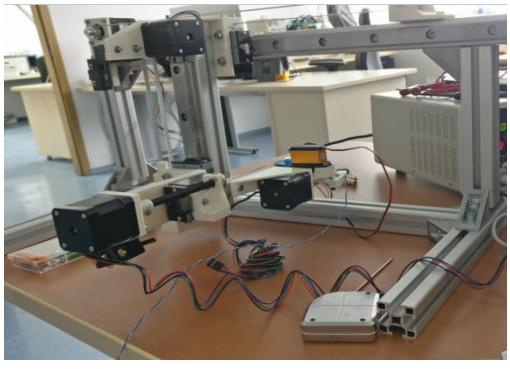
Slika 4: Sestavljena Z os, testiranje.

Uporabil sem koračne motorje *Nema 17*. Ti motorji nimajo povratne informacije o svoji poziciji *(Open-loop)*. Potrebno je uporabiti končno stikalo. Tako lahko po ponovnem zagonu krmilnika, z uporabo končnega stikala najdemo izhodiščno točko. Koračnemu motorju lahko nato posredujemo število korakov, ki naj jih motor opravi v izbrano smer.



Slika 5: Sestavljena konstrukcija

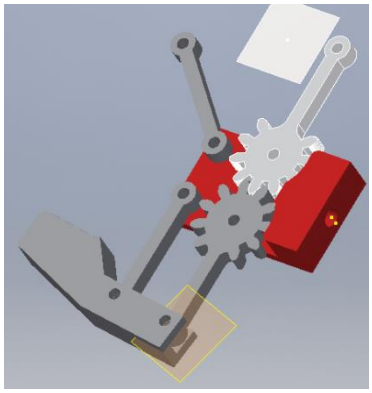
Nosilec za mehanizem je izdelan tako, da je 4. rotacijska os manipulatorja orientirana horizontalno.



Slika 6: Končan manipulator montiran na nosilec

2.3 Prijemalo

Najprej sem izdelal prijemalo za manipulacijo z objekti. Prijemalo je moralo biti zmožno objekt prijeti in spustiti.



Slika 7: 3D model prijemala



Slika 1: Končano prijemalo

Za odpiranje/zapiranje prijemala sem uporabil 6 V servo motor.

Želel sem, da bi prijemalo lahko opravljalo čim več različnih funkcij.

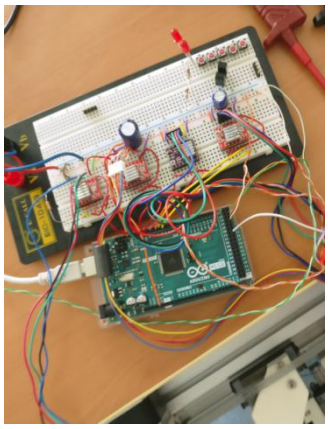
V "prst" prijemala sem namestil še fotoupor, ki omogoča zaznavanje osvetljenosti. S pomočjo fotoupora bi lahko preverjali ali je prijemalo v pravilni poziciji.



Slika 2: Fotoupor v prstu

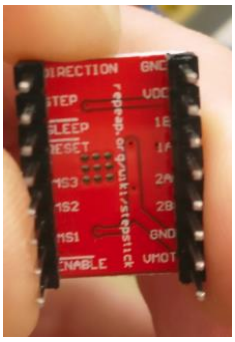
2.4 Električna vezava

Uporabil sem krmilnik *Arduino Mega 2560*, ki upravlja koračne motorje ter servo motor na prijemalu. Uporabljena so tudi končna stikala, ki zagotovijo določitev ničelne točke.
Za krmiljenje koračnih motorjev sem uporabil dodatne gonilnike *A4988*.



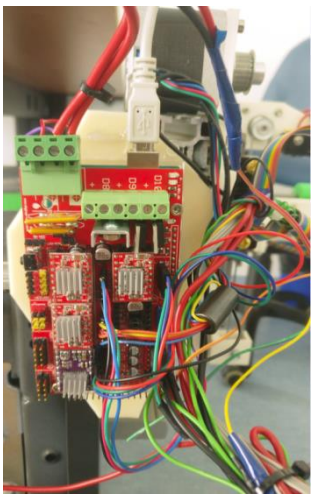
Slika 30: Vezava

Gonilnik se enostavno poveže na arduino preko treh povezav (*DIR, STEP, GND*).
V programu, ki je naložen na krmilnik, lahko enostavno definiramo, kako naj se naš koračni motor giblje. *DIR* definira smer sukanja robota (0V-levo, 5V-desno), medtem ko *STEP* (*PWM signal*) določa, koliko korakov naj motor izvede v prej definirani smeri.



Slika 41: A4988

2.4.1 Ramps 1.4



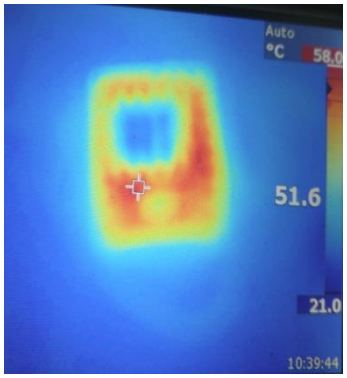
Slika 52: Ramps 1.4

Z osnovno vezavo (*Mega+A4988*) sem potrdil delovanje sistema. Hotel sem še dodatno izboljšati zanesljivost in kompaktnost elektronike z *Ramps 1.4* ploščo.

Ramps 1.4 je razširitvena plošča za *Arduino Mega*, namenjena gonilnikom koračnih motorjev. Najpogosteje se uporablja za krmilje 3D tiskalnikov, priklopi se direktno na *Arduino ploščo*.
Podpira priklop do pet koračnih motorjev, hkrati pa ima veliko število različnih vhodov (analog, digital) ter izhodov (3 x 12V tranzistor). Gonilniki *A4988* se priklopijo na *Ramps 1.4* (*slika 12)*.

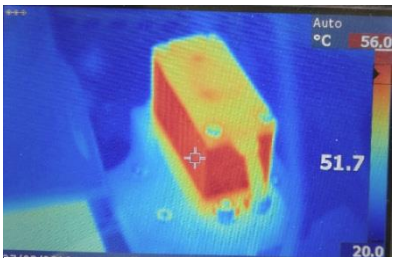
2.5 Termična stabilnost mehanizma

Ob daljšem delovanju sistema, se je gonilnik *A4988* Y osi (nosi največjo težo), pregrel do termične zaščite, ki je vgrajena v čipu, tako, da je prenehal delovati (dokler je zaščita vklopljena).
Problem je nastal zaradi prevelikih koračnih motorjev, ki so imeli nizko notranjo upornost, posledično pa so porabili preveč amperov iz gonilnika. Zamenjal sem Y koračni motor za enak, vendar manjši motor. Zamenjava motorjev je zmanjšala temperaturo gonilnika iz 110 °C na 60 °C.



Slika 63: Termična slika gonilnika

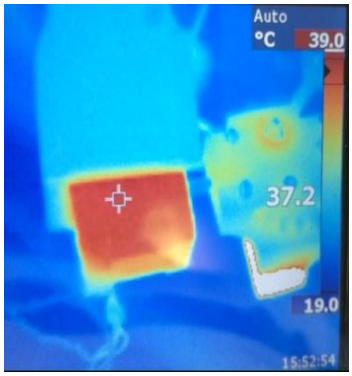
Servo motor na prijemalu se je ob navadnem delovanju na 6 V segrel na 76 °C. To temperaturo lahko zmanjšamo, da nekoliko znižamo napajalno napetost do točke, ko servo motor še lahko normalno opravlja nalogo.



Slika 14: Termična slika servo motorja

Po zmanjšanju napajalne napetosti iz 6 V na 4,8 V, se je servo motorju temperatura zmanjšala za 25 ºC.

Nema 17 koračni motorji se segrejejo do 40 °C, medtem ko se Y motor zaradi drugačne izvedbe (manjši motor, enak gonilnik) segreje do 65 °C pri polni obremenitvi.



Slika 157: Termična slika Nema17 motorja
17 motorja

2.6 Program

Program je napisan v programskem jeziku C. Za nadzor koračnih motorjev se uporablja knjižnica *"AccelStepper"*, ki omogoča enostavno definicijo maksimalne hitrosti, pospeškov ter pojemkov. S to knjižnico lahko istočasno vodimo več motorjev.

Napisal sem program, kjer lahko vpišemo koordinate (v milimetrih), robotski mehanizem pa se pomakne v želene koordinate.

Void loop zanka programa je sledeča:

```
void loop() {
  while(Serial.available() > 0) // Don't read unless avaliable
  {
     TravelX=(Serial.parseInt()*64); //multiplied for input in mm.
     Serial.println(TravelX/64);
     TravelY=(Serial.parseInt()*64); //multiplied for input in mm.
Serial.println(TravelY/64);
     TravelZ=(Serial.parseInt()*156);
     Serial.println(TravelZ/156);
     TravelZr=(Serial.parseInt()); //3200=360deg, 1600=180deg//
     Serial.println(TravelZr);
//////////////////////////////LIMITS////////////////////////////
        if (TravelX>20000){TravelX=20000;}  //X stepper limits
           else if(TravelX<0) {TravelX=0;}
        if (TravelY>7000){TravelY=7000;}    //Y stepper limits
           else if(TravelY<0) {TravelY=0;}
        if (TravelZ>153615){TravelZ=153615;}  //Z stepper limits
          else if(TravelZ<0) {TravelZ=0;}
        if (TravelZr>153615){TravelZr=153615;}  //Zr stepper limits
           else if(TravelZr<0) {TravelZr=0;}
//////////////////////////////////////////////////////////////////
   stepperX.moveTo(TravelX);  // Set new move position for X Stepper
   stepperY.moveTo(TravelY);  // Set new move position for Y Stepper
   stepperZ.moveTo(TravelZ);  // Set new move position for Z Stepper
   stepperZr.moveTo(TravelZr);//Set new move position for Zr Stepper
}
  if ((stepperX.distanceToGo() != 0) || (stepperY.distanceToGo() !=0)
|| (stepperZ.distanceToGo() !=0) || (stepperZr.distanceToGo() !=0)) {
        stepperX.run();  // Move Stepper X into position
        stepperY.run();  // Move Stepper Y into position
        stepperZ.run();  // Move Stepper Z into position
        stepperZr.run();  // Move Stepper Zr into position
       }
}
```

Program omogoča vodenje mehanizma preko vpisanih koordinat (X,Y,Z,Zr). Koordinate vpišemo v serijski vmestnik v okolju Arduino. Istočasno lahko premikamo vse štiri osi. Uporabljen je absolutni koordinatni sistem. Koordinate vpisujemo po naslednjem zgledu:

```
0,0,0,0                 //x=0mm; y=0mm; z=0mm; zr=0º;
100,56,75,1800          // x=100mm; y=56; z=75; zr=90º;
100,70,75,0             //premakneta se samo Y os ter Zr.
```

3 Zaključek

Manipulator je dokončan. Sledi še izdelava različnih "glav", ki bi mehanizmu omogočale opravljanje različnih funkcij.

Razlaga

FDM: (Fused deposition modelling), navadni 3D tiskalniki, ki talijo plastično žico.
Assembly: skupek 3D modeliranih delov zloženih skupaj, da formirajo skupno geometrijo,
ABS: tip plastike, ki se uporablja pri FDM 3D tiskalnikih,
Open-loop: Odprto-zančni način krmiljenja.

Literatura

[1] https://reprap.org/ (citirano 1.7.2018)

[2] https://www.arduino.cc/en/Main/ArduinoBoardMegaADK (citirano 1.7.2018)

[3] https://www.pololu.com/product/1182 (citirano 1.7.2018)

Štetje in merjenje hitrosti prometa z dostopnim mikrovalovnim senzorjem

Gal Pavlin

Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1000 Ljubljana
gp4228@student.uni-lj.si, gal@pavlin.si

Affordable speed and traffic flow sensing using microwave sensor module

Abstract. *Radar sensor modules have become very affordable. With the right choice of signal processing electronics good enough results can be expected without hurting the system's affordability. Possible use cases range from room activity sensing to measuring traffic speed and flow.*

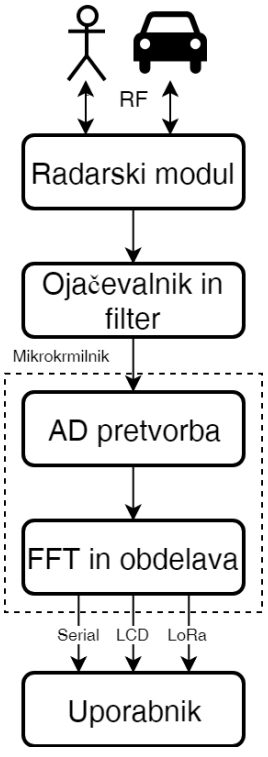
1 Uvod

CDM324 je majhen mikrovalovni radarski modul. Ne večji kot 3x3 cm je z nekaj dodatne elektronike zmožen zaznati premikanje in hitrost večjih objektov v njegovem dosegu.

Uporabnih vrednosti je veliko. Pokazal bom, da je radar uporaben za merjenje aktivnosti v zaprtem prostoru, merjenje hitrosti ter štetje prometa.

2 Strojna oprema

Slika 1 prikazuje potek signalov v celotnem sistemu.



Slika 1. Blokovna shema celotnega sistema

2.1 Radarski modul

Čeprav je izvor CMD324 modula vprašljiv, ga je mogoče kupiti na znanih spletnih avkcijah. Modul je po funkcionalnosti in izgledu precej podoben modulu IMP-165 nemškega podjetna Innosent [1].

Senzorski modul vsebuje mikrovalovni oddajnik in sprejemnik, ki delujeta na 24 Ghz [2]. Signal se konstantno oddaja in sprejema preko dveh mikrostrip anten na sprednji strani. Izhod iz modula je analogni signal, ki je frekvenčna razlika oddanega in sprejetega signala. Če se pred modulom dovolj velik objekt premika, nastane med oddanim in sprejetim signalom frekvenčna razlika, ki je posledica frekvenčnega premika zaradi Dopplerjevega pojava [3][4].

Radarski modul potrebuje za delovanje 5V pri porabi 60mA [2]. Napajanje mora biti čim bolj čisto, saj se vsaka motnja na napajanju direktno preslika na izhod. Po navadi so motnje v napajanju periodične, kar povzroči napačne odčitke ali pa močno dvigne zaznavni prag oziroma zniža občutljivost.

Antena na modulu je usmerjena. Sevalni diagram je sploščen stožec s kotom 80° vodoravno in kotom 35° navpično. [2] Zaradi dokaj široke antene je zaznavanje majhnih objektov na daljše razdalje zahtevno. Z mojo implementacijo modula v sistem sem lahko zaznal in izmeril hitrost osebnega avtomobila na približno 15m.

2.2 Ojačevalnik signala

Analogni izhod iz modula je potrebno pred digitalno obdelavo ojačiti. Za dizajn ojačevalnika sem uporabil spletno orodje WEBENCH® Filter Designer podjetja Texas Instruments.

Ojačevalno vezje ima dve stopnji, ki sta realizirani s pomočjo dvojnega operacijskega ojačevalnika. Vezje hkrati služi tudi kot nizkoprepustni filter, ki odreže signale višjih frekvenc, saj nas ti pri obdelavi ne zanimajo. Orodje se je izkazalo za zelo uporabno in ojačevalnik je deloval zadovoljivo dobro že ob prvi implementaciji. [5]

Zaradi fiksnega ojačanja pride pri velikih vhodnih signalih ojačevalnik v nasičenje. To se zgodi npr., ko je velik objekt zelo blizu radarja. V frekvenčnem spektru je poleg glavne komponente še več harmonikov, ki zasenčijo ostal uporaben spekter. To težavo bi lahko odpravil z večkanalnim ojačevalnikom z različnimi ojačanji, pri čemer bi izbiral najbolj optimalnega. To bi lahko rešil tudi z avtomatskim uravnavanjem ojačanja, kot to počnejo nekateri mikrofonski ojačevalniki. [7]

2.3 Mikrokrmilnik

Mikrokrmilnik služi kot enota za obdelavo radarskega signala. Izbral sem 32-bitni STM32F4 podjetja ST micro. Ta družina mikrokrmilnikov se ponaša s hitrim AD pretvornikom in FPU enoto, ki zelo pohitri izvajanje matematičnih operacij za obdelavo signala.

Naloga mikrokrmilnika je, da izvaja AD pretvorbo in zajete podatke v časovnem prostoru s pomočjo hitre fourierjeve transformacije pretvori v podatke v frekvenčnem prostoru. V slednjem je obdelava signala bistveno lažja. [3] Računanje poteka v realnem času.

Tabela 1. Parametri zajema in obdelave s FFT

Opis	Parameter	Vrednost
Frekvenca zajema	f_s	20 Khz
Velikost FFT	FFT_{SIZE}	2048 vzorcev
Ločljivost frekvence	Δf	19.5 Hz

Zajem podatkov in računanje FFT potekata istočasno. S t.i. dvojnim predpomnilnikom se doseže, da se zajame in pretvori celoten signal. Slabost je, da se v sistem doda časovni zamik, ki je enak zajemu ene velikosti FFT. V mojem primeru je ta zamik okrog 100 ms.

Poleg zajema in pretvorbe signala je naloga mikrokrmilnika tudi, da iz dobljenih podatkov izloči bistvene komponente. To je v mojem primeru bodisi hitrost objekta pred radarjem ali pa aktivnost v prostoru, ki sem jo določil kot oceno energije celotnega spektra odbitega signala.

Pri pisanju programske kode sem se trudil čim bolj uporabljati periferijo. Uporabil sem interni DMA (Direct Memory Access), ki poskrbi, da se zajeti podatki iz internega AD pretvornika zapišejo v delovni spomin brez posredovanja procesorja. Na ta način zajem podatkov z vzorčenjem 20 kHz ni povročal nobenih težav. Vzorčenje bi lahko še povečal in izboljšal ločljivost, vendar se to ni izkazalo za potrebno.

3 Izdelava sistema

3.1 Vezje

Prvotna ideja je bila, da bi se celoten radar napajal iz ene litijeve celice. Vezje je načrtovano s polnilcem, ki polni baterijo iz USB napajanja. Za zagotovitev potrebnih napetostnih nivojev je na vezju stikalni pretvornik, ki napetost baterije (3-4.2 V) pretvori na 5 V. Za njim je linearni regulator (LDO), ki napetost zniža še na 3.3 V za napajanje mikrokrmilnika.

Za preprosto merjenje frekvence sem na vezje dodal komparator, ki analogni sinusni signal pretvori v kvadratnega z isto periodo. Ta signal je nato speljan na merilec periode (timer) v mikrokrmilniku. S to metodo mi ni uspelo dobiti zadovoljivih rezultatov, kljub implementaciji histereze. Razlog za to je bil prevelik šum v signalu. Direktno vzorčenje z AD pretvornikom in izračun FFT se je izkazalo za mnogo bolj zanesljivo metodo določanja frekvence z najvišjo amplitudo.



Slika 2: Tiskano vezje z elementi

Na vezju je tudi večsmerna navigacijska tipka, ki služi za premikanje po menijih prikazovalnika. Z njo lahko uporabnik nastavlja merilno območje in parametre prikaza. Uporablja se tudi za vklop in izklop.

Na sliki 2 je prvo izdelano vezje. Radarski modul se nahaja na spodnji strani. Usmerjenost antene modula kaže skozi vezje navzdol. Na sliki 2 je v zgornjem levem kotu videti žične prevezave, ki so posledica nepravilne vezave navigacijske tipke.

Vezje sem cinil ročno s spajkalnikom s fino konico. Elemente brez nogic sem zaspajkal z vročim zrakom.

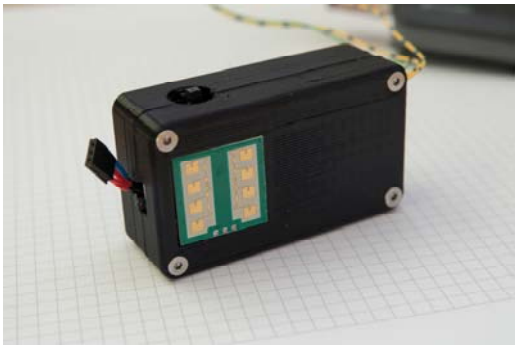
3.2 Problemi

Imel sem težave z motnjami na napajanju. Stikalni boost pretvornik je v kombinaciji z radarskim modulom na 5 V napajanju povzročal periodične špice velikosti do 500 mV, kar je močno dvignilo zaznavni prag za celoten sistem. Težavo mi je uspelo delno omejiti s filtriranjem napajanja, skoraj dokončno odpraviti pa tako, da sem uporabil zunanje USB napajanje. Zaradi tega radar na koncu ni bil več avtonomen in je za delovanje potreboval zunanje napajanje. V naslednji iteraciji načrtovanja vezja bom skušal ta del izboljšati in motnje na napajanju odpraviti, ne da bi spremenil prvotno zamisel avtonomnega delovanja radarja.

3.3 Ohišje

Radarski modul se skupaj s pripadajočim vezjem nahaja v plastičnem ohišju. Ohišje je sestavljeno iz dveh delov, ki sta med seboj povezana s štirimi vijaki. Na vrhu ohišja je odprtina za navigacijsko tipko. Ob straneh je odprtina za USB napajanje, konektor za programiranje in konektor za serijska vrata. Ohišje je bilo natisnjeno s 3D tiskalnikom v Makerlab Ljubjana.

Na sliki 3 je druga verzija ohišja. Prva je bila približno dvakrat ožja in porezkana iz lesa. Zaradi pomanjkanja prostora in nepraktičnosti sem to verzijo opustil. To verzijo ohišja je videti na sliki 2. Manjka še aluminijast pokrovček z zadnje strani z odprtino za manjši prikazovalnik.



Slika 3: Ohišje s sprednjo odprtino za radarsko anteno

3.4 Prikaz in povezljivost

Za prikaz radarskih podatkov v živo služi grafični OLED prikazovalnik velikosti 2.42 palca na zadnji strani. Na prikazovalniku se izpisuje izmerjena hitrost, če se pred radarjem objekt premika. Če v okolici ni premikajočih objektov, radar z napisom »Low signal« sporoči, da ni odbitega signala. Na spodnji strani prikazovalnika se v živo izrisuje graf komponent signala v kompleksnem spektru. Za grob odčitek iz grafa je na voljo tudi skala. Z visoko frekvenco osveževanja (50Hz) je slika zelo tekoča.



Slika 4: OLED prikazovalnik v delovanju

Z večsmerno navigacijsko tipko lahko uporabnik upravlja z grafom na prikazovalniku. Lahko ga razteguje in krči ter premika po časovni osi, tako da menja merilno območje. Če uporabnik tipko kratko pritisne, se na prikazovalniku pokaže meni, kjer se lahko nastavlja tako parametre za FFT kot parametre algoritma za izračun hitrosti. Uporabnik lahko tudi izbira enote za prikaz izmerjene hitrosti in vključi oziroma izključi prikaz aktivnosti pred radarjem.

Radar se lahko poveže preko serijskega vodila s TTL nivoji na drugo napravo. Sprogramiral sem ga tako, da vsake tri sekunde pošlje najvišjo izmerjeno hitrost in povprečno aktivnost. Na drugi strani serijskih vrat lahko preprosto priključimo osebni ali »System On Chip« računalnik s programsko opremo, ki beleži poslana sporočila in jim doda časovno oznako. Sistem je s tem pripravljen na beleženje hitrosti.

3.5 Možne izboljšave

Namesto direktne žične povezave bi lahko uporabil brezžično LoRa povezljivost. Kot modul se priključi na radar preko serijskih vrat in na spletni strežnik pošilja radarske podatke. Na spletu se lahko podatki naknadno obdelajo in prikažejo v obliki primerni za širšo publiko. Z dodatkom solarnega napajanja bi lahko sistem postavil kamorkoli v mestu, kjer je dovolj sonca. To bi močno poenostavilo uporabo radarja na mestih, kjer ni dostopne električne in internetne infrastrukture.

4 Rezultati

4.1 Računanje hitrosti in ocena aktivnosti v prostoru

Hitrost objekta, ki se premika pred radarjem, se določi s formulo 1. [4] [6] Ločljivost hitrosti je pri izbranem Δf 0.12 m/s.

$$v = f_{dop} \frac{c_o}{2 f_0 \cos(\alpha)} \quad (1)$$

Tabela 2. Opis veličin v enačbi (1)

Oznaka	Veličina
f_{dop}	Frekvenčni zamik
f_0	Oddajna frekvenca (24.125 Ghz)
c_0	Hitrost svetlobe
α	Kot med hitrostjo objekta in usmerjenostjo mikrovalov

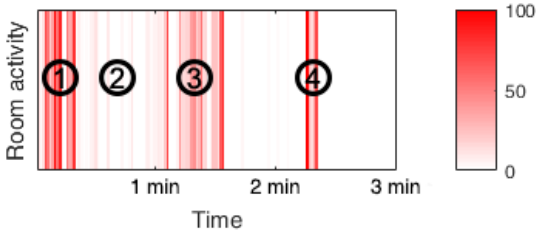
Aktivnost okolice pred radarjem se lahko oceni s povprečno vrednostjo signala v času zajema podatkov za eno velikost FFT. Vse prispevke signala v frekvenčnem prostoru med seboj geometrijsko seštejemo s formulo (2).

$$U = \sqrt{\sum_{i=i}^{FFT_{size}} U_i^2} \quad (2)$$

4.2 Merjenje aktivnosti v zaprtem prostoru

Radar sem postavil v prostor na mizo in ga povezal na računalnik preko serijskih vrat. Računalnik je beležil 3 sekundno povprečje signala. Meritev je potekala 3 minute. V tem času sta se dve osebi premikali po prostoru ali sedeli pri miru.

Zajete podatke sem obdelal v Matlab-u in iz njih sestavil toplotno karto na sliki 2. Enote so relativne glede na najvišjo izmerjeno vrednost.



Slika 5: Toplotna karta aktivnosti v prostoru

Tabela 3. Opis dogajanja v prostoru za sliko 2

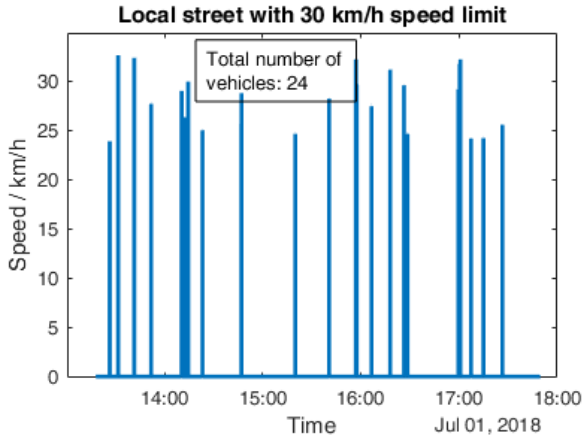
Št.	Dogajanje v prostoru
1	Dve osebi prideta v prostor.
2	Osebi sedita.
3	Ena oseba hodi po prostoru in ga nato zapusti
4	Prostor zapusti še druga oseba

4.3 Merjenje hitrosti in štetje prometa

Algoritem po vsakem končanem izračunu FFT najprej v frekvenčnem prostoru poišče komponento z najvišjo amplitudo. Nato preveri, če je ta komponenta višja od nastavljene spodnje vrednosti, da se izloči šum. Če je, se izračuna hitrost po enačbi (1), s tem da se za kot α vzame 0. Kot med hitrostjo objekta in umerjenostjo mikrovalov je potrebno upoštevati kasneje. Pri algoritmu je še veliko prostora za izboljšave.

Za prikaz izmerjenih vrednosti sem uporabil tekoče povprečje zadnjih štirih izmerjenih vrednosti. Na ta način sem odpravil skakanje prikaza. Rezultat je zaokrožen na dve decimalni mesti.

Radar sem postavil ob cesto. Zaradi širokega snopa oddajne in sprejemne antene je efektivni domet omejen na približno 15 m za meritev hitrosti osebnega avtomobila. Zaradi kratkega dometa, sem radar obrnil približno 40° glede na smer vozil. Meritev se je izvajala 3 ure na nedeljsko popoldne, zato je bilo naštetih avtomobilov le 24.



Slika 6: Graf hitrosti prometa za 1.7.2018

5 Zaključek

Uspelo mi je pokazati, da je z malo iznajdljivosti za preproste meritve uporaben tudi kitajski radarski modul vprašljivega porekla. To sem dosegel z uporabo ustreznega algoritma in skrbno načrtovanim vezjem. Z IoT integracijo in solarnim napajanjem bi lahko sistem postal uporaben za merjenje hitrosti in gostote prometa v mestih ali kot npr. učni pripomoček za poučevanje v srednjih tehniških šolah.

Literatura

[1] Limpkin's blog, 11.7.2018 https://www.limpkin.fr/index.php?post/2017/02/22/Making-the-Electronics-for-a-24GHz-Doppler-Motion-Sensor

[2] Data Sheet IPM-165, 11.7.2018 https://www.innosent.de/fileadmin/media/dokumente/DATASHEETS_2016/Datenblatt_IPM-165_V8.5.pdf

[3] Li S., Wang Y. (2012) The Design of Radar Signal Processor Based on FPGA. In: Zhang Y. (eds) Future Wireless Networks and Information Systems. Lecture Notes in Electrical Engineering, vol 143. Springer, Berlin, Heidelberg

[4] Innosent Application note I, 11.7.2018 https://www.innosent.de/fileadmin/media/dokumente/Downloads/Application_Note_I_-_web.pdf

[5] WEBENCH® Filter Designer, 11.7.2018 http://www.ti.com/design-tools/signal-chain-design/webench-filters.html

[6] K. Ma, K. Zhang and Y. Zhang, "A signal processing practice method based on LFMCW radar," 2017 6th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, 2017, pp. 449-452.

[7] L. González-Delgado, D. Valencia-Redrován, V. Robles-Bykbaev, N. González-Delgado and T. Panzner, "Fuzzy controller for automatic microphone gain control in an autonomous support system for elderly," 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom), Natal, 2014, pp. 77-81.

Potek napake funkcije atan2 ob popačenju vhodnih signalov

Mitja Ališ

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana
E-pošta: mitja1357@gmail.com

The course of error due to distortion of input signal to the atan2 function

Distortion of sine and cosine values, used for angle determination with the atan2 function, can result in numerical error. According to the performed review of literature, error is normally presented by taking only the basic harmonic into account. This paper however presents determination of error by also taking into account higher harmonics, which are non-negligible at larger distortion of sine or cosine. Error is going to be expressed with infinite series, which expand the domain of distortion parameter.

1 Uvod

Dandanes je potreba po kakovostni regulaciji elektromotorskih pogonov neizogibna in tako prisotna v številnih aplikacijah. Za kakovostno nadzorovanje rotirajočega pogona se v povratnih zankah uporablja senzorje zasuka [1]. Zasuk se meri na različne principe: inkrementalni dajalnik, resolver, enkoder [2][3][4]... Izhoda enkoderja ali resolverja sta signala sinus in kosinus. Za izračun informacije o zasuku se uporabi funkcijo inverz tangensa [8]. V aplikacijah se običajno uporablja funkcija atan2, katere zaloga vrednosti je $[-\pi, \pi]$ [9].

Senzorji niso idealni in posledično je iluzorno pričakovati delovanje povsem brez napake. Primer napake je popačena oblika zajetih signalov sinus in kosinus. Različne literature [5] [6] [7] opisujejo, kako popačenje vhodnih signalov vpliva na napako. Napako so izrazili le z najočitnejšimi harmoniki. V napaki nastopajo tudi višji harmoniki, ki so lahko predstavi kot napako v obliki neskončne vrste. V raznih aplikacijah je možen dostop le do izhodnega podatka o zasuku. Z referenčnim dajalnikom, se lahko določi napako, njen izvor ostaja neznan.

S poznavanjem vplivov popačenja signalov sinus in kosinus na napako se lahko kasneje iz napake oceni kaj nanjo vpliva in kako jo zmanjšati. V tem delu bom predstavil kako se pri popačenih vhodnih signalih sinusa in kosinusa izraža napaka izračunanega kota v obliki neskončne vrste.

2 Metodologija in rezultati

Osnovna vhodna signala v funkcijo sta sinus in kosinus. Popačena sta lahko na več načinov:

$$Sin = B_0 + B_1 \sin(\theta + \varphi_s) \tag{1}$$

$$Cos = A_0 + A_1 \cos(\theta + \varphi_c) \tag{2}$$

Vsebujeta lahko različne amplitude, fazne zamike ali enosmerne komponente. Idealno so v (1) in (2), B_0, A_0, φ_s in φ_c, enaki 0, B_1 in A_1 sta enaka. Signaloma se lahko zaradi enega parametra spremeni tako amplituda kot faza. To se zgodi ob upoštevanju enega od parametrov v (3) in (4).

$$Sin = \sin(\theta) + \Delta_c \cos(\theta) + \Delta_s \sin(\theta) \tag{3}$$

$$Cos = \cos(\theta) + \Delta_c \cos(\theta) + \Delta_s \sin(\theta) \tag{4}$$

Definicija merjenega kota φ in napake ε je:

$$\varphi = \mathrm{atan2d}(Sin, Cos) \tag{5}$$

$$\varepsilon = \varphi - \mathrm{atan2d}(\sin(\theta), \cos(\theta)) \tag{6}$$

pri čemer θ predstavlja referenco. atan2d() je funkcija, ki je že privzeto definirana v programskem okolju MATLAB in je namenjena izračunu vrednosti arkus tangensa v štirih kvadrantih. Izhod funkcije je podan v kotnih stopinjah [10]. Napaka bo predstavljena v obliki neskončne vrste (7).

$$\varepsilon(x) = C_0(x) + \sum_{n=1}^{\infty} C_n(x) \sin(n\Theta + \varphi_n(x)) \tag{7}$$

Parameter x predstavlja neodvisno spremenljivko, katera popači vhodna signala sinus in kosinus. C_0 ki predstavlja enosmerno komponento napake, C_n amplitudo posameznega člena vrste ter φ_n fazni zamik harmonika napake. Vsi členi so odvisni od x.

V nadaljevanju je predstavljen eden od možnih načinov določanja napake. Izbran parameter (A_0, B_0, A_1, B_1, φ_s, φ_c, Δ_s, Δ_c), ki sem ga upošteval pri napaki, sem limitiral v neskončnost. Pri limitiranju parametra v neskončnost se napaka izrazi v obliki npr:

$$\varepsilon = \begin{cases} 90^\circ - \theta, & \theta \in \{0^\circ, 180^\circ\} \\ 270^\circ - \theta, & \theta \in \{180^\circ, 360^\circ\} \end{cases} \tag{8}$$

Napako se razvije v Fourierovo vrsto in s tem ugotoviti konvergenco amplitud ter katere harmonike je potrebno opazovati. Sledi iskanje analitične funkcije, ki opisuje potek amplitude harmonika ter faznega zasuka posameznega harmonika ob spremembi parametra. Raziskave so pokazale, da so višji harmoniki potence osnovnega harmonika napake. Napaka se izrazi kot potenčna vrsta.

V nadaljevanju so predstavljeni principi določanja napake ob različnih amplitudah, faznih zamikih, enosmernih komponentah ter kombinaciji različnih amplitud ter faz zaradi parametra Δ_s ali Δ_s.

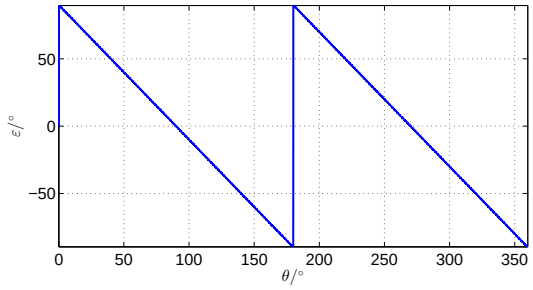
2.1 Določanje napake ob različnih amplitudah

Definirajmo razmerje amplitude sinusa in kosinusa s k. Sprememba obeh amplitud za določen faktor ne bo ustvarila dodatne napake.

$$Sin = k\sin(\theta) \tag{9}$$
$$Cos = \cos(\theta) \tag{10}$$

Ob povečanju k proti neskončnosti se pojavi napaka (slika 1),

$$\lim_{k\to\infty}(\operatorname{atan2}(k\sin\theta,\cos\theta) - \operatorname{atan2}(\sin\theta,\cos\theta)) \tag{11}$$



Slika 1: Napaka ε ob limiti k proti neskončnosti

ki se v Fourierovi vrsti izrazi kot:

$$\varepsilon = \frac{180}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin 2n\theta. \tag{12}$$

Največji vpliv ima drugi harmonik, njegov potek v odvisnosti od k je predstavljen na sliki 2 kot C_1, saj nastopajo le sodi harmoniki.

Z uporabo Curve Fitting Toolbox se poteku najbolje prilega racionalna funkcija. Napaka se izrazi s potenčno vrsto:

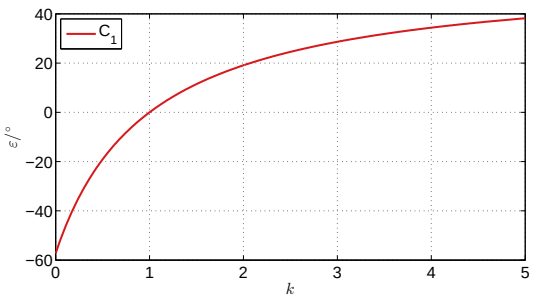
$$\varepsilon(k) = \frac{180}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}(\frac{k-1}{k+1})^n\sin 2n\theta \tag{13}$$

in velja, če je k večji od 0.

$$k \geq 0$$

V (13) namesto k vstavimo razmerje amplitud:

$$\varepsilon(k) = \frac{180}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}(\frac{B_1 - A_1}{B_1 + A_1})^n\sin 2n\theta \tag{14}$$



Slika 2: Potek drugega harmonika v odvistnosti od k

kar velja pri pogoju:

$$\frac{B_1}{A_1} \geq 0.$$

2.2 Določanje napake ob spremembi faznega zamika

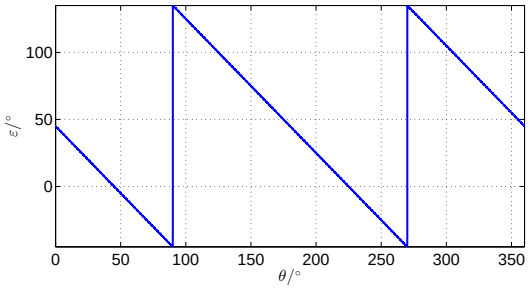
Vhodna signala imata naslednjo obliko:

$$Sin = \sin(\theta + \varphi_s) \tag{15}$$
$$Cos = \cos(\theta + \varphi_c) \tag{16}$$

Napako se določi posamično za vsakega od parametrov. Drugi je takrat enak 0. Na koncu se enačbi združi. Za določanje limite ni potrebno iti proti neskončnosti, ampak le do najslabše možnosti, ki je pri $\pm 90°$:

$$\varepsilon = \lim_{\varphi_s\to 90°}\operatorname{atan2}(Sin, Cos) - \operatorname{atan2d}(\sin(\theta),\cos(\theta)) \tag{17}$$

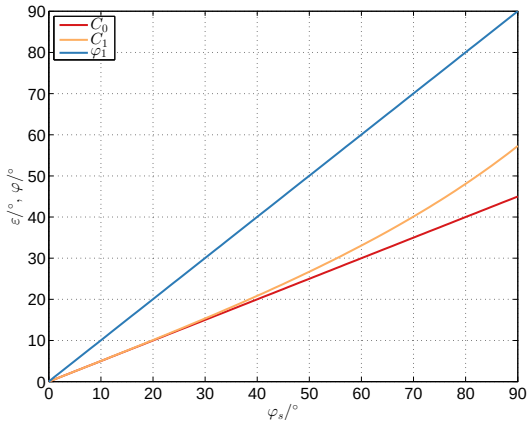
Potek napake ε s slike 3 predstavi vrsta (18).



Slika 3: Napaka ε ob limiti $\varphi_s \to 90°$

$$\varepsilon = 45° - \frac{180}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin(2n\theta) \tag{18}$$

Iz izraza je vidno nastopanje enosmerne komponente in sodih harmonikov. Na sliki 4 so prikazani: potek enosmerne komponente C_0, potek amplitude drugega harmonika C_1 in φ_1 fazni zamik drugega harmonika glede na kosinusno obliko. Za predstavitev poteka s sinusi je potrebno prišteti še $90°$. Ordinatna skala grafa na sliki 4 je v stopinjah. Pri C_0 in C_1 stopinje predstavljajo amplitudo napake, pri φ_1 skala prikazuje fazni zamik v stopinjah. Poteku enosmerne komponente se najbolje prilega funk-



Slika 4: Potek enosmerne komponente napake C_0, amplitude drugega harmonika C_1 in faznega zamika φ_1 glede na kosinus, v odvisnosti od faznega zamika sinusa φ_s

cija premice. Potek amplitude drugega harmonika napake ima obliko funkcije tangens. Fazni zamik drugega harmonika narašča linearno, vendar je za predstavitev s sinusi potrebno prišteti še $90°$. Enako se lahko ponovi za fazni zamik kosinus signala. Enačba končnega poteka se glasi:

$$\varepsilon(\varphi_s, \varphi_c) = \frac{\varphi_s + \varphi_c}{2} + \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (\tan\frac{\varphi_s - \varphi_c}{2})^n \sin(2n\theta + n(90° + \varphi_s + \varphi_c)) \tag{19}$$

pri čemer, (19) velja za :

$$\varphi_s - \varphi_c \in [-90°, 90°]$$

2.3 Določanje napake ob spremembi enosmerne komponente

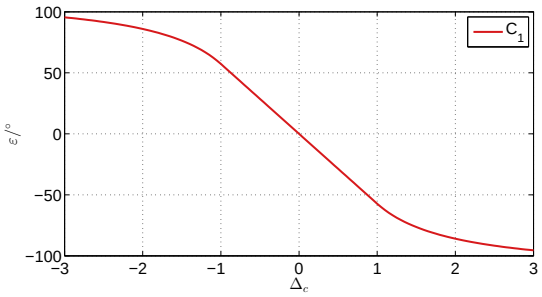
Amplitudi signalov sinus in kosinus sta enaki 1. Z limito A_0 in razvojem napake v Fourierovo vrsto napako izrazimo z:

$$\varepsilon = \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{2}{n} \sin(n\theta + 90°n). \tag{20}$$

Enosmerne komponente ni, največji je prvi harmonik. Potek s slike 5 se razdeli na 3 dele in aproksimira napako z naslednjim izrazom, upoštevajoč tudi enakosti amplitud:

$$\varepsilon(A_0) = \begin{cases} \frac{180}{\pi} \sum_{n=1}^{\infty} (-1)^n \frac{1}{n} (2 - |\frac{A_0}{A_1}|^{-n}) \sin(n\theta), & \frac{A_0}{A_1} \le -1 \\ \frac{180}{\pi} \sum_{n=1}^{\infty} (-1)^n \frac{1}{n} (\frac{A_0}{A_1})^n \sin(n\theta), & |\frac{A_0}{A_1}| \le 1 \\ \frac{180}{\pi} \sum_{n=1}^{\infty} (-1)^n \frac{1}{n} (2 - (\frac{A_0}{A_1})^{-n}) \sin(n\theta), & \frac{A_0}{A_1} \ge 1 \end{cases} \tag{21}$$

Enako se lahko stori tudi za B_0 (22) in za aproksimacijo napake, ko imata sinus in kosinus enako enosmerno komponento (23).



Slika 5: Potek amplitude prvega harmonika od enosmerne komponente kosinus signala, pri amplitudah enakih 1

$$\varepsilon(B_0) = \begin{cases} \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (2 - |\frac{B_0}{B_1}|^{-n}) \sin(n\theta - 90°n), & \frac{B_0}{B_1} \le -1 \\ \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (\frac{B_0}{B_1})^n \sin(n\theta + 90°n), & |\frac{B_0}{B_1}| \le 1 \\ \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (2 - (\frac{B_0}{B_1})^{-n}) \sin(n\theta^\circ + 90n), & \frac{B_0}{B_1} \ge 1 \end{cases} \tag{22}$$

$$\varepsilon(A_0, B_0 = A_0) = \begin{cases} \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (2 - |\sqrt{2}\frac{A_0}{A_1}|^{-n}) \sin(n\theta + 90°n), & \frac{A_0}{A_1} \le -\frac{\sqrt{2}}{2} \\ \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (\sqrt{2}\frac{A_0}{A_1})^n \sin(n\theta - 90°n), & |\frac{A_0}{A_1}| \le \frac{\sqrt{2}}{2} \\ \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (2 - (\sqrt{2}\frac{A_0}{A_1})^{-n}) \sin(n\theta - 90°n), & \frac{A_0}{A_1} \ge \frac{\sqrt{2}}{2} \end{cases} \tag{23}$$

2.4 Različni amplitudi in fazi v odvistnosti od enega parametra

Fourierova vrsta limite (6) pri upoštevanju (3) in (4), ko gre Δ_c v neskončnost, je

$$\varepsilon = 45° - \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin(2n\theta). \tag{24}$$

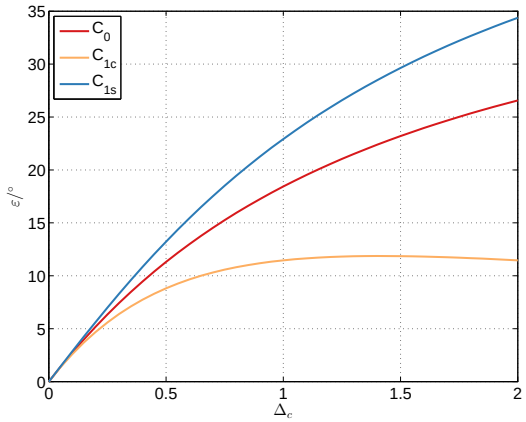
Potek drugega harmonika od Δ_c se lahko zapiše kot vsota sinusa in kosinusa.

$$C_{1s} \cdot \sin(2\theta) + C_{1c} \cdot \cos(2\theta) \tag{25}$$

Poteki enosmerne komponente ter drugega harmonika, predstavljenega kot seštevek sinusa in kosinusa, so prikazani na sliki 6.

Enosmerna komponenta ima obliko racionalne funkcije v funkciji atan. C_{1s} in C_{1c} sta racionalni funkciji. Amplituda je geometrijska vsota členov $C_n = \sqrt{C_{ns}^2 + C_{nc}^2}$, faza za sinusni potek napake je $\varphi_n = atan(\frac{C_{nc}}{C_{ns}})$. Končna enačba z upoštevanjem amplitude osnovnega signala sinus in kosinus se glasi:

$$\varepsilon(\Delta_c) = \text{atand}\frac{\Delta_c}{\Delta_c + 2A_1} + \frac{180}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (\frac{\Delta_c}{\sqrt{\Delta_c^2 + 2A_1\Delta_c + 2A_1}})^n \sin(2n\theta + n(90° + \text{atan}(\frac{\Delta_c + A_1}{A_1}))) \tag{26}$$



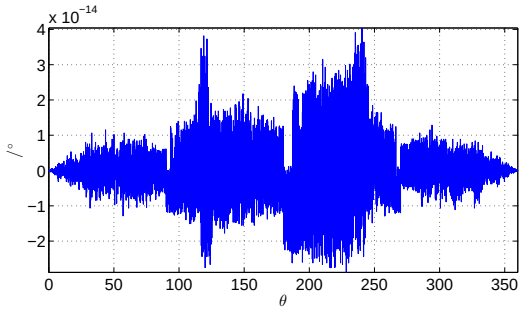
Slika 6: Potek enosmerne komponente in drugega harmonika napake, v odvisnosti od Δ_c

$$\varepsilon(\Delta_s) = \text{atand}\frac{-\Delta_s}{\Delta_s + 2A_1} + \frac{180}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{\Delta_s}{\sqrt{\Delta_s^2 + 2A_1\Delta_s + 2A_1}}\right)^n \sin\left(2n\theta + n\left(90^\circ + \text{atan}\left(\frac{\Delta_s + A_1}{A_1}\right)\right)\right) \quad (27)$$

$$\Delta_s, \Delta_c > -A_1$$

3 Komentar rezultatov

Pri modeliranju potekov je bilo uporabljeno prvih 15 členov potenčne vrste. Razlika med dejansko napako in predvideno napako je le numerična (slika 7). Opravil sem FFT predvidene in dejanske napake. Razlika med posameznima amplitudama harmonika je le numerična. V primeru, da parameter dosega večje vrednosti ali se približuje robnim pogojem, dejanska napaka postane nezvezna in je s 15 členi ni mogoče povsem določiti. Omeniti je potrebno, da so kljub izpeljavam, predstavljene vrste napake posameznih deformacij med seboj odvisne.



Slika 7: Razlika med predvideno (13) in dejansko napako pri $k = 1.1$

4 Zaključek

V članku so predstavljeni poteki napake v odvisnosti od različnih amplitud, enosmernih komponent, faznih zamikov in kombinacije različnih amplitud ter faznih zamikov signalov sinusa in kosinusa. Napaka vsebuje tudi višje harmonike, ki ob večjih razlikah med sinusom in kosinusom niso več zanemarljivi. Za majhna popačenja vhodnih signalov zadostuje napako izraziti z enosmerno komponento in prvim členom potenčne vrste. Napako, izraženo z enosmerno komponento in prvim členom potenčne vrste, je potrdila tudi druga literatura [5]. Izraze se lahko uporabi za ugotovitev nepravilne montaže resolverja ali enkoderja v vgrajenih aplikacijah, kjer ni dostopa do signalov sinus in kosinus. V signalih sinus in kosinus se lahko pojavijo tudi višji harmoniki, ki prav tako vplivajo na napako. Vpliv popačenj vhodnih signalov v funkcijo atan2 na izhodno napako ponuja še veliko izzivov za nadaljne delo.

Literatura

[1] Gachter J., Hirz M.,Seebacher R., ”Impact of Rotor Position Sensor Errors on Speed Controlled Permanent Magnetized Synchronous Machines”, IEEE 12th International Conference on Power Electronics and Drive Systems (PEDS), pp.822-830, 12-15 Dec. 2017

[2] Brugnano F., Concari C., Imamovic E., Savi F., Toscani A., Zanichelli R., ”A simple and accurate algorithm for speed measurement in electric drives using incremental encoder”, IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society, pp. 8551-8556, 29 Oct.-1 Nov. 2017

[3] Reddy B.P., Murali A., Shaga G., ”Low Cost Planar Coil Structure for Inductive Sensors to Measure Absolute Angular Position”, 2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST), pp.14-18, 14-16 April 2017

[4] Zhang Z., Ni F., Liu H., Jin M., ”Theory analysis of a new absolute position sensor based on electromagnetism”, International Conference on Automatic Control and Artificial Intelligence, pp.2204-208, 3-5 Mar. 2012

[5] Qi Lin, T. Li, Z. Zhou, ”Error Analysis and Compensation of the Orthogonal Magnetic Encoder”, Proceedings of IEEE ICMCC Conference, pp.11-14, 21-23 Oct. 2011

[6] Hanselman D.C., ”Resolver Signal Requirements for High Accuracy Resolver-to-Digital Conversion”, IEEE Transactions on Industrial Electronics, vol.37, no.6, pp.556-561, Dec. 1990

[7] Demierre M., “Improvements of CMOS Hall Microsystems and Application for Absolute Angular Position Measurements”, PhD. thesis, pp. 152-161, Federal Polytechnic School of Lausanne, Switzerland, 2003

[8] Dolinar G. ”Matematika 1”, Fakulteta za elektrotehniko, Založba FE in FRI, 2010

[9] https://www.mathworks.com/help/matlab/ref/atan2.html, dostop junij 2018

[10] https://www.mathworks.com/help/matlab/ref/atan2d.html, dostop junij 2018

Prepoznavanje pisanja s pomočjo konvolucijskih nevronskih mrež in LSTM

Blaž Bertalanič

Fakulteta za elektrotehniko Univerze v Ljubljani, Tržaška 25, 1000 Ljubljana
E-pošta: blaz.bertalanic@protonmail.com

Using CNNs with Long-Short-Term Memory for recognition of writing

Abstract. *This paper addresses the problem of automatic recognition of human behavior, specifically writing, from the image sequences. We propose to use Convolutional Neural Networks utilizing long-short term memory (LSTM) layer to recognize writing of students during a lecture. We present a dataset consisting of more than 13 hours of video involving 22 persons, and describe a structure of the proposed CNN. We present a comparison of recognition accuracy of the classical CNN structure using a single image, and the proposed CNN with LSTM operating on a batch of 10 images. The results indicate that the proposed structure clearly outperforms the baseline method in the detection of writing.*

1 Uvod

Razpoznavanje vzorcev in klasifikacija slik s pomočjo konvolucijskih nevronskih mrež (CNN) [1] v zadnjih letih dosega hiter napredek zahvaljujoč dvigu računskih moči s pomočjo grafičnih kartic. Rešitve, ki uporabljajo to tehnologijo, najdemo praktično na vsakem koraku v našem življenju, od telefonov, do razpoznavanja vzorcev v industriji.

Čeprav konvolucijske nevronske mreže zelo dobro delujejo pri klasifikaciji objektov na slikah (recimo ali je na sliki oseba, pes, mačka, ...), pa pride do težav pri razpoznavanju dogodkov ali akcij, ki so časovno odvisni oziroma trajajo dalj časa. Pri teh je potrebno upoštevati sekvenco zaporednih slik, da lahko podamo končno oceno. To je tudi težava, kateri se bomo posvetili v tem članku.

V članku predstavljamo model dvo-dimenzionalne konvolucijske nevr. mreže v kombinaciji z Long short term memory (LSTM) slojem [2], ki uspešno razpoznava pisanje oseb na video posnetku.

Problem razpoznavanja aktivnosti pisanja je prisoten pri opazovanju in avtomatiziranem ocenjevanju aktivnosti in pozornosti učencev ali študentov v razredu [11], kjer je pisanje ena od pomembnih indikacij. Avtomatizirano razpoznavanje pisanja želimo uporabiti za dolgoročno ocenjevanje aktivnosti in pozornosti učenca v razredu.

2 Pregled literature

Uporaba nevronskih mrež za razpoznavanje aktivnosti ljudi še ni zelo razširjena, saj je potrebno upoštevati časovno komponento.

Skupina indijskih raziskovalcev [3] je uporabila CNN mreže za razpoznavanje gest, vendar so pred slojem LSTM uporabili 3D konvolucijske mreže, ki so precej bolj zahtevne za računanje od naše metode. S svojim modelom so poizkušali razpoznati različne geste ljudi iz video posnetka. Podobna ideja je bila uporabljena tudi za razpoznavo dejanj iz videa zgolj s pomočjo 3D konvolucijskih mrež [4].

Ena od možnosti za analizo video posnetka je tudi kombinacija slik in optičnega toka (ang. Optical flow), kar so uporabili raziskovalci pri razpoznavanju aktivnosti iz video posnetkov [5], želimo pa uporabiti optični pretok v nadaljnjem razvoju prepoznavanja pisanja.

Raziskovalcem je uspelo narediti tudi model, ki je s pomočjo časovne distribucije obdelal vsako sliko videa posebej in na koncu naredil odločitev s polno povezanim slojem [6].

Uporaba 2D konvolucijskih mrež in LSTM je bila uporabljena pri generiranju opisov video posnetkov [7]. Razlika glede na našo metodo je v tem, da pri nas LSTM mreža umesti akcijo v eno izmed dveh skupin, pri njih pa opiše dogajanje na posnetku.

3 Generiranje zbirke podatkov

Zbirka podatkov za prepoznavanje pisanja je bila posneta v laboratoriju LUCAMI na Fakulteti za elektrotehniko. Za namen analize vedenja študentov je bilo izvedeno snemanje študentov s Kinect One senzorjem med spremljanjem lekcije na temo obdelave signalov (DSP) v dolžini 25 minut. V več poskusih je bilo skupno posnetih 22 udeležencev, ki so izpolnjevali teste znanja pred in po lekciji. Analiza vpliva pozornosti na znanje pa je bila tudi objavljena [12].

Video posnetke študentov med lekcijo smo uporabili kot vhodne podatke za učenje modela mreže za prepoznavanje pisanja, pri čemer smo ročno za vsako osebo označili ustrezen razred (pisanje, ne-pisanje) za celotni posnetek. Za prvo skupino šestih oseb smo uporabili celotnih 33 minut video posnetka, medtem ko smo za ostalih 16 uporabili 8 minut dolge izseke iz originalnega video posnetka. Oba video posnetka sta vsebovala 10 sličic na sekundo.

a)



b)

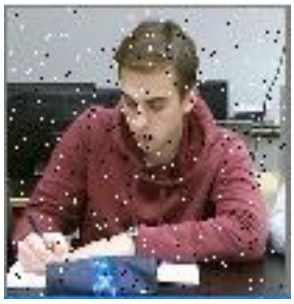


Slika 1. Primer dveh posnetkov predavanj. Prvi dolžine 8 minut in drugi dolžine 33 minut.

Posamezne osebe v video posnetku smo nato izrezali s fiksno pozicijo kvadratnega okna, ki smo ga postavili tako, da je čim bolj zajemal celotno gibanje posamezne osebe. Glede na postavo oseb smo tudi izbrali različne velikosti oken, nato pa pridobljene slike pomanjšali na enotno velikost 100x100 pikslov s tremi RGB kanali. Iz tega smo pripravili dve zbirki podatkov. V prvi smo zaporedne slike združili skupaj v sekvence po 10, kar je v našem primeru pomenilo enosekundni odsek, in dobili končno obliko enega vhodnega podatka (10, 100, 100, 3).



a)



b)

Slika 2. Na sliki a) je primer izreza osebe. Na sliki b) je primer slike z dodanim šumom

Pri drugi zbirki smo slike obdržali samostojno in za referenco poizkušali sistem naučiti prepoznavati pisanje le na slikah (referenčna metoda). Pri obeh zbirkah smo tudi zrcalili vse primere preko vertikalne osi in tako še povečali število primerov. Hkrati smo na tak način poskrbeli tudi, da smo imeli v učnem setu osebe, ki pišejo z levo in desno roko.

Težava pri naših podatkih je bila neenakomerna predstavitev obeh oznak, saj je bilo ne-pisanje prisotno v približno 2/3 vseh podatkov. Ta problem smo rešili na način, da smo primerom pisanja dodali šum in jih nato dodali v zbirko podatkov za učenje. Skupno smo na koncu imeli za približno 13 ur enosekundnih sekvenc pisanja in ne-pisanja, kar je za uporabo v globokem učenju (ang. Deep learning) še vedno relativno majhna zbirka podatkov.

Akcije pisanja smo z ogledi videoposnetkov označili sami za vsako sliko video posnetka, ter nato po 10 zaporednih ocen (na 1 sekundo posnetka) s pomočjo median filtra združili v eno oceno, ki pripada enemu vhodnemu podatku v prvi zbirki in je predstavljala povprečno vrednost pisanja v eni sekundi, medtem ko smo pri drugi zbirki uporabili prvotne oznake.

4 Predlagana metoda

4.1 Oblika mreže

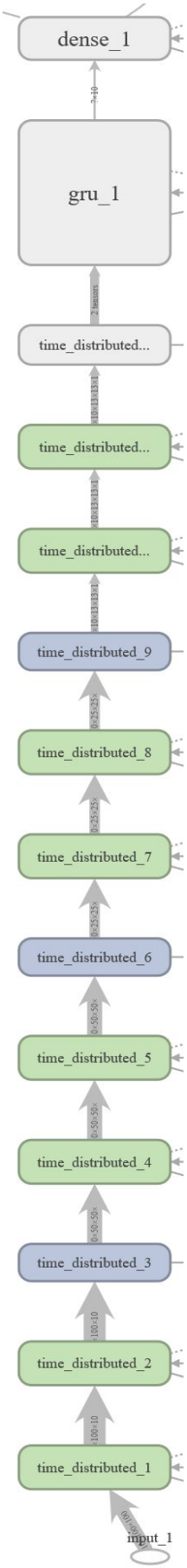
Metode razpoznavanja pisanja smo se lotili na dva načina. Prvi način je napovedoval dejanje na statičnih slikah in smo ga uporabili kot referenco za našo predlagano metodo, medtem ko smo pri drugi metodi upoštevali odvisnost slik med sabo in tako tudi zasnovali mrežo s slojem LSTM, ki nam je pri odločitvi upošteval celotno sekundno sekvenco.

Mreža, ki smo jo uporabili v prvem poskusu (osnovna metoda na posameznih slikah), je imela 13 slojev. Najprej smo postavili tri sklope s po dvema konvolucijskima in enim MaxPooling slojem, temu so sledili Dropout, Flatten in FullyConnected sloj, ki se je nato povezal na izhod, kjer smo dobili našo napoved stanja na sliki.

V drugem primeru (predlagana metoda) je naša mreža ponovno imela tri sklope s po dvema konvolucijskima in enim MaxPooling slojem, ki sta mu sledila še dva konvolucijska sloja, Flatten in nato LSTM sloj, ki pa je bil kasneje zamenjan s slojem Gated recurrent unit (GRU)[8], ki deluje podobno kot LSTM, le da je arhitektura nekoliko preprostejša in posledično olajša računanje, hkrati pa bolje deluje na manjših zbirkah podatkov [9]. Razlog za izbor takšne arhitekture mreže se skriva ravno v zadnjem sloju, saj si LSTM/GRU sloj zapomni prejšne vhodne podatke in jih posledično upošteva pri končni napovedi dejanja. V našem primeru si zapomni vseh 10 slik, ki sestavljajo eno sekundo in tako lahko bolj natančno napove, ali oseba v tem času piše ali ne. Vsi sloji so delovali s časovno distribucijo, saj smo na ta način sekvenco pripeljali do sloja GRU, ki si je nato zapomnil vseh 10 slik in podal končno napoved.

Z uporabo druge metode zmanjšamo možnost, da bi pretentali sistem, ki uporablja statične slike za napovedovanje. Za primer bi lahko oseba držala okvirno pozo na kateri bi delovala kot da piše, v resnici pa roke sploh ne bi premaknila.

Vsi naši konvolucijski sloji so imeli 16 filtrov in vsi, razen v prvem sloju, so bili velikosti 3x3, medtem ko je že omenjeni sloj imel velikost filtrov 7x7. Naš GRU sloj je imel 10 celic. Mreža je bila majhna zaradi majhnega števila oznak in smo se s tem izognili preprileganju in tako dosegli boljšo generalizacijo.



Slika 3. Oblika predlagane mreže za učenje

4.2 Učenje mreže

Mrežo smo učili na Nvidia grafični kartici Titan Xp, ki vsebuje 3840 CUDA jeder s frekvenco delovanja 1.6 GHz in 12 GB grafičnega pomnilnika.

Učenje je potekalo na množici za učenje (ang. Batch) v velikosti 24 vzorcev in z učnim korakom (ang. Learning rate) velikosti 0.000025. Proces je tekel 5 ciklov (ang. Epoch).

Za optimizacijo smo uporabili SGD algoritem z momentumom 0.9, ki daje najboljše rezultate, če želimo, da je model čim bolj generalen [10].

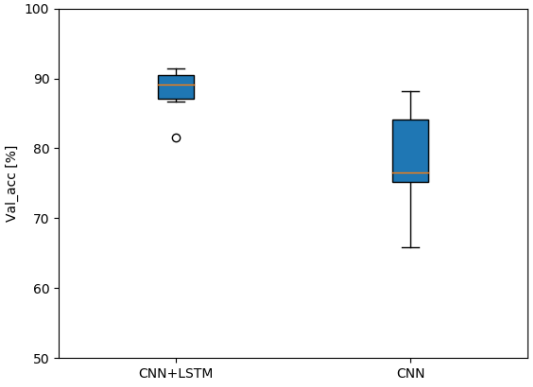
5 Rezultati

Našo rešitev smo testirali tako, da smo za vse poizkuse izločili vse slike neke osebe iz učne množice podatkov. Oseba, ki smo jo izbrali, je imela označenih 2944 primerov nepisanja in 1036 pisanja za testiranje LSTM metode. Pri navadnih slikah smo imeli 29246 oznak nepisanja in 10554 pisanja.



Slika 4. Primer slike iz validacijskega seta

Ker gre v osnovi še vedno za majhno zbriko podatkov, je bilo naše učenje nekoliko odvisno tudi od sreče, saj se ob začetku procesa uteži določijo naključno in kot take, zaradi manjšega števila podatkov, naš algoritem lahko usmerijo v različne lokalne minimume. Kljub temu se končni rezultati v večini primerov ob ponovitvah razlikujejo le za nekaj odstotkov in imajo povprečno vrednost pravilne napovedi na validacijskem setu 89.1%.



Slika 5. Statistična predstavitev natančnosti (validacijski podatki) pri predlagani metodi CNN+LSTM (levo) in referenčni mreži (desno)

Podobno je bilo z učenjem na navadnih slikah, kjer je bil povprečni odstotek nižji, vendar se nam težava pojavi pri ponavljivosti. V primeru učenja zgolj na posameznih slikah smo zgolj v enem primeru imeli to srečo, da se je model solidno naučil razlikovati med obema akcijama, medtem ko je v ostalih poizkusih večinoma dobro napovedoval ne-pisanje, medtem ko je bilo razmerje med pravilno in nepravilno napovedanimi primeri pisanja okoli polovične uspešnosti.

Za predstavitev rezultatov smo uporabili modela, ki sta na koncu dala najboljši delež pravilnih napovedi na validacijskem setu. Zaradi neenakomerne porazdelitve primerov za pisanje in ne-pisanje pri validacaijskem setu, smo se osredotočili predvsem na uspešnost napovedovanja pisanja.

Izkaže se, da naša metoda bolje napoveduje dejanje pisanja, medtem ko se pri ne-pisanju uspešnost napovedi razlikuje za približno 0.5% in je kot taka praktično nespremenjena.

Tabela 1. Matrika razvrstitev na slikah (primerjalna metoda)

/	Ne-pisanje	Pisanje
Ne-pisanje	**95.6%**	4.4%
Pisanje	32.3%	**67.7%**

Tabela 2. Matrika razvrstitev na sekvenci 10 slik (predlagana metoda)

/	Ne-pisanje	Pisanje
Ne-pisanje	**96.1%**	3.9%
Pisanje	24.2%	**75.8%**

Kot manjši test smo rezultate primerjalne metode združili v enako dolge sekvence, kot smo jih imeli pri testiranju naše metode, in z uporabo median filtra določili skupno napoved na tej sekvenci. Izkaže se, da je rezultat še nekoliko slabši kot na primerjalni metodi.

Naša metoda je izboljšala napoved pisanja kar gre pripisati predvsem temu, da smo z njo upoštevali tudi gibanje oseb, predvsem rok, ki je ključna značilnost tega dejanja. Še vedno pa naša uspešnost ni na dovolj visoki ravni, da bi jo lahko uporabili v splošne namene. Razlog za to gre iskati predvsem v različnih načinih gibanja oseb, drugačnih pozicijah telesa in tudi v različnih slogih pisanja.

6 Zaključek

Pokazali smo, da predlagan model bolje rešuje naš problem kot zgolj navadna CNN mreža. Naša rešitev tako izrablja prednost CNN, ki je razpoznavanje objektov na sliki in hkrati s pomočjo LSTM upošteva časovno odvisnost slik med seboj.

V dosedanjem delu smo se osredotočali zgolj na razpoznavo na video posnetku, zato imamo v nadaljevanju raziskave namen uporabiti tudi globinske slike in izračunan optični tok (ang. Optical flow). Dodatni množici podatkov bi lahko za mrežo predstavljali nove informacije, na podlagih katerih bi potekala končna odločitev ali oseba piše ali ne. Načrtujemo tudi dodatna snemanja, s katerimi bi še povečali velikost in raznolikost učne množice.

Literatura

[1] Y. LeCun and Y. Bengio. »Convolutional networks for images, speech, and time-series,« In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

[2] Sepp Hochreiter and J¨urgen Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, 1997.

[3] Koustav Mullick and Anoop M. Namboodiri, »Learning deep and compact models for gesture recognition,« https://cdn.iiit.ac.in/cdn/cvit.iiit.ac.in/images/ConferencePapers/2017/GESTURE_RECOGNITION.pdf, [Na spletu; dostopano 10. 05. 2018].

[4] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen, “Sign Language Recognition Using Convolutional Neural Networks,” v *European Conference on Computer Vision Workshops*, 2015.

[5] Karen Simonyan and Andrew Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” v *Neural Information Processing Systems*, 2014.

[6] G¨ul Varol, Ivan Laptev, and Cordelia Schmid, “Long-term Temporal Convolutions for Action Recognition,” *arXiv:1604.04494*, 2016.

[7] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell, »Long-term Recurrent Convolutional Networks for Visual Recognition and Description«, arXiv: 1411.4389v4, 2016

[8] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. »On the properties of neural machine translation: Encoder-decoder approaches,« arXiv: 1409.1259, 2014.

[9] Junyoung Chun, Caglar Gulcehr, KyungHyun Cho, Yoshua Bengio, »Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,« arXiv: 1412.3555, 2014.

[10] Nitish Shirish Keskar, Richard Soche, »Improving Generalization Performance by Switching from Adam to SGD,« arXiv: 1712.07628, 2017.

[11] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features". *EURASIP Journal on Image and Video Processing* , ISSN 1687-5176, 2017, vol. 2017, 80, str. 1-12

[12] BURNIK, Urban, ZALETELJ, Janez, KOŠIR, Andrej. Video-based learners' observed attention estimates for lecture learning gain evaluation. *Multimedia tools and applications*, ISSN 1380-7501, 2017.

Zahvala

Rad bi se zahvalil dr. Janezu Zaletelju za vso pomoč pri analizi in reševanju opisanega problema.

Intuitivno vodenje robotske roke s kretnjami

Domen Ulbl[1],
[1]*Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru*
E-pošta: domen.ulbl@student.um.si

Gesture based intuitive control of a robotic arm

Abstract. *In this paper we present a method for intuitive control of a collaborative robotic arm using hand and arm gestures performed by the user. The proposed system uses two Myo Armband units to capture the motion of the arm and a personal computer to process the measurements and simulate the motion of the Kinova Mico manipulator in the V-REP simulation environment. A mathematical model of the human arm is built in order to compute the target position for the robot to reach. The resulting simulation is presented in video form as well.*

1 Uvod

Roboti so bili dolgo časa uporabljeni le na področjih vojske, proizvodnje in visokotehnoloških raziskav. Dandanes se ta trend spreminja, saj se roboti uveljavljajo na številnih novih področjih kot sta medicina in pomoč v domačem okolju. Ta sprememba je povzročila potrebo, da robote nadzirajo in varno uporabljajo ljudje brez strokovnega znanja na področju robotike. Tako sta tradicionalni metodi programiranja robotov z nadzorno konzolo ali programiranje v simulaciji neustrezni.

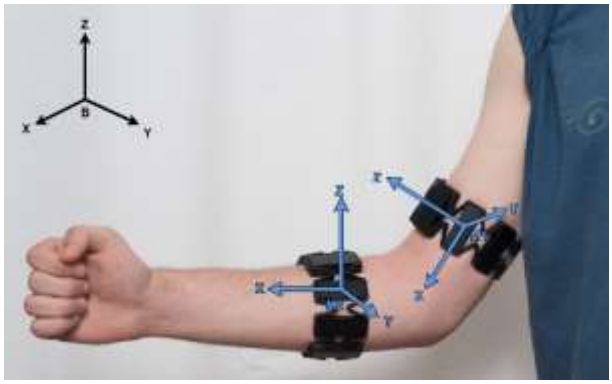
V člankih [5] in [10] je predlagan nov način nadzora nad robotskimi manipulatorji, ki izkorišča kretnje uporabnikovih rok, identificirane s sistemi za zajem slike. Namen naše raziskovalne naloge pa je razviti sistem za intuitivno vodenje manipulatorja s kretnjami uporabnikove roke brez uporabe kamer. S takšnim pristopom ni več potrebe po strukturiranem okolju z nameščenimi kamerami in ustrezno osvetlitvijo, ampak zadoščajo le nosljivi senzorji gibanja, nameščeni na uporabniku.

Na področju vodenja robotov s senzorji na uporabniku je bilo v zadnjem času izvedenih precej raziskav, ki uporabljajo človeške kretnje za nadzor mobilnega robota [3], robotske roke [8] ali simulacije [6]. Skupna točka omenjenih člankov je zajemanje človeških gibov in mišične aktivnosti z Myo Armband, inovativno zapestnico za interakcijo med ljudmi in napravami.

V našem članku bomo predstavili vodenje robotske roke Kinova Mico s senzorsko zapestnico Myo s pomočjo kinematičnega modela človeške roke in programskega vmesnika na osebnem računalniku za pošiljanje podatkov med zapestnico in simulacijo. V prvem delu bomo predstavili sestavne dele sistema in povezavo med njimi, v drugem delu bomo opisali kinematični model človeške roke, v tretjem delu pa bomo prikazali eksperimentalne rezultate v obliki videa simulacije.

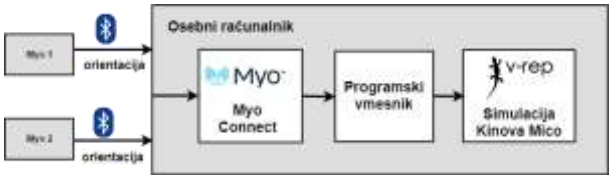
2 Opis sistema

Sistem za vodenje je sestavljen iz dveh zapestnic Myo, ki jih uporabnik namesti na roko, kot je prikazano na sliki 1. Oznaki M1 in M2 predstavljata koordinatna sistema zapestnic, katerih rotaciji sovpadata z rotacijama uporabnikove nadlakti in podlakti.



Slika 1: Roka uporabnika nameščenima zapestnicama

Zapestnici sta na osebni računalnik povezani preko BLE 4.0 (Bluetooth Low Energy) povezave in Bluetooth sprejemnika. Gonilniki in programska oprema *Myo Connect* skrbijo za povezavo med strojno opremo in programskim vmesnikom, ki smo ga napisali v C++ jeziku. S funkcijami zapestnice upravljamo preko *libmyo* API (Application Program Interface), ki je del programske opreme Myo Connect, kot je prikazano na sliki 2. Podatke iz zapestnic pošiljamo v simulacijsko okolje V-REP s pomočjo aplikacijskega programskega vmesnika API za »Remote client« programe. V okolju V-REP se izvaja simulacija manipulatorja Mico.



Slika 2: Shematični prikaz sistema

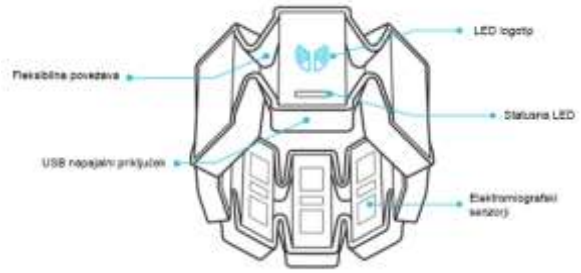
2.1 Zapestnica Myo

Zapestnica Myo podjetja Thalmic Labs omogoča meritev pospeška, kotne hitrosti in orientacije v treh oseh kot tudi zaznavo gest uporabnikove dlani. Centralna enota zapestnice je procesor Cortex M4, ki preko I2C komunicira z Invensense MPU 9150 enoto za zajem gibanja in NRF51822 Bluetooth modul [7].

Enota za zajem gibanja vsebuje 3-osni MEMS pospeškometer, žiroskop in magnetometer ter nam tako

nudi meritve pospeška, kotne hitrosti in smeri Zemljinega magnetnega polja, iz katerih pridobimo meritev orientacije.

Kretnje dlani se zaznavajo preko mišične aktivnosti, merjene z 8 elektromiografskimi senzorji na notranji strani zapestnice (slika 3).



Slika 3: Deli zapestnice Myo (povzeto po [9])

2.2 Programski vmesnik

Programski vmesnik, ki smo ga napisali, se preko *libmyo* API odziva na dogodke, ki jih generira *Myo Connect* glede na stanje zapestnice. Zajemanje orientacije poteka s frekvenco 50Hz. Program izpisuje podatke o stanju zapestnice v komandno vrstico in shranjuje meritve v datoteke .csv. Uporabniku omogoča nastavitev začetne orientacije zapestnic, ki določa referenco za vse nadaljnje rotacije. Preko V-REP API pošilja ukaze simulaciji robotske roke.

2.3 Simulacija Kinova Mico v V-REP

Simulacijsko okolje V-REP proizvajalca Coppelia Robotics [2] omogoča natančno simulacijo kinematike in dinamike poljubnih robotskih mehanizmov. Pri našem projektu smo se omejili na simulacijo kinematske robotske roke.



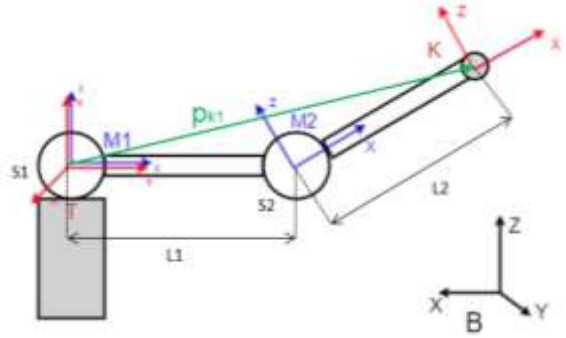
Slika 4: Robotska roka Kinova Mico z označenimi sklepi S1-S6

Za vodenje smo izbrali model šestosne kolaborativne robotske roke Kinova Mico (slika 4), ker je njena kinematična struktura podobna človeški roki in je njen maksimalni doseg le malo manjši od človeškega. To robotsko roko imamo tudi v Laboratoriju za industrijsko robotiko, UM-FERI, tako da je v prihodnje mogoča tudi implementacija na realnem sistemu.

3 Princip vodenja

Za vodenje robotske roke uporabimo meritvi orientacije obeh zapestnic, iz katerih s pomočjo matematičnega modela človeške roke izračunamo pozicijo konca roke. Preko programskega vmesnika pošljemo meritvi orientacije v simulacijsko okolje V-REP. Format meritev je rotacijski kvaternion.

Enostaven model človeške roke na sliki 5 je sestavljen iz dveh sferičnih sklepov S1 in S2, ki predstavljata ramenski in komolčni sklep, ter dveh vmesnih segmentov L1 in L2, ki prestavljata nadlaket in podlaket.



Slika 5: Poenostavljen model človeške roke z označenima koordinatnima sistemoma trupa in konca roke

Orientaciji koordinatnih sistemov M1 in M2 sta določeni z meritvama orientacije zapestnic, nameščenih kot je prikazano na sliki 1. Če upoštevamo še dolžini nadlakti in podlakti L1 in L2, lahko pozicijo koordinatnega sistema konca K glede na koordinatni sistem trupa T določimo iz transformacijskih matrik:

$$ {}^{\mathbf{K}}\mathbf{T}_{\mathbf{T}} = {}^{1}\mathbf{T}_{\mathbf{T}} \cdot {}^{2}\mathbf{T}_{1} \cdot {}^{\mathbf{K}}\mathbf{T}_{2} \tag{2.1} $$

$$ \begin{bmatrix} {}^{1}_{T}r_{11} & {}^{1}_{T}r_{12} & {}^{1}_{T}r_{13} & 0 \\ {}^{1}_{T}r_{21} & {}^{1}_{T}r_{22} & {}^{1}_{T}r_{23} & 0 \\ {}^{1}_{T}r_{31} & {}^{1}_{T}r_{32} & {}^{1}_{T}r_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} {}^{2}_{1}r_{11} & {}^{2}_{1}r_{12} & {}^{2}_{1}r_{13} & L1 \\ {}^{2}_{1}r_{21} & {}^{2}_{1}r_{22} & {}^{2}_{1}r_{23} & 0 \\ {}^{2}_{1}r_{31} & {}^{2}_{1}r_{32} & {}^{2}_{1}r_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & L2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = $$

$$ \begin{bmatrix} {}^{1}_{T}r_{11}{}^{2}_{1}r_{11} + {}^{1}_{T}r_{12}{}^{2}_{1}r_{21} + {}^{1}_{T}r_{13}{}^{2}_{1}r_{31} & {}^{1}_{T}r_{11}{}^{2}_{1}r_{12} + {}^{1}_{T}r_{12}{}^{2}_{1}r_{22} + {}^{1}_{T}r_{13}{}^{2}_{1}r_{32} & {}^{1}_{T}r_{11}{}^{2}_{1}r_{13} + {}^{1}_{T}r_{12}{}^{2}_{1}r_{23} + {}^{1}_{T}r_{13}{}^{2}_{1}r_{33} & L1{}^{1}_{T}r_{11} + L2({}^{1}_{T}r_{12}{}^{2}_{1}r_{11} + {}^{1}_{T}r_{12}{}^{2}_{1}r_{21} + {}^{1}_{T}r_{13}{}^{2}_{1}r_{31}) \\ {}^{1}_{T}r_{21}{}^{2}_{1}r_{11} + {}^{1}_{T}r_{22}{}^{2}_{1}r_{21} + {}^{1}_{T}r_{23}{}^{2}_{1}r_{31} & {}^{1}_{T}r_{21}{}^{2}_{1}r_{12} + {}^{1}_{T}r_{22}{}^{2}_{1}r_{22} + {}^{1}_{T}r_{23}{}^{2}_{1}r_{32} & {}^{1}_{T}r_{21}{}^{2}_{1}r_{13} + {}^{1}_{T}r_{22}{}^{2}_{1}r_{23} + {}^{1}_{T}r_{23}{}^{2}_{1}r_{33} & L1{}^{1}_{T}r_{21} + L2({}^{1}_{T}r_{21}{}^{2}_{1}r_{11} + {}^{1}_{T}r_{22}{}^{2}_{1}r_{21} + {}^{1}_{T}r_{23}{}^{2}_{1}r_{31}) \\ {}^{1}_{T}r_{31}{}^{2}_{1}r_{11} + {}^{1}_{T}r_{32}{}^{2}_{1}r_{21} + {}^{1}_{T}r_{33}{}^{2}_{1}r_{31} & {}^{1}_{T}r_{31}{}^{2}_{1}r_{12} + {}^{1}_{T}r_{32}{}^{2}_{1}r_{22} + {}^{1}_{T}r_{33}{}^{2}_{1}r_{32} & {}^{1}_{T}r_{31}{}^{2}_{1}r_{13} + {}^{1}_{T}r_{32}{}^{2}_{1}r_{23} + {}^{1}_{T}r_{33}{}^{2}_{1}r_{33} & L1{}^{1}_{T}r_{31} + L2({}^{1}_{T}r_{32}{}^{2}_{1}r_{11} + {}^{1}_{T}r_{32}{}^{2}_{1}r_{21} + {}^{1}_{T}r_{33}{}^{2}_{1}r_{31}) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2} $$

Pri tem je:

${}^{1}\mathbf{T}_{\mathbf{T}}$ je transformacijska matrika M1 glede na T,

${}^{2}\mathbf{T}_{1}$ je transformacijska matrika M2 glede na M1,

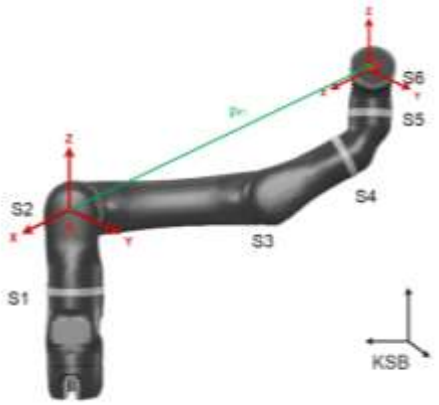
${}^{\mathbf{K}}\mathbf{T}_{2}$ je transformacijska matrika K glede na M2,

${}^{K}\mathbf{T}_{\mathbf{T}}$ je transformacijska matrika K glede na T.

Rotacija v $\mathbf{T}_{\mathbf{T}}^{1}$ in $\mathbf{T}_{1}^{2}$ je predstavljena z rotacijsko matriko, ki jo lahko izračunamo iz kvaterniona po postopku, opisanem v [1].

Vektor P_{k1} dobimo iz pozicijskega dela matrike (2.2). Ta vektor nam predstavlja pozicijo konca modela človeške roke glede na koordinatni sistem trupa.

Končna točka vektorja P_{k1} hkrati predstavlja ciljno pozicijo konca robotske roke Mico, ki naj jo ta zavzame glede na koordinatni sistem A (slika 6). Koordinatni sistem A je pozicioniran v sklepu S2, vendar nanj ni pritrjen in tako ne spreminja orientacije, ko robotski sklep S2 rotira.



Slika 6: Model Kinova Mico z vektorjem pozicije konca robotske roke (prirejeno po [2])

Ker je ciljna pozicija podana v kartezičnih koordinatah, jo manipulator doseže v načinu inverzne kinematike, pri čemer orientacije konca manipulatorja ne nadziramo. Za izračun pozicij sklepov iz dane ciljne pozicije konca manipulatorja uporabljamo Jakobianov pseudoinverz. S tem postopkom izračunane pozicije blizu singularne lege in izven dosegljivega področja manipulatorja neprimerno oscilirajo, kar zmanjšuje robustnost vodenja. Robustnost izboljšamo, če zagotovimo, da so vse ciljne pozicije znotraj dosegljivega področja dovolj oddaljene od singularnih leg. V primeru robotske roke Kinova Mico je dosegljivo področje podano v [4].

Obliko dosegljivega območja v grobem aproksimiramo s sfero s centrom v točki A in polmerom r_{max}. Velikost maksimalnega polmera smo določili eksperimentalno z merjenjem velikosti vektorja P_{k1} v trenutku, ko se pojavijo oscilacije v rešitvah inverzne kinematike.

Za doseganje večje robustnosti vodenja moramo preveriti, če dana ciljna pozicija leži znotraj doseglijvega območja, kar enostavno storimo z uporabo sferičnih koordinat.

Vsako ciljno pozicijo $T = (x, y, z)$ lahko v sferičnih koordinatah opišemo kot:

$$r = \sqrt{x^2 + y^2 + z^2} \tag{2.3}$$

$$\theta = \arctan\left(\frac{\sqrt{x^2 + y^2}}{z}\right) \tag{2.4}$$

$$\phi = \arctan\left(\frac{y}{x}\right), \tag{2.5}$$

pri čemer sta kota θ in ϕ določena kot na sliki 8.

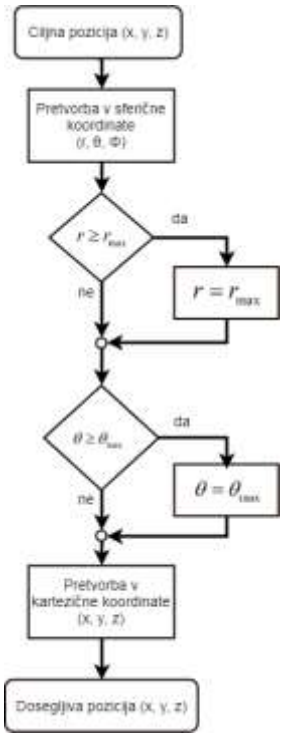
Kot θ je zaradi omejitev gibanja sklepa S2 omejen na interval $0 \le \theta \le \theta_{\max}$.

Za vsako ciljno pozicijo izvedemo preverjanje, prikazano na sliki 7. Najprej pretvorimo kartezične koordinate v sferične. Če ciljna točka leži izven dosegljivega območja, v sferičnih koordinatah izberemo najbližjo točko znotraj dosegljivega območja. Nato pretvorimo sferične koordinate v kartezične po enačbah:

$$x = r\sin(\theta)\cos(\phi) \tag{2.6}$$

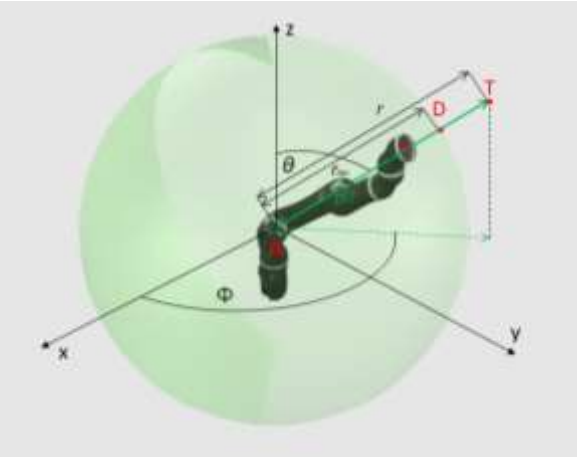
$$y = r\sin(\theta)\sin(\phi) \tag{2.7}$$

$$z = r\cos(\theta) \tag{2.8}$$



Slika 7: Diagram poteka preverjanja dosegljivosti ciljne pozicije

S tem postopkom ciljno točko izven dosegljivega območja projiciramo na površino sfere, ki omejuje dosegljivo področje.



Slika 8: Mico v z označenim dosegljivim področjem

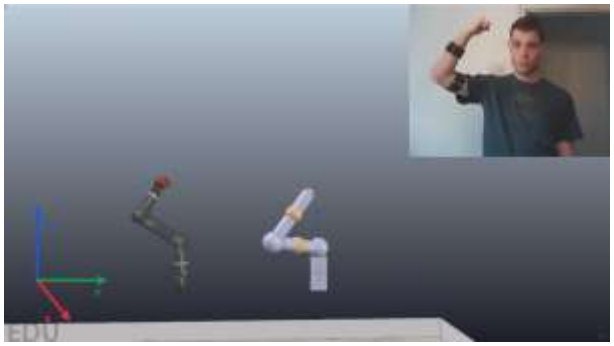
Točka T na sliki 8 predstavlja ciljno pozicijo, določeno z vrhom modela človeške roke, točka D je projekcija točke T v dosegljivi prostor manipulatorja. Tako lahko manipulator Mico postavimo v smiselno pozicijo za vsako ciljno točko.

4 Rezultati

Rezultate smo predstavili v obliki videa (slika 9), ki je dosegljiv na naslovu: https://www.youtube.com/watch?v=5wUPvvUxtF4.

Video je sestavljen iz zajete zaslonske slike med potekom simulacije in posnetka uporabnika iz spletne kamere. V prvem delu posnetka se preko komandne vrstice nastavi začetna orientacija modela človeške roke. Center rdeče krogle predstavlja vrh vektorja ciljne pozicije P_{k1}.

Na simulaciji lahko vidimo, da robot Mico precej natančno in hitro sledi premikom modela človeške roke in se v okviru delovnega območja želeni poziciji čim bolj približa, tudi ko je cilj izven dosegljivega področja.


Slika 9: Posnetek iz videa simulacije

5 Zaključek

V prispevku smo predstavili možen princip vodenja robotske roke z uporabnikovimi kretnjami s pomočjo zapestnice Myo. Izdelali smo kinematični model človeške roke, s katerim smo vodili pozicijo robotske roke Kinova Mico v načinu inverzne kinematike. Sledenje poteka dovolj natančno in hitro. Možne izboljšave vključujejo implementacijo nadzora nad orientacijo konca robota. Geste dlani bi lahko vključili za nadzor dodatnih funkcij kot je na primer nadzor prijemala. Prihodnje delo bo usmerjeno v implementacijo vodenja na dejanskem robotu.

6 Literatura

[1] Baker M., Maths - Conversion Quaternion to Matrix. 2017.
Dostopno na:
http://www.euclideanspace.com/maths/geometry/rotations/conversions/quaternionToMatrix/index.htm
[2] Coppelia Robotics, http://www.coppeliarobotics.com/
[3] Ganoev A., Shin H., Lee K., Study on Virtual Control of a Robotic Arm via a Myo Armband for the Self Manipulation of a Hand Amputee, International Journal of Applied Engineering Research ISSN 0973-4562, Južna Korea, 2016.
Dostopno na:
https://www.ripublication.com/Volume/ijaerv11n2.htm
[4] Kinova Inc., KINOVA Ultra lightweigh robotic arm User Guide, Kanada, 2018.
Dostopno na: https://www.kinovarobotics.com/en
[5] Lavanya K., Shree D., Nitschitha B., Gesture controlled robot, ICEECCOT 2017,Indija, 2017
Dostopno na:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8284549
[6] Sathiyanarayanan M., Mulling T., Nazir B., Controlling a Robot Using a Wearable Device (MYO), IEDR,Velika Britanija, 2015.
Dostopno na:
http://vixra.org/pdf/1703.0013v1.pdf
[7] Stern B. Inside Myo. New York City: Adafruit, 2016.
Dostopno na https://learn.adafruit.com/myo-armband-teardown/inside-myo
[8] Ulkir O., Akgun G. Kaplanoglu E. Real Time Robotic Arm Control Using Wearable Gesture Armband, IMSEC 2017, Adrana, Turčija, 2017.
Dostopno na:
https://www.researchgate.net/publication/321134113_Real_Time_Robotic_Arm_Control_Using_Wearable_Gesture_Armband
[9] Thalmic Labs. Myo SDK Manual. 0.9.0. Ontario, Kanada: Thalmic Labs, 2013.
Dostopno na:
https://developer.thalmic.com/docs/api_reference/platform/index.html
[10] Wilson F., Cueva C. Hugo S., 7 DOF industrial robot controlled by hand gestures using microsoft kinect, CCAC 2017, Kolumbija, 2017.
Dostopno na:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8276455

7 Zahvala

Iskreno se zahvaljujem mentorju izr. prof. dr. Alešu Hacetu za vodenje in vse koristne nasvete pri nastajanju tega prispevka.

Numerični pristop k izračunu polj v kovinskem valovodu splošnega prereza in izračun smernega diagrama iz valovoda izpeljane valovodne lijakaste antene

Kristjan Vuk Baliž[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
E-pošta: kristjan@entropija.si

Numerical aproach to computing fields' distribution inside arbitrary cross-section metalsheet waveguide and radiation pattern computation of from such waveguide extracted horn antenna

Abstract. *Present paper introduces finite differences numerical approach in frequency domain (FDFD) for the purpose of solving metal sheet waveguide problem, considering dispersion diagram and transverse field distribution within the structure. Additionally, the method for computing radiation pattern of a horn antenna is developed. Finally, analytical and numerical results for a particular pyramidal horn are metrically compared.*

1 Uvod

Valovodna lijakasta antena je ena od osnovnih elektromagnetnih struktur, ki omogoča sistematično sevanje elektromagnetne eneregije v prostor. Ko je sprejemna antena ali nabor več sprejemnih anten prostorsko omejen, je namreč smiselno sevano moč oddajne antene skoncentrirati v ustrezni del prostora. Tedaj torej razmšljamo o (anizotropnih) antenah z izraženo smernostjo (D > 0 dBi). Sevanje takšnih anten v prostor najbolj nazorno opiše smerni diagram, skozi tega pa tudi prepoznavamo njihovo primernost za posamezno aplikacijo.

Te so s porastjo potrebe po brezžičnem podatkovnem prometu v vzhajajoči peti generaciji (5G) informacijsko-komunikacijske dobe čedalje zahtevnejše, pogojujoč predvsem energetsko učinkovitost in ustrezen domet brezvrvične komunikacijske zveze. Na nivoju anten je relevanten parameter, ki zadosti obema pogojema, podatek o dobitku antene. Pri izboru antene za dano uporabo nas torej zanima predvsem njen smerni diagram, iz katerega izhaja omenjeni parameter in nasploh celoten tehniški profil antene.

Pred izdelavo, v fazi načrtovanja valovodne antene, lahko za lepe geometrije (pravokotni in krožni prerezi) smerni diagram pridobimo analitično, v splošnem pa nam ostanejo numerične metode, ki zagotavljajo poljubno, vendar končno natančnost, in jih je danes moč poganjati že na vsakem sodobnem osebnem računalniku. Ko anteno izdelamo, je možna tudi električna meritev, ki da unikaten smerni diagram ob vseh neidealnostih praktičnega izdelka. Kot merilo vrednotenja kakovosti numerične metode bomo vzeli primerjavo numeričnega rezultata z analitičnim, uporabljajoč ustrezno metriko.

V nadaljevanju je opravljen razvoj numerične metode končnih razlik za namen numerične rešitve valovodne valovne enačbe, kar da vplogled v retardirano polje v valovodu. Prvih nekaj zaporednih rodov TE predstavimo v numerično zgeneriranem disperzijskem diagramu. Sledi razdelek z razvojem metode za numerični izračun smernega diagrama valovodne antene v spremstvu ustreznih grafičnih uprizoritev. Za konec pa je narejena še primerjava numeričnega rezultata z analitičnim.

2 Numerična metoda končnih razlik

2.1 Razvoj numerične metode končnih razlik za izračun polja v valovodu

Ker z analitiko obvladujemo le nekatere enostavnejše geometrije, si privoščimo razvoj univerzalnejše, numerične metode; metode končnih razlik (angl. finite differences - FD). Valovodno anteno lahko fizikalno obravnavamo kot problem razširjanja elektromagnetnega valovanja po omejenem ter naprej v bolj ali manj neomejen, prazen prostor. Najprej se lotimo prirobnice (valovoda), ki vodi valovanje od vzbujalnega (sevalnega) telesa dalje v razpirajoči se lijak, ki valovanje sklaplja v prazen prostor. Rešitev sistema Maxwell-ovih enačb da za tak slučaj valovno enačbo v treh dimenzijah, kot sledi:

$$\Delta_{\mathrm{xyz}}\psi(\vec{r},t) = \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} \tag{1}$$

Enačbo (1) preoblikujemo za slučaj harmoničnega stacionarnega stanja pri eni frekvenci tako, da krajevno-časovno funkcijo preslikamo v frekvenčni prostor, kar da Helmholtzovo enačbo:

$$(\Delta_{\mathrm{xyz}} + k^2)\psi = 0 \tag{2}$$

Vzemimo za razvoj numerične metode pravokotni valovod, ki se lepo prilega kartezičnemu koordinatnemu sistemu. Tedaj lahko pišemo:

$$\psi = \psi(x,y,z) = \psi(x,y)e^{-j\beta_z z} \tag{3}$$

Po vstavljenju zgornje enakosti v enačbo (2) dobimo valovodno valovno enačbo:

$$(\Delta_{xy} + k^2 - \beta_z^2)\psi = 0, \tag{4}$$

pri čemer je k valovno število v danem mediju (običajno zrak) in β_z njegova aksialna komponenta.

Ker lahko v valovodu skladno s teorijo ravninskega vala v splošnem obstajata dva načina razširjanja

valovanja, tj. glede na neizraženost longitudinalnih komponent bodisi vektorja električne poljske jakosti (TE način) bodisi vektorja magnetne poljske jakosti (TM način), in njima pripadajoči rodovi, rešujemo enačbo (4) za vsakega od načinov posebej. V nadaljevanju se osredotočimo na dominantni, osnovni rod TE, ki se kot prvi pojavi pri najnižji frekvenci. Od tod dalje se tako spremenljivka oz. neznanka ψ nanaša na vzdolžno komponento vektorja magnetne poljske jakosti H_z.

Vpeljimo sedaj neekvidistantno diskretizacijsko mrežo vozlov in jo napnimo na presek valovoda. Odslej torej računamo s končnim številom točk v prostoru, razmaknjenih za korak δx oz. δy ter indeksiranih s parom indeksov (i,j), ki pritičejo zaporednima diskretnima korakoma vzdolž koordinat x in y.

$$\psi(x_i, y_j) \rightarrow \psi_{i,j} \qquad i = 1,2,3,\dots,N \qquad j = 1,2,3,\dots,M \tag{5}$$

Druga parcialna odvoda po obeh transverzalnih prostorskih spremenljivkah nadomestimo z ustreznimi numeričnimi približki, ki jih izpeljemo iz diferenčnih kvocientov v matematični definiciji odvoda.

$$\frac{\partial\psi(i_+,j)}{\partial x} = \left.\frac{\partial\psi}{\partial x}\right|_{x=\frac{x_i+x_{i+1}}{2},\, y=y_j} \approx \frac{\psi_{i+1,j}-\psi_{i,j}}{\delta x} \tag{6}$$

$$\frac{\partial\psi(i_-,j)}{\partial x} = \left.\frac{\partial\psi}{\partial x}\right|_{x=\frac{x_{i-1}+x_i}{2},\, y=y_j} \approx \frac{\psi_{i,j}-\psi_{i-1,j}}{\delta x} \tag{7}$$

$$\frac{\partial\psi(i,j_+)}{\partial y} = \left.\frac{\partial\psi}{\partial y}\right|_{x=x_i,\;\; y=\frac{y_j+y_{j+1}}{2}} \approx \frac{\psi_{i,j+1}-\psi_{i,j}}{\delta y} \tag{8}$$

$$\frac{\partial\psi(i,j_-)}{\partial y} = \left.\frac{\partial\psi}{\partial y}\right|_{x=x_i,\;\; y=\frac{y_{j-1}+y_j}{2}} \approx \frac{\psi_{i,j}-\psi_{i,j-1}}{\delta y} \tag{9}$$

$$\frac{\partial^2\psi(i,j)}{\partial x^2} \approx \frac{\frac{\partial\psi(i_+,j)}{\partial x}-\frac{\partial\psi(i_-,j)}{\partial x}}{\delta x} \tag{10}$$

$$\frac{\partial^2\psi(i,j)}{\partial y^2} \approx \frac{\frac{\partial\psi(i,j_+)}{\partial y}-\frac{\partial\psi(i,j_-)}{\partial y}}{\delta y} \tag{11}$$

Zapišimo sedaj enačbo (4) s pomočjo pridobljenih približkov:

$$\frac{\psi_{i+1,j}+\psi_{i-1,j}-2\,\psi_{i,j}}{\delta x^2} + \frac{\psi_{i,j+1}+\psi_{i,j-1}-2\,\psi_{i,j}}{\delta y^2} + \\ +(k^2-\beta_z^2)\psi_{i,j} = 0 \tag{12}$$

$$\frac{\delta y}{\delta x}\left(\psi_{i+1,j}+\psi_{i-1,j}\right) + \frac{\delta x}{\delta y}\left(\psi_{i,j+1}+\psi_{i,j-1}\right) + \\ +\left(\delta x\delta y(k^2-\beta_z^2) - 2\left(\frac{\delta y}{\delta x}+\frac{\delta x}{\delta y}\right)\right)\psi_{i,j} = 0 \tag{13}$$

Enačba (13) velja za notranje vozle mreže, medtem ko terjajo robni vozli ločeno obravnavo – v duhu mejnih pogojev za oba poljska vektorja – šele njih poznavanje pa sploh dovoljuje obstoj rešitve. Iz Maxwell-ovih enačb v integralni obliki izhaja, da normalna komponenta vektorja magnetne poljske jakosti ($\vec{H}$) preko meje prehaja zvezno, torej mora na meji z idealnim prevodnikom biti enaka nič. Iz tega skladno s splošnima valovodnima enačbama sledi, da je normalni odvod aksialne komponente vektorja $\vec{H}$ ob stenah valovoda enak nič:

$$\nabla \mathrm{H_z} \cdot \vec{1}_\mathrm{n} = 0 \tag{14}$$

Če to prevedemo v numeriko za primer zgornjega oz. levega robu, dobimo:

$$\left.\frac{\partial^2\psi}{\partial x^2}\right|_{x=0} \approx \frac{\psi_{2,j}+\psi_{0,j}-2\psi_{1,j}}{\delta x^2} = \frac{2\psi_{2,j}-2\psi_{1,j}}{\delta x^2} \tag{15}$$

$$\left.\frac{\partial^2\psi}{\partial y^2}\right|_{y=0} \approx \frac{\psi_{i,2}+\psi_{i,0}-2\psi_{i,1}}{\delta y^2} = \frac{2\psi_{i,2}-2\psi_{i,1}}{\delta y^2} \tag{16}$$

Ko naposled zapišemo ustrezno enačbo za vsakega izmed vozlov mreže, pridobimo sistem n enačb z n neznankami, kjer je n enak številu vseh vozlov mreže $n = N\times M$. Če ga zapišemo v matrični obliki, dobimo kvadratno matriko koeficientov dimenzij $n \times n$, pomnoženo s stolpičnim vektorjem neznank, ter stolpični vektor znanih vrednosti – ničel – na destni strani enačaja.

Diagonalni elementi matrike koeficientov so enaki $\delta x\delta y(k^2-\beta_z^2) - 2\left(\frac{\delta y}{\delta x}+\frac{\delta x}{\delta y}\right)$. Če matriko koeficientov razstavimo na razliko dveh matrik, tako da iz le-te izvzamemo nekonstantni del diagonalnih členov, torej

$$\alpha = \delta x\delta y(k^2-\beta_z^2) \tag{17}$$

dobimo:

$$(\alpha\boldsymbol{I}-\boldsymbol{A})\boldsymbol{\psi} = \boldsymbol{0} \tag{18}$$

V enačbi (18) prepoznavamo enačbo lastnih vrednosti in lastnih vektorjev. Rešitev le-te nam postreže s pari lastnih vrednosti α in lastnih vektorjev $\boldsymbol{\psi}$, pri čemer iz posamezne α izluščimo $\beta_z(\omega)$ kot funkcijo frekvence, posamezen $\boldsymbol{\psi}$ pa nam da diskretno porazdelitev iskane poljske jakosti. Posamezna rešitev nam pravzaprav razkriva karakteristike posameznega rodu valovanja v valovodu. Ko pri danem α vsebujočo aksialno komponento faznega števila postavimo na nič $\beta_z = 0$, dobimo kritično frekvenco ω_k, tj. frekvenco, pri kateri je posamezni rod vzbujen; nad to frekvenco se rod v smislu hitrosti propagacije po valovodu pokorava krivulji, ki jo v t.i. disperzijski diagram zarisuje funkcija $\beta_z(\omega)$, pod to frekvenco pa rod ne obstaja.

2.2 Izračun in prikaz disperzijskega diagrama za pravokotni valovod – prirobnico antene

Vzemimo v obravnavo prirobnico valovodne antene, piramidnega lijaka, ki jo bomo obravnavali tudi v nadaljevanju tega pisanja.

Tabela 1. Podatki o valovodu in parametri za numerično obravnavo

Tip prirobnice	WR-90
Širina (a)	22,86 mm
Višina (b)	10,16 mm
Št. vozlov v x smeri (m)	80
Št. vozlov v y smeri (n)	80
Korak mreže v x smeri (δx)	0,28575 mm
Korak mreže v y smeri (δy)	0,12700 mm

Za potrebe nadaljnih izvajanj si dovolimo poenostavitev; in sicer privzemimo, da so stene valovoda kakor tudi celotne antenske strukture iz idealnega prevodnika,

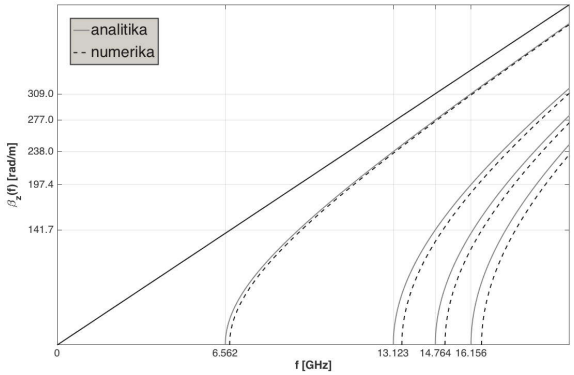
znotraj katerega sta obe poljski jakosti ničelni. V resnici se polje zažira reda nekaj vdornih globin v prevodnik, kar za primer bakra v mikrovalovnem področju znaša kvečjemu nekaj $10\mu m$.

Za TE_{mn} rodove disperzijski diagram analitično izračunamo po naslednjem obrazcu:

$$\beta_z(\omega) = \sqrt{\omega^2\mu\varepsilon - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{19}$$

Numerični račun disperzijskega diagrama izpeljemo iz (17), in sicer:

$$\beta_z(\omega) = \sqrt{\omega^2\mu\varepsilon - \frac{\alpha}{\delta x \delta y}} \tag{20}$$



Slika 1. Primerjava numerično in analitično pridobljenega disperzijskega diagrama za prve štiri TE rodove

2.3 Izračun in prikaz porazdelitve polja za dominatni rod v transverzalni ravnini valovoda

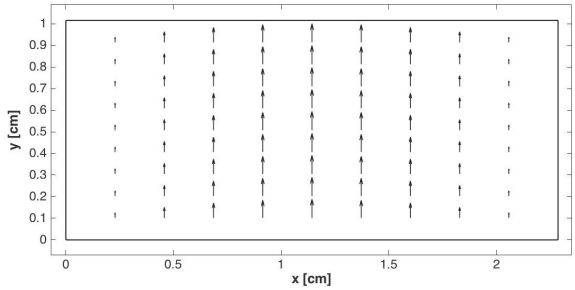
Elemente vektorja $\boldsymbol{\psi}$ moramo po izračunu najprej spraviti iz 1D zapisa (stolpičnega vektorja) nazaj v 2D zapis (matriko), ki pravzaprav razkriva porazdelitev aksialne komponente ustrezne poljske jakosti po prerezu valovoda v diskretnih točkah. Pot do transverzalnih komponent poljskih vektorjev je določena s splošnima valovodnima enačba, ki ju oklesteni za primer TE valovanja zapišimo numerično:

$$\vec{E}_t \cdot \vec{1}_x = \frac{\partial H_z}{\partial y} \approx \frac{H_{z\,i,j+1} - H_{z\,i,j-1}}{2\delta y} \tag{21}$$

$$\vec{E}_t \cdot \vec{1}_y = -\frac{\partial H_z}{\partial x} \approx \frac{H_{z\,i-1,j} - H_{z\,i+1,j}}{2\delta x} \tag{22}$$

$$\vec{H}_t \cdot \vec{1}_x = \frac{\partial H_z}{\partial x} \approx \frac{H_{z\,i+1,j} - H_{z\,i-1,j}}{2\delta x} \tag{23}$$

$$\vec{H}_t \cdot \vec{1}_y = \frac{\partial H_z}{\partial y} \approx \frac{H_{z\,i,j+1} - H_{z\,i,j-1}}{2\delta y} \tag{24}$$



Slika 2. Prikaz transverzalne porazdelitve vektorja električne poljske jakosti za dominantni rod TE_{10}

3 Smerni diagram valovodne antene

Vzemimo najprej za primer odprti konec valovoda, ki, tako kot katera koli druga oblika prenosnega voda, vodi in prostorsko bolj ali manj omejuje valovanje od svojega začetka proti koncu. Podobno kot velja za odprte sponke prenosne linije (npr. nezaključeni prostozračni dvovod ali koaksialni kabel), da imamo na njenem koncu opravka z nezveznostjo impedance, velja tudi za valovod, vendar se tukaj večji del energije izseva; le majhen del valovanja se odbije nazaj po valovodu. Za razliko od klasičnega koncepta odprtih sponk, kjer impedanca napravi skok z neke značilne končne vrednosti na idealizirano neskončno vrednost, se valovod s svojo karakteristično (valovno) impedanco že sam po sebi posrečeno približa impedanci praznega prostora.

Če valovod sedaj postopno razširimo v lijak z ustrezno presečno geometrijo, dobimo t.i. valovodno lijakasto anteno, ki skoraj vso energijo napredujočega vala v valovodu izseva v prostor. Kakšno je *postopno* razširjenje in kakšna *ustrezna* geometrija sta vprašanji, ki presegata okvir tega pisanja. V idealnem pa na odprtini lijaka dobimo ravninski val, katerega valovne fronte so vzporedne ekvifazne ravnine z izračunljivo amplitudno porazdelitvijo, ki jih po Huygens-Fresnelovem principu obravnavamo kot mrežo diferencialnih sekundarnih, Huygensovih izvorov, kateri v prostor skladno s superpozicijo izsevajo energijo v valovodu napredujočega vala. Huygensovi izvori so koherenti izvori in kot takšni dajejo prostorsko porazdelitev sevanega polja v smislu značilnega interferenčnega vzorca, ki v splošnem bolj ali manj izrazito odlikuje smerni diagram vsake valovodne antene.

Izkaže se, da se »(ne)uspešnost« pri skonstruiranju lijaka kaže ravno skozi obliko smernega diagrama (globina minimumov, širina snopov), sicer pa imamo za ta namen ustrezno metriko fazne napake, ki opredeljuje maksimalni relativni zasuk faze na odprtini antene.

3.1 Smerni diagram Huygensovega izvora ter smerni diagram valovodne lijakaste antene

Huygensov izvor je zgolj teoretični računski pripomoček, v resnici pa v svoji prvobitni obliki ne obstaja. Kot takšen seva anizotropno sevano polje in mu pripada smerni diagram, ki ga do sorazmernostne konstante natančno opredelimo z izrazom

$$dE(\vec{r}) \propto \frac{E_0(x,y)}{r}\cos(kr)\,(1 + cos\theta)\,\mathrm{dx\,dy} \tag{25}$$

Izraz je med drugim razbremenjen časovne odvisnosti, saj ta na amplitudna razmerja posameznih prispevkov v točki računanja (zanimanja) tako nima vpliva, zanima nas tam le fazni zasuk sevanega polja posameznega prispevka, ki pa je funkcija ustreznih prostorskih spremenljivk oz. preprosto razdalje od izvora r in pa frekvence, implicitno vsebovane v valovnem številu k. Predvsem fazna in v manjši meri tudi amplitudna razmerja so namreč tista, ki nam v točkah računanja smernega diagrama (sfera s polmerom R) dajejo bodisi konstruktivno bodisi destruktivno interferenco.

Zapišimo sedaj smerni diagram valovodne antene pravokotnega prereza, ki ga dobimo analitično, potom integracije izraza (21):

$$F(\theta,\phi) = (1+cos\theta)\left[\left(\frac{\sin\left(\frac{kb}{2}sin\theta cos\phi\right)}{\frac{kb}{2}sin\theta cos\phi}\right)\left(\frac{\cos\left(\frac{ka}{2}sin\theta cos\phi\right)}{1-\left(\frac{ka}{\pi}\right)^2 \sin^2\theta\cos^2\phi}\right)\right], \quad (26)$$

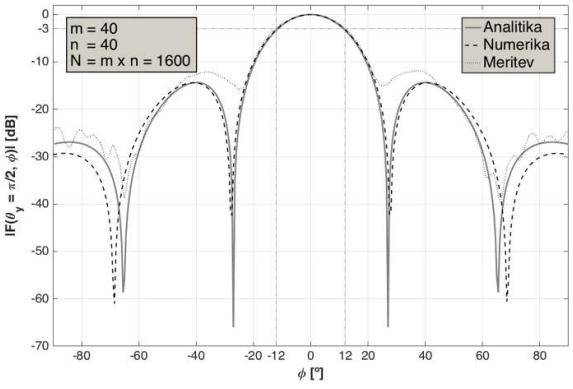
pri čemer sta a in b dimenziji odprtine lijaka.

3.2 Numerični izračun in prikaz smernega diagrama obravnavane valovodne antene

Sevanje valovodne antene v prostor je po Huygensu v celoti določeno s porazdelitvijo polja na odprtini lijaka. Kako do slednjega rezultata pridemo numerično, je podrobno opisano v razdelku 2, za pridobitev smernega diagrama nam tako ne preostane nič drugega, kot da seštejemo kvečjemu vse prispevke k sevanem polju s strani tistih končno točk na odprtini lijaka, kjer je polje pač znano. Polje na odprtini je sedaj diskretna funkcija dveh prostorskih spremenljivk, velikost Huygensovih izvorov pa končno majhnih izmer. Postopek temelji na izrazu (21) in je sledeča:

$$\tilde{F}(\theta_m,\phi_n) = (1+\cos\theta_m) \sum_{i=1}^{X}\sum_{j=1}^{Y}\frac{E_0[i,j]}{r_{i,j}(\theta_m,\phi_n)}\cos\left(k\, r_{i,j}(\theta_m,\phi_n)\right)\delta x\,\delta y \quad (27)$$

Zgornji izraz (23) je zavoljo njegove univerzalnosti smiselno normirati na njegovo vršno vrednost. Pogosto ga postavimo še v logaritemsko merilo in ga v dveh ravninah, v vsaki posebej, predstavimo v kartezičnem ali polarnem koordinatnem sistemu.



Slika 3. Prikaz smernega diagrama valovodnega lijaka v E ravnini, pridobljenega analitično, numerično ter z meritvijo (mreža 40 x 40 točk)

Za konec se vprašajmo še, kako dober in s tem uporaben je sploh naš numerični rezultat. Za vrednotenje tega se poslužimo metrike srednje kvadratne napake (angl. Mean squared error – MSE), ki bo kot vhodni argument vzela istoležne vrednosti numeričnega in analitičnega rezultata smernega diagrama v eni od ravnin:

$$MSE\left(\tilde{F},F\right) = \frac{1}{N}\sum_{n=1}^{N}\left(F(\phi_n)-\tilde{F}(\phi_n)\right)^2 \quad (28)$$

Tabela 2. Prikaz metrike MSE pri primerjavi numeričnega in analitičnega rezultata za smerni diagram antene

MSE [10^{-4}]				
n \ m	20	40	60	80
20	10,06	3,457	2,240	1,817
40	10,15	3,457	2,243	1,819
60	10,16	3,510	2,243	1,842
80	10,16	3,460	2,244	1,820

Tabela 2 prikazuje metrične vrednosti MSE za primer različno goste diskretizacijske mreže. Očitno se točnost rezultata numerične metode povečuje zlasti vzdolž vrstic tabele, tj. z zgoščevanjem diskretizacijske mreže po y-osi; to seveda velja za smerni diagram v E ravnini – obratno bi veljalo za H ravnino. Ta ugotovitev se kaže kot dobro izhodišče za optimizacijo računske zahtevnosti naloge računanja smernega diagrama v eni ravnini.

4 Zaključek

Z nekaj matematične spretnosti smo razvili numerično metodo za računanje polj v kovinskem votlovodu. Uporabili smo metodo končnih razlik v frekvenčnem prostoru (FDFD), ki daje vpogled v elektromagnetno dogajanje znotraj valovoda pri posamezni frekvenci. Numerični pristop nam za razliko od analitičnega ponuja možnost razširjenja obravnave na neenostavne geometrije, kar poenostavlja načrtovanje in konstruiranje naprednejših valvodnih anten.

Za razširjanje valovanja po valovodu je značilen fenomen (rodovne) disperzije, ki je po večini nezaželen pojav. Slednji je za prve štiri TE rodove predstavljen v disperzijskem diagramu, ki smo ga z razvojem numerične metode pridobili tudi numerično.

Posamezne rodove na eni strani okarakteriziramo z adekvatnimi krivuljami v disperzijskem diagramu, na drugi strani pa jih prepoznavamo tudi skozi značilno pripadajočo porazdelitev poljskih vektorjev po prerezu valovoda, ki smo jo za osnovni dominantni rod TE pridobili numerično (Slika 2).

Z virtualnim razširjenjem transverzalne poljske slike TE_{10} rodu v valovodu na izmere odprtine obravnavanega piramidnega lijaka smo na numerični način pridobili tudi pripadajoči smerni diagram v 2D in ga v preseku z E ravnino prikazali na sliki (Slika 3). Na slednji se k primerjavi pridružujeta še praktično izmerjena in pa analitična inačica; primerjava s slednjo je metrično ovrednotena v tabeli 2.

Literatura

[1] Antenna theory: Analysis and Design; Balanis, Constantine A., 3rd ed., Hoboken: Wiley-Interscience, 2005

[2] Elektromagnetno valovanje; Mlakar, Jože, 2. izd., Ljubljana: Fakulteta za elektrotehniko, 2002

[3] ELMG polje; Sinigoj, Anton R., 1. izd., Ljubljana: Fakulteta za elektrotehniko, 2002

[4] Antene in razširjanje valov; Matjaž Vidmar, Ljubljana, 2017

Integracija naprav za merjenje spremenljivosti srčnega utripa (HRV) voznika v simulator vožnje

Timotej Gruden

Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana
E-pošta: timotej.gruden.si@ieee.org

Integration of devices for measuring drivers' heart rate variability (HRV) in a driving simulator

Abstract. *In this paper I present the process of integrating two different devices for measuring drivers' physiological signals into a driving simulator. Heart rate variability (HRV) is found to be a promising parameter in describing drivers' state. We used Bittium's Faros 360 and Empatica's E4 devices for measuring HRV through ECG or PPG respectively.*

The NERVteh's driving simulator demanded smooth process of starting / ending measurements and simulation. Both Empatica and Faros 360 have different operating modes of which both provide an option for streaming data via Bluetooth. Application Programming Interfaces (APIs) for communication between systems were defined and prototypes for integration were developed. Our work was done in C#.

At the end we tested our integration with a user study that reported on measuring drivers' physiological signals. No problems with integration were found.

1 Uvod in pregled področja

Analiza in pregled stanja varnosti v cestnem prometu za leto 2017 [1] kažeta, da smo imeli na cestah Republike Slovenije v zadnjih petih letih med 17500 in 19000 prometnih nesreč vsako leto. Izmed teh smo samo v zadnjem letu (2017) zabeležili 104 prometne nesreče s smrtnim izidom in 7901 nesrečo z lažjimi ali hujšimi telesnimi poškodbami. Z vidika števila prometnih nesreč in posledic je bilo leto 2017 »najvarnejše« izmed zadnjih petih let. Javna agencija Republike Slovenije za varnost prometa (AVP) navaja celo, da je bilo število umrlih v preteklem letu najmanjše v zadnjih 60 letih.

K postopnem zmanjševanju števila nesreč in posledičnem izboljšanju varnosti v cestnem prometu gotovo močno pripomore razvoj novih tehnologij. Študije kažejo, da lahko s proučevanjem podatkov o voznikih (demografski podatki, fiziološki signali, upoštevanje predpisov, navade ipd.) definiramo različne razrede varnosti voznikov. [2] primerja demografske podatke z uporabo varnostnih mehanizmov v avtomobilu (varnostni pas, omejitev hitrosti, …) in zavrača mnogo stereotipov. Študija [3] v simulatorju vožnje je pokazala, da imajo ljudje z motnjo pozornosti zaradi hiperaktivnosti (ADHD) več možnosti za povzročitev prometne nesreče kot ostali. Študija [4] z uporabo mobilnega telefona zajema parametre vožnje posameznika in ga ob tem opozarja na določene kritične trenutke. Z zajemom fizioloških signalov voznikov želijo raziskovalci zaznati predvsem vpliv stresa na vožnjo [5] in ga na različne načine analizirati [6]. Fiziološki signali (predvsem EEG in EKG) so uporabni tudi za merjenje utrujenosti voznikov, ki bi lahko privedla do prometne nesreče [7].

Skratka, različnim meritvam signalov se pri analizi in profiliranju voznikovega obnašanja ne moremo izogniti. Poleg meritev, ki jih že opravlja vozilo samo (hitrost, čas vožnje, …), so nujno potrebne tudi meritve signalov voznikov. Med njimi najpogosteje merimo srčno aktivnost (srčni utrip, spremenljivost srčnega utripa – HRV), možgansko aktivnost, prevodnost kože, telesno temperaturo in velikost zenice.

Za proučevanje srčne aktivnosti se v zadnjem času vedno pogosteje pojavlja spremenljivost srčnega utripa (angl. *Heart Rate Variability – HRV*). Ta se izkazuje kot najuspešnejša nemoteča metoda za merjenje delovanja avtonomnega živčnega sistema, še posebej parasimpatičnega živčevja [8]. Raziskave HRV signalov so že tako razširjene, da obstaja mnogo priporočil, kako kar najbolje načrtovati študije, katere spremenljivke kažejo na statistično pomembne razlike ipd. [9].

V pričujočem delu predstavljam, kako smo naprave za merjenje HRV integrirali v simulator vožnje. V sodelovanju s podjetjem NERVteh [10] želimo namreč profilirati različne vzorce obnašanja voznikov skozi simulacijo. Uporaba simulacije za poustvarjanje kritičnih dogodkov je povsem na mestu, saj je izpostavljanje voznikov nevarnostim v resničnem vozilu praktično neizvedljivo ne samo zaradi nevarnosti temveč tudi zaradi dejstva, da se kritične situacije dogajajo ob časovno nepredvidljivih trenutkih.

Za razumevanje HRV signala v naslednjem poglavju na hitro opisujem delovanje srčne mišice in metode za njegovo merjenje, nato predstavljam uporabljeni dve napravi za merjenje HRV, v četrtem poglavju pa povzemam zahteve za integracijo naprav v simulator. Peto poglavje opisuje dogovorjeni način uporabe vmesnikov za komunikacijo med simulatorjem in sistemom za zajem fizioloških signalov, sledeče (6.) pa predstavlja njihovo izvedbo. V zaključnem poglavju je podanih še nekaj možnosti za nadaljnje delo.

2 Delovanje srčne mišice

Bistvena posledica delovanja srca je pošiljanje oz. razširjanje krvi po žilah. Pri tem se srčna mišica krči in razteza ter poleg tega oddaja električen signal značilne oblike.

2.1 Elektrokardiografija (EKG)

Že davno odkrita metoda za merjenje lastnosti delovanja srca je elektrokardiografija (EKG). Navadno z večjim številom elektrod, postavljenih okoli srca opazovanca, zaznavamo spremembe v električnem potencialu – električno napetost.

V EKG signalu se periodično z vsakim utripom pojavlja značilen PQRST vzorec. Posamezna črka predstavlja posamezno točko v signalu. Za nas je tokrat pomemben samo vrh posameznega vzorca – točka R.

2.2 Optična pletizmografija (PPG)

Druga možnost za merjenje srčne aktivnosti je novejša metoda optične pletizmografije (angl. *photoplethysmography – PPG*), ki z meritvijo absorpcije svetlobe (navadno različnih barv) v arterijah meri pulz krvne prostornine (angl. *Blood Volume Pulse – BVP*). Meritev se zaradi neinvazivnosti najpogosteje izvaja na zapestju. V signalu je možno zaznati vrhove in doline, ki predstavljajo posamezne udarce srca.

2.3 R–R interval in IBI

Kot že omenjeno, lahko z namenskimi algoritmi v vzorcu EKG signala določimo posamezne značilne točke. Iz njih lahko izračunamo posamezne intervale, ki se v raziskavah pojavljajo bodisi kot odvisne bodisi kot neodvisne spremenljivke. Nam najpomembnejši signal je zaporedje dolžin časovnih intervalov med dvema točkama R, tj. R–R intervalov. Obratna vrednost teh predstavlja trenutno frekvenco srčnega utripanja.

Naprave, ki ne merijo EKG signala, navadno podajajo interval med udarci (angl. *Inter-Beat Interval – IBI*), ki ga pridobijo z metodo PPG ali kako drugače. Lastnosti IBI so primerljive R–R intervalom, niso pa popolnoma enake (lahko pride do razlik zaradi različne dolžine poti krvi, elastičnosti žil ipd.).

2.4 Spremenljivost srčnega utripa (HRV)

Pomembno je omeniti tudi, da interval med zaporednimi udarci srca še zdaleč ni konstanten. Ta pojav imenujemo spremenljivost srčnega utripa (angl. *Heart Rate Variability – HRV*). Zanimiva značilnost HRV signala je, da se R–R intervali krajšajo med vdihom in daljšajo med izdihom zraka (respiratorna sinusna aritmija).

3 Napravi za merjenje HRV

Za uporabo v simulatorju vožnje na čim večjem številu ljudi je še posebej pomembno, da uporabljene naprave omogočajo neinvazivne tehnike merjenja. Pri načrtovanju integracije smo se zato odločili za Faros 360 s tekstilnim trakon, ki ga preprosto namestimo pod prsi, in Empatico E4, ki jo namestimo na zapestje nedominantne roke voznika.

3.1 Faros 360

Produkt finskega podjetja Bittium, tj. Faros [11], je majhna naprava, ki jo pritrdimo na nosljive elektrode. Prav zaradi svoje majhnosti jo lahko povsem neopazno nosimo pod obleko skozi vsakdan. Proizvajalec ponuja tri različice naprave: 90, 180 in 360; uporabili smo le Faros 360, saj samo ta ponuja komunikacijo preko bluetooth vmesnika.

Faros vsebuje EKG senzor, ki lahko deluje s frekvenco vzorčenja do 1000 Hz, pospeškometer s frekvenco vzorčenja do 100 Hz in temperaturni senzor. Poleg surovih podatkov podaja še R–R intervale, izmerjene v milisekundah.

3.2 Empatica E4

Med napravami, ki se pritrdijo na zapestje, je Empatica E4 [12] ena izmed redkih raziskovalnih naprav s certifikatom za uporabo v medicinske namene. Vsebuje 3-osni pospeškometer s frekvenco vzorčenja 32 Hz, optični infrardeči temperaturni senzor s frekvenco vzorčenja 4 Hz, posrebreni elektrodi za merjenje galvanskega odziva (električne prevodnosti) kože in senzor optične pletizmografije (PPG). Z njim meri BVP in iz njega preračunava IBI, v kolikor je to mogoče.

4 Zahteve za integracijo

Delovanje simulatorja vožnje NERVteh mora biti zaradi uporabe naprav za merjenje HRV popolnoma nemoteno. Pred vstopom v simulator uporabnikom namestimo napravi in ju prižgemo. Po tem z njima ne smemo biti več v fizičnem stiku, dokler ju uporabnik ne preneha uporabljati.

Tudi sam sistem za delovanje simulatorja naj bi bil pripravljen po filozofiji »one-click«, kjer lahko z enim klikom brez inženirskega poznavanja delovanja tehnologije sprožimo celotno simulacijo, ki vključuje tudi merjenje fizioloških signalov. Potrebi po nameščanju naprav in prižigom vsake posebej se na žalost ne moremo popolnoma izogniti, lahko pa ju poskusimo kar se da olajšati.

4.1 Možnosti simulatorja

Simulator NERVteh uporablja program Oktal Scanner Studio [13] za izvajanje simulacije in različnih scenarijev. Za zagon in upravljanjem z njim so inženirji pri NERVtehu razvili svojo rešitev, poimenovano OktalControl. Izvajalni operacijski sistem je Winodws 10, OktalControl je bil razvit v .NET tehnologiji (C#). Programski vmesnik za komunikacijo s simulacijo (Oktal) je izdan samo za jezik C#.

4.2 Možnosti Farosa 360

Faros ponuja dva načina delovanja. Prvi, nepovezani način, shranjuje izmerjene surove podatke na notranji pomnilnik naprave. Po končani meritvi jih moramo ročno preko USB prenesti na računalnik in pretvoriti v format, primeren za nadaljnjo analizo.

Drugi način omogoča povezavo Farosa in simulatorja s tehnologijo bluetooth. Napravi se brez težav seznanita, na računalniku se ustvarijo nova navidezna zaporedna vrata za komunikacijo (COM port). Komunikacijski protokol je podrobneje opisan v Bittiumovem dokumentu št. 810378. Z računalnika

lahko nastavimo parametre meritev (merjene veličine in frekvence vzorčenja) ter zaženemo meritev. Faros nato vsakih 200 ms pošlje paket z do tedaj zbranimi surovimi podatki.

Ostalih omejitev glede uporabe programskih jezikov oz. operacijskih sistemov ni, saj je za uporabo naprave potrebno le pošiljanje in prejemanje podatkov z navideznih zaporednih vrat. Potrebuje pa računalnik delujoč bluetooth vmesnik.

4.3 Možnosti Empatice E4

Podobno kot Faros, tudi Empatica E4 ponuja način brez povezave, t. i. snemalni način. Ta se vključi, če v roku 20 sekund po prižigu naprava ne zazna nobene vhodne povezave. Podatke iz snemalnega načina preko USB z uporabo namenske aplikacije E4 manager prenesemo na spletno platformo E4 connect od koder jih lahko prenesemo na računalnik v obliki csv datoteke.

Empatico E4 lahko uporabljamo tudi v kombinaciji z mobilnimi napravami. Na voljo sta aplikaciji, ki podatke prenašata v spletno platformo, ter razvojni orodji (angl. *Software Development Kit – SDK*) za Android in iOS. Povezava z mobilnimi napravami poteka preko BLE – Bluetooth Low Energy. Za razvoj lastne aplikacije mora razvijalec pridobiti API ključ, ki se preveri ob vsakokratni povezavi Empatice z mobilno napravo. Slaba stran tega je, da je internetna povezljivost posledično pogoj za vzpostavitev povezave z napravo.

Tretja možnost je povezava Empatice E4 z Windows računalnikom preko posebnega BLE vmesnika [14] in namenskega programa (pretočnega strežnika). Na pretočni strežnik se lahko povežemo s predhodno nastavljenimi TCP vrati. Kot omenjeno, za uporabo te možnosti potrebujemo operacijski sistem Windows.

5 Aplikacijski programski vmesniki (APIs)

Za neodvisen razvoj posameznih elementov integracije in zagotovilo delovanja končnega produkta je potrebno dobro definirati vmesnike in postopke komunikacije tako med napravami kot tudi med deli programov.

5.1 Razredi »Manager«

Za vsako izmed naprav smo naredili svoj razred »Manager« (»FarosManager« in »EmpaWinManager«), ki upravlja s posamezno napravo. Ostali deli celote lahko do naprave dostopajo samo preko teh razredov.

Dogovorili smo se, da se ob prižigu simulatorja vožnje iz Manager razreda ustvari objekt, nad njim pa je potrebno za povezavo naprave (npr. ob pritisku gumba) klicati funkcijo »connect()«. Za začetek oz. konec meritve je določena posebna funkcija »startComm()« oz. »stopComm()«. Če ob koncu vožnje naprave ne bomo več uporabljali, lahko povezavo prekinemo s klicem funkcije »disconnect()«.

5.2 Vmesnik »Listener«

Določiti je potrebno še sprejem oz. shranjevanje pridobljenih podatkov. Podjetje je za takojšnjo obdelavo izrazilo željo po shranjevanju v realnem času oz. najkasneje ob zaključku posamezne vožnje.

Za vsako napravo smo se tako dogovorili za vmesnika (angl. *interface*) »Listener« (»FarosListener« in »Empa-Listener«), ki določata po eno funkcijo za vsako vrsto podatka, ki ga je poslala naprava. Razvijalec je moral nato sam izvesti vmesnik in tako v njem določiti, kaj bo storil s prebranimi podatki (npr. shranil).

Prej omenjena razreda Manager vsebujeta tudi funkciji »addListener(Listener)« oz. »removeListener (Listener)«. Za uporabo naj bi najprej v obliki novega objekta izvedli vmesnika in ju dodali razredoma Manager. Ta ob sprejemu vsakega podatka z naprave kliče ustrezno funkcijo na vseh dodanih vmesnikih.

5.3 Časovna sinhronizacija

Faros ne ponuja možnosti časovnih žigov, zagotavlja pa, da bo znotraj vsake sekunde poslal 5 paketov s točno določeno vsebino – podatki vnaprej določene dolžine. Empatica sicer sproti ob vsakem vzorcu meritve pošilja časovne žige, a je sinhronizacija le teh s simulatorjem zelo težka. Proizvajalec sicer trdi, da se uri sinhronizirata ob priklopu USB povezave, a po izkušnjah sodeč ostaneta sinhronizirani le za kratek čas.

V izogib težavam smo se dogovorili, da bo simulator (računalnik) glavni časovni izvor. Pri shranjevanju surovih podatkov iz merilnih naprav bomo ob prvem prejetem vzorcu zapisali čas s simulatorja, ostale časovne žige pa bomo preračunali glede na frekvenco vzorčenja posameznega podatka. Pri podatkih, kjer ta ni določljiva, se uporabi čas sprejetja na simulatorju.

6 Integracija

Glede na dejstvo, da je NERVtehov nadzorni program za simulacijo narejen v programskem jeziku C#, smo tudi ostale elemente za lažjo integracijo pripravili v istem jeziku.

6.1 FarosManager

Izdelani razred FarosManager predpostavlja, da med računalnikom in Farosom 360 že obstaja bluetooth povezava in z njo navidezna zaporedna vrata. Prej omenjena funkcija connect() se trenutno pokliče samodejno ob ustvarjenem novem objektu FarosManager. Ta preko zaporednih vrat odpre povezavo z bitno hitrostjo 115200 bit/s in poizve po identifikacijski številki naprave, ki jo zapiše v dnevniško datoteko.

Za predvideno delovanje moramo narediti še razred, ki bo vključeval vmesnik FarosListener. Ta predvideva dve možnosti pridobivanja in shranjevanja podatkov. Sprejema lahko celotne pakete vzorcev, ki jih Faros pošlje vsakih 200 ms, ali pa vzorce posameznih signalov, ki jih FarosManager že izlušči iz paketa. Prednost zadnjega načina je, da se razvijalcu ni potrebno še enkrat obremenjevati z nastavitvami naprave, številom vzorcev posameznega signala in izračunom CRC, za kar poskrbi FarosManager. Razred, ki vključuje vmesnik FarosListener, moramo seveda dodati FarosManagerju s funkcijo addFarosListener().

Ob klicu funkcije za začetek meritve FarosManager preko zaporednih vrat pošlje ukaz za začetek meritve v nepovezanem načinu in zažene bralno zanko, ki prebira

podatke in jih razporeja v pakete. Iz vsakega prebranega paketa glede na nastavitve naprave in specifikacijo protokola izlušči vzorce posameznega signala in jih posreduje vsem dodanim Listener objektom.

Funkcija za zaključek meritve pošlje Farosu ukaz za zaključek in prekine bralno zanko. Če naprave ne bomo več uporabljali, kličemo še funkcijo disconnect(), ki sprosti navidezna zaporedna vrata.

Poleg naštetih funkcionalnosti lahko FarosManager še preveri trenutne nastavitve naprave, pošlje nove nastavitve ali sproži utrip svetlečih diod na Farosu.

6.2 EmpaWinManager

Izdelani razred EmpaWinManager predpostavlja, da na računalniku teče namenski program Empatica BLE Server (pretočni strežnik), s katerim je Empatica E4 že povezana. Ob ustvarjenem objektu EmpaWinManager se vzpostavi TCP povezava na 127.0.0.1:28000, kjer je dostopen pretočni strežnik. Ob klicu funkcije connect() EmpaWinManager preveri, katere E4 naprave so povezane na pretočni strežnik, in se poskuša povezati na prvo izmed njih.

Ob klicu funkcije startComm() za začetek meritve se EmpaWinManager pri pretočnem strežniku naroči (angl. *subscribe*) na vse možne pretočne signale. Vmesnik EmpaListener predvideva ločene funkcije za vsakega izmed merjenih signalov, ki se kličejo ob vsakem sprejetem vzorcu. Podatki se vedno pošiljajo vsem dodanim vmesnikom.

Ob zaključku meritve se EmpaWinManager le odjavi (angl. *unsubscribe*) od prejemanja pretočnih podatkov. Funkcija disconnect() zapre TCP povezavo s pretočnim strežnikom, E4 pa nanj ostane povezana.

6.3 Testiranje delovanja

Integracijo obeh naprav skupaj smo prvič preizkusili v uporabniški študiji: Razlike med fiziološkimi signali voznikov v različnih situacijah [15]. Ta je pokazala, da komunikacija med napravami in njihova uporaba potekata gladko. Več težav povzroča sam program za nadzor simulacije, OktalControl, še več pa simulacijsko ogrodje Oktal Scanner Studio. Zadnja dva sta bila vzroka večine težav.

Poudariti je potrebno tudi, da kljub uspešni integraciji Empatica E4 za meritve HRV med vožnjo ni vedno najbolj primerna. Med mirovanjem voznika naprava podaja verodostojne rezultate meritev, med gibanjem (npr. zavijanjem, prestavljanjem, …) pa so podatki pomanjkljivi. Verjetno notranji algoritem za izračun IBI ni dovolj robusten, premiki pa vnesejo preveč motenj v BVP signal. Faros 360 ne povzroča nobenih težav.

7 Zaključek in nadaljnje možnosti

Za konec naj omenim še, da samo delujoča integracija dveh sistemov nikakor ni zadosten pogoj za delovanje sistema. Vedno se je potrebno zavedati omejitev, tako iz fizikalnega ozadja, kot zaradi človeka, ki bo dani sistem upravljal. Vsaka uporaba simulatorja vožnje z napravami za merjenje fizioloških signalov voznikov mora biti vnaprej dobro premišljena, da teh omejitev ne preseže.

V nadaljevanju želimo dobljeni produkt uporabiti za zajem množice podatkov večjega števila ljudi. Z nadaljnjo obdelavo bomo poskušali poiskati tiste lastnosti fizioloških signalov voznikov, ki čim bolje odražajo lastnosti vožnje oz. nanjo pomembno vplivajo.

Literatura

[1] Javna agencija Republike Slovenije za varnost prometa. Analiza in pregled stanja varnosti v cestnem prometu za leto 2017. https://www.avp-rs.si/ (30. 6. 2018).

[2] Shinar, D. et al. (2001). Self-reports of safe driving behaviors in relationship to sex, age, education and income in the US adult driving population. Accident Analysis & Prevention, 33(1), 111-116.

[3] Fischer, M. et al. (2007). Hyperactive children as young adults: Driving abilities, safe driving behavior, and adverse driving outcomes. Accident Analysis & Prevention, 39(1), 94-105.

[4] Fazeen, M. et al. (2012). Safe driving using mobile phones. IEEE Transactions on Intelligent Transportation Systems, 13(3), 1462-1468.

[5] Healey, J., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on intelligent transportation systems, 6(2), 156-166.

[6] Singh, R. R. et al. (2013). A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. Biomedical Signal Processing and Control, 8(6), 740-754.

[7] Roman, B. et al. (2001, October). Fatigue indicators of drowsy drivers based on analysis of physiological signals. In International Symposium on Medical Data Analysis (pp. 62-68). Springer, Berlin, Heidelberg.

[8] Sztajzel, J. (2004). Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system. Swiss medical weekly, 134(35–36).

[9] Laborde, S. et al. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting. Frontiers in psychology, 8, 213.

[10] NERVteh. https://www.nerv-teh.com/ (4. 7. 2018).

[11] Bittium Faros 360. https://www.bittium.com/ (3. 7. 2018).

[12] Empatica E4. https://www.empatica.com/research/e4/ (4. 7. 2018).

[13] Oktal. Simulation in Motion. http://www.oktal.fr/en/automotive/range-of-simulators/software (19. 6. 2018).

[14] Bluegiga BLED112 Bluetooth® Low Energy Dongle. https://www.silabs.com/products/wireless/bluetooth/bluetooth-low-energy-modules/bled112-bluetooth-smart-dongle (4. 7. 2018).

[15] Gruden, T. (2018). Uporaba komercialno dostopnih naprav za merjenje fizioloških signalov voznikov. Pridobljeno od https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=101771 (13. 7. 2018).

Porazdeljeno vodenje časovno kritičnega sistema zračne levitacije preko industrijskega protokola PROFINET

Ana Kerbler[1], Uroš Zajšek[1], Boštjan Mlakar[1]

[1]*Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,*
Koroška cesta 46, 2000 Maribor, Slovenija
E-pošta: ana.kerbler@student.um.si, uros.zajsek@student.um.si, bostjan.mlakar@student.um.si

Distributed management of the time critical air levitation system through the industrial protocol PROFINET

Povzetek. *Članek predstavlja raziskovalni projekt porazdeljenega zaprto-zančnega vodenja sistema zračne levitacije. Sistem zračne levitacije sestavljajo vetrovnik ter dva krmilnika SIMATIC S7-1200, ki sta med seboj povezana z industrijskim protokolom PROFINET. Cilj raziskovalnega projekta je preučiti ali je s pomočjo porazdeljenega vodenja, preko industrijskega protokola, možno stabilizirati in voditi časovno kritični sistem zračne levitacije in projekt fizično implementirati. Porazdeljeno vodenje smo zasnovali na način, da prvi krmilnik vodi zračni tok vetrovnika. Drugi krmilnik, pa izvaja regulacijo v realnem času in je zadolžen za merjenje višine lebdečega objekta, komunikacijo s prvim krmilnikom ter ostalimi zunanjim napravim, ki so ključne za delovanje sistema. Na tak način smo dosegli popolnoma porazdeljeno vodenje, ki je praktično neodvisno od lokacije vodenja sistema ter obdelave ter zajemanja podatkov iz sistema. Model objekta vodenja ter zaprto-zančno strukturo smo verificirali s pomočjo programskega paketa MATLAB. Nastavljene parametre regulatorja ter izbrano strukturo vodenja smo implementirali preko TIA portala, s katerim smo sprogramirali krmilnika SIMATIC S7-1200.*

1 Uvod

Z razmahom novih računalniških in informacijskih tehnologij, tako na področju komunikacij, kakor tudi sistemov realnega časa, se v zadnjem času veča potreba po porazdeljenih sistemih, ki so med seboj povezani toda dislocirani [1]. Porazdeljeni sistemi ter porazdeljeno vodenje predstavljata nov trend v razvoju industrijskih sistemov ter aplikacij. Porazdeljeno vodenje je digitalni avtomatiziran nadzorni sistem, ki uporablja geografsko porazdeljene krmilne zanke, ki vodijo in nadzirajo proizvodnjo, stroje in management podjetja. Za razliko od centraliziranega sistema vodenja, kjer vsak proces, stroj ali naprava potrebuje svojo krmilno enoto [2], [3]. Centraliziran sistem vodenja se nanaša na zaprto-zančno strukturo vodenja, pri čemer samostojna enota nadzoruje vhodno in izhodne enote ter izvaja procesiranje v realnem času. Porazdeljeno vodenje nudi mnoge prednosti ter je učinkovitejše od centraliziranega sistema, toda ima določene slabosti, katere so danes predmet mnogih raziskav ter inženirskega razvoja. Glavni problem takšnega sistema je zagotoviti zanesljivost komunikacijskih kanalov, ki morajo zagotavljati pravilne, varne in zanesljive ter pravočasne realno-časovne informacije v točno določenem časovnem intervalu. Problem je torej predvsem zagotoviti komunikacije z nizko zakasnitvijo (latenco), dovolj veliko bitno hitrostjo ter regulatorje (krmilnike), ki so sposobni voditi hitre dinamične, v osnovi nestabilne procese v realnem času.

Namen predstavljenega dela je raziskati porazdeljeno strukturo zaprto-zančnega vodenja časovno kritičnega sistema zračne levitacije. Sistem zračne levitacije je odprto-zančno nestabilni sistem, kar pomeni, da zaprto zančno vodenje mora zagotavljati vzbujanje sistema v natančno določenih časovnih intervalih. Vsaka nepredvidena zakasnitev preko komunikacijskega kanala ali sprememba časovnega intervala, oziroma časa odziva določenega podsistema lahko vodi v nestabilno delovanje celotnega sistema. Za naš sistem porazdeljenega vodenja, smo izbrali dva industrijska krmilnika SIMATIC S7-1200 ter komunikacijski protokol PROFINET. PROFINET je industrijski protokol, ki omogoča komunikacijo in prenos podatkov preko industrijskega Eterneta. Namenjen je za vodenje industrijskih sistemov in je primeren za časovno kritične sisteme s kritičnim časovnim oknom (zakasnitvijo, latenco) do 1 ms ali še manj.

Za zaprto-zančno vodenje sistema zračne levitacije smo najprej izpeljati matematični model objekta vodenja. Izpeljan model je idealiziran in ne vključuje motenj ter odstopanj modela. Ker motnje težko določimo ali izmerimo, jih ne moremo posredno vključit v model. Nastavitev vodenja v tem primeru zahteva določeno mero inženirskih izkušenj.

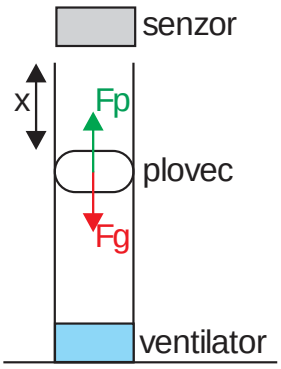
Delo, ki smo ga zvedli vključuje poznavanje modeliranja fizikalnih procesov, znanje daljinskega vodenja in zaprto-zančnega vodenja in industrijskih komunikacij. Sistem zračne levitacije smo izdelali sami po naslednjih korakih:

- raziskava in izpeljava matematičnega modela zračne levitacije,
- raziskava in načrtovanje zaprto-zančne strukture ter nastavitev PID- regulatorja,
- fizična izdelava sistema zračne levitacije z vetrovnikom,
- programiranje industrijskih krmilnikov,
- preizkus delovanja.

V članku bo v drugem poglavju predstavljen matematični model zračne levitacije, nato sledi poglavje tri, ki predstavlja porazdeljeno vodenje. V četrtem poglavju bodo predstavljeni rezultati vodenja, za katerim sledi zaključek.

2 Modeliranje sistema

Za izvedbo simulacij, v katerih želimo doseči čim bolj realno obnašanje modela in določiti parametre PID-regulatorja za zračno levitacijo plovca v vetrovniku, Slika 1., je uporabljen idealiziran lineariziran matematični model, ki je zapisan na osnovi 2. Newton-ovega zakona [4], [5].



Slika 1: Model zračne levitacije

Nelinearni model zračne levitacije:

$$m\ddot{x} = F_p - F_g, \\ = \frac{1}{2} C_d \cdot \rho \cdot A \cdot \left(v_{eq} - \dot{x}\right)^2 - m \cdot g. \quad (1)$$

Liberaliziran model, predstavljen s prenosno funkcijo (1) je:

$$G(s) = \frac{x(s)}{v_{eq}(s)} = \frac{1}{s} \cdot \frac{a \cdot K_V}{(s+a) \cdot (s \cdot T_V + 1)} \quad (2)$$

Posamezni parametri prenosne funkcije (1) so podani v naslednjih enačbah:

$$a = \frac{2 \cdot g}{v_{eq}} \quad (3)$$

$$v_{eq} = \sqrt{\frac{g \cdot m}{\alpha}} \quad (4)$$

$$\alpha = \frac{C_d \cdot \rho \cdot A}{2} \quad (5)$$

Kjer je:

- x – višina plovca,
- $\dot{x}$ – hitrost plovca,
- K_v – ojačenje ventilatorja,
- T_v – časovna konstanta ventilatorja,
- g – težnostni pospešek,
- m – masa plovca,
- C_d – viskoznost zraka,
- v_{eq} – hitrost ventilatorja, pri kateri je plovec v ravnovesju,
- A – površina plovca,
- ρ – gostota zraka.

Izdelan sistem vetrovnika ima naslednje določene parametre:

- m = 7 g,
- g = 9,81 m/s2,
- d = 41 mm,
- C_d = 1,568 10^{-5},
- ρ =1,184 kg/m3.

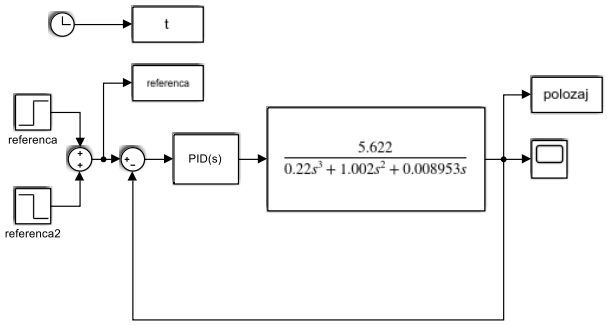
3 Vodenje zračne levitacije

Glede na izpeljan matematični model ter določene parametre modela, smo s pomočjo programa MATLAB določili parametre PID- regulatorja. Regulator smo nastavili na osnovi minimizacije performančnega učinka občutljivostne prenosne funkcije zaprto-zančnega sistema. Minimizacija občutljivostne performančne karakteristike nam zagotavlja zanesljivo vodenje v primeru delovanja nemerljivih izhodnih motenj ter odstopanje parametrov modela [6]. Določeni končni parametri regulatorja so: K_P = 0,06596, K_I = 0,00473, K_D = 0,23.

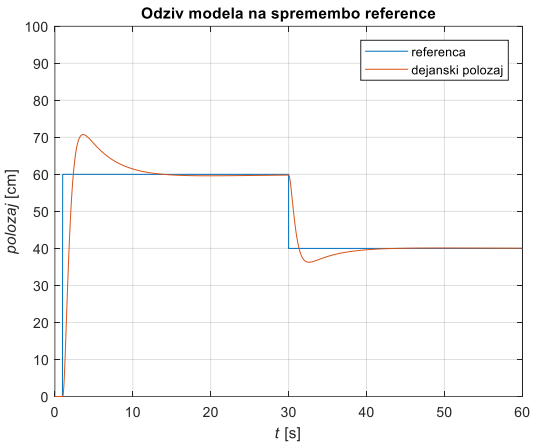
Simulacijska shema vodenja procesa je prikazana na sliki 2. Modeliran proces je sestavljen iz prenosne funkcije objekta vodenja ter PID- regulatorja. Odziv modela zračne levitacije je prikazan na sliki 3. Pri času 1 s stopnično spremenimo referenco višine na 60 cm. Odziv modela ima prenihaj in doseže referenčno vrednost pri času 14 s. Pri času 30 s stopnično zmanjšamo referenco iz 60 cm na 40 cm. V odzivu pride do podnihaja in po času 11 s ponovno doseže referenčno vrednost.

Izračunano prenosno funkcijo regulatorja smo uporabili v programu TIA PORTAL, kjer smo izvedli eksperimentalni del. V izbranem programskem okolju smo uporabili PID- blok s sledečo prenosno funkcijo (6), [7].

$$R(s) = K_P \cdot \left((b \cdot w - x) + \frac{1}{T_I \cdot s} \cdot (w - x) + \frac{T_D \cdot s}{a \cdot T_D \cdot s + 1} \cdot (c \cdot w - x) \right) \quad (6)$$



Slika 2: Simulacijski model procesa s PID regulatorjem.



Slika 3: Simuliran odziv modela vetrovnika na stopnično spremembo reference višine.

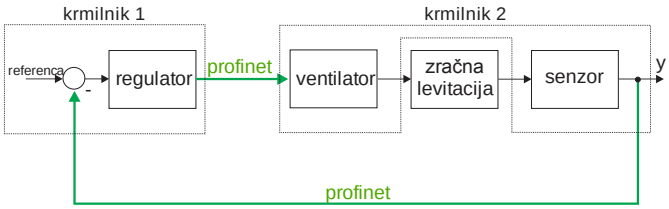
Kjer je:

- K_P - ojačenje,
- T_I - časovna konstanta integratorja,
- T_D - časovna konstanta diferenciatorja,
- a – zakasnitev diferencialnega dela,
- c – konstante diferencialnega dela ,
- x– dejanska vrednost,
- w – željena vrednost,
- b – konstanta proporcionalnega dela.

4 Komunikacijski protokol PROFINET in porazdeljeno vodenja

Komunikacijski protokol PROFINET uporablja TCP/IP standard za časovno nekritične operacije – za diagnostiko in upravljanje. Za časovno kritične operacije pa uporablja posebni RT- standard. To pomeni da se uporablja poseben PROFINET okvir, cikličen prenos podatkov, alarmov in nadzorne komunikacije. Prenosna hitrost Profinet-a znaša 100 Mbit/s. Uporablja full duplex komunikacijo in podpira tudi brezžično povezavo. Naslavljanje poteka preko IP in MAC naslova.

Prenos informacij poteka tako, da gospodar pošlje zahtevo, suženj jo obdela, če je zanj, pošlje podatke nazaj, če ni, pa pošlje podatke do naslednjega sužnja. Zato je PROFINET nekoliko počasnejši od EtherCAT-a. PROFINET uporablja posamezni okvir za naslavljanje posameznega sužnja in ima omejeno število vozlišč. Primeren je za komunikacijo med industrijskimi krmilniki, nadzornimi moduli (SCADA) in upravljalnimi moduli (MES). Enostavno ga združujemo z že obstoječimi vodili [8]. V sistemu regulacije zračne levitacije na sliki 5, PROFINET predstavlja kabel pod številko 5.



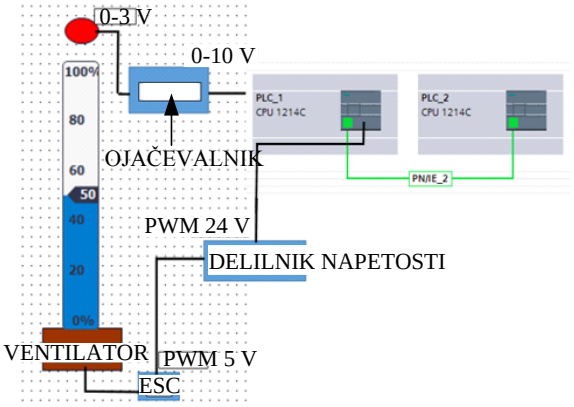
Slika 5: Zaprto-zančna struktura porazdeljenga vodenja.

5 Eksperimentalni rezultati

Fizično realiziran sistem je sestavljen iz:

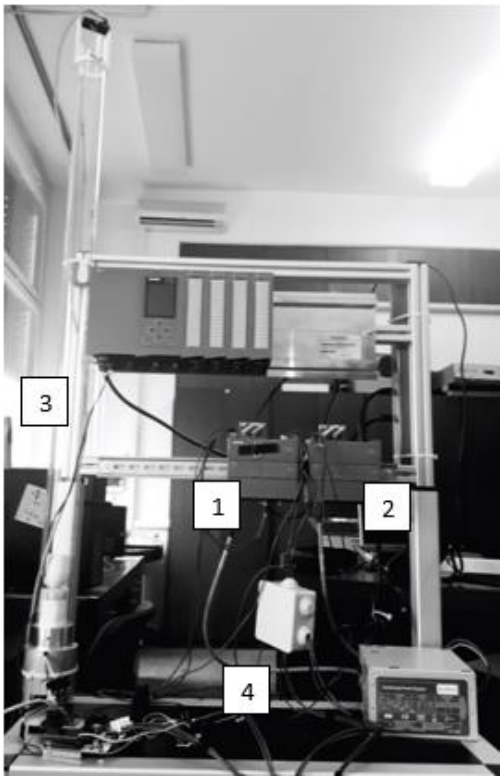
- vetrovnika s cevjo, višine 1 m,
- brez- krtačnega enosmernega motorja, ki je nameščen v aluminijasti podstavek nosilca cevi,
- dveh programirljivih krmilnikov SIMATIC S7-1200,
- plovca, natisnjenega s 3D tiskalnikom,
- senzorja višine tipa X-NUCLEO-53L0A1,
- RC- filtra, ki gladi izhodno napetost senzorja, saj le ta generira PWM- signal, na krmilniku, pa je uporabljen analogni vhod,
- operacijskega ojačevalnika s faktorjem ojačenja K=3, kateri je ojačal izhodni signal senzorja pred analognim vhodom v krmilnik,
- 12 V napajalnika,
- krmilnika motorja, tipa ESC ZTW 20 A BEC,
- napetostnega delilnika za znižanje izhodnega PWM- signala iz krmilnika iz 12 V na 5 V,
- IP- usmerjevalnika, kateri je uporabljen za komunikacijo PROFINET.

Vetrovnik je voden preko industrijskega vodila PROFINET. Na sliki 6 je prikazana shema izdelanega grafičnega vmesnika sistema.



Slika 6: Prikaz procesa zračne levitacije

Za vodenje na daljavo sta uporabljena dva SIEMENS-ova krmilnika SIMATIC S7-1200. Prvi je uporabljen kot aktuator, s pomočjo katerega prejemamo signal iz senzorja. Drugi krmilnik je uporabljen za regulacijo, ki regulirano veličino pošilja na prvi krmilnik, le ta pa vodi motor. Sistem realnega časa deluje s časom tipanja 20 ms. Končni izdelek je prikazan na sliki 7.



Slika 7: Sistem zračne levitacije

Slika 7. prikazuje naslednje elemente: aktuator – prvi krmilnik (št. 1), regulacija – drugi krmilnik (št. 2), vetrovnik (št. 3) in kabel za vzpostavitev komunikacije z industrijski protokolom PROFINET (št. 4)

6 Zaključek

V predstavljenem članku smo predstavili raziskavo in implementacijo porazdeljenega sistema zaprto-zančnega vodenja objekta zračne levitacije plovca v vetrovniku. Končni cilj projekta je bil izvesti regulacijo plovca v vetrovniku na realnem fizičnem modelu preko dveh porazdeljenih (medsebojno oddaljenih) programirljivih krmilnikov, povezanih s komunikacijo PROFINET. Matematični model levitacije z dodano regulacijo in odziv reguliranega procesa v programskem okolju MATLAB, ob upoštevanju dimenzij ter mase plovca, nekaterih vplivov iz okolja, kot sta viskoznost in gostota zraka smo tako preizkusili še v realnem okolju z industrijskima krmilnikoma. Na dno cevi vetrovnika, nad ventilator, smo namestili še paralelno zalepljene slamice, da smo v vetrovniku zagotovili laminarni tok zraka.

Iz prikazanih rezultatov je razvidno, da smo nalogo uspešno opravili, saj sistem deluje zanesljivo in učinkovito. Porazdeljeno vodenje preko protokola PROFINET se je izkazala, kot zanesljiva rešitev za časovno kritični sistem zračne levitacije z intervalom tipanja 20 ms.

Literatura

[1] Chen, X.; Song, G.; Zhang, Y. Virtual and Remote Laboratory Development: A Review. In Proceedings of the 12th International Conference on Engineering, Science, Construction, and Operations in Challenging Environments; American Society of Civil Engineers: Reston, VA, USA, 2010; pp. 3843–3852.

[2] Ma, J.; Nickerson, J.V. Hands-On, Simulated, and Remote Laboratories: A Comparative Literature Review. ACM Comput. Surv. 2006, 38, 7.

[3] Antsaklis, P.; Basar, T.; DeCarlo, R.; McClamroch, H.; Spong, M.; Yurkovich, S. NSF/CSS Workshop on New Directions in Control Engineering Education; Technical Report; University of Illinois at Urbana-Champaign: Champaign, IL, USA, 1998.

[4] Jesus Chacon, Jacobo Saenz, Luis de la Torre, Jose Manuel Diaz and Francisco Esquembre, Design of a Low-Cost Air Levitation System for Teaching Control Engineering, Sensors 2017, 17(10), 232.1

[5] Mejías, A.; Herrera, R.S.; Márquez, M.A.; Calderón, A.J.; González, I.; Andújar, J.M. Easy Handling of Sensors and Actuators over TCP/IP Networks by Open Source Hardware/Software. Sensors 2017, 17 (1), 94.

[6] Norman S. Nise, Control Systems Engineering, John Wiley & Sons, 2011

[7] Bernard Friedland, Control System Design, Dover Publications, 2005

[8] Komunikacijski protokol PROFINET. Dostopno na: https://us.profinet.com/technology/profinet/#prettyPhoto [15. 7. 2018]

Index

Aleksic Danijela: 78
Alič Mitja: 463
Aljaž Barbara: 388
Alp Ilgaz Mehmet: 71, 74
Amon Matej: 151
Andonovski Goran: 194
Antončič Mitja: 241, 245
Astorino Antonio: 71, 74
Babič Jan: 159
Babić Zdenka: 303
Babuder Klemen: 93
Baines García Álvaro: 143
Balent Jošt: 182
Banfi Luka: 51
Batagelj Boštjan: 63, 67, 71, 74
Begeš Gaber: 307
Beguš Samo: 423
Berkopec Ana: 116
Bertalanič Blaž: 467
Bevc Jakob: 25
Bezovšek Luka: 100
Bizjak Boris: 237
Bizjak Žiga: 155
Blažič Aljaž: 190
Blažič Boštjan: 241, 245
Bogovič Tomaž: 299
Bohak Ciril: 335, 348, 384
Borko Matej: 404
Bošnak Matevž: 190
Bovcon Borja: 396
Brandon RichardWebster: 367
Bratina Božidar: 143
Brezočnik Zmago: 13
Bujok Petr: 344
Celestina Metod: 313
Chowdhury Amor: 31
Cikač Jaka: 380
Čadež Matej: 273
Čegovnik Tomaž: 112, 291
Čehovin Zajc Luka: 372, 388
Čermelj Vid: 291
Damir Žarko: 3
Demirovic Nedzmija: 225, 229
Dežman Tomaž: 35
Dobrišek Simon: 368
Donaj Gregor: 404
Drnovšek Janko: 295
Drobnič Klemen: 269, 273, 277
Dubravic Amila: 127, 135
Emeršič Žiga: 400
Felzer Gregor: 31
Finc Matjaž: 5
Fišer Rastko: 269, 273, 277
Fraundorfer Friedrich: 380
Fürst Luka: 340, 445
Gams Andrej: 139, 167
Gašpar Jan: 249
Gašpar Timotej: 170
Gašparac Greta: 376
Geršak Gregor: 295, 423, 431
Glavič Boštjan: 307
Golob Mitja: 147
Gomez Iñaki: 186
Grm Klemen: 368
Gruden Timotej: 479
Gržinič Frelih Nina: 431
Guillaume Peter: 167
Guna Jože: 100
Gungl Ernest : 13
Gyergyek Tomaž: 186
Hace Aleš: 147
Hak Kim Byeong: 384
Hamler Anton: 233
Hasic Zerina: 127
Herman Leopold: 241, 245
Horvat Marina: 419
Humar Iztok: 39
Huremovic Alma: 225
Idzig Jure: 93
Igrec Dalibor: 31
Ilic Sinisa: 352
J. Martínez T. Guillermo: 74
Jakovljević Tamara: 43
Jakus Grega: 35, 108, 419
Jamšek Janez: 449
Jamšek Marko: 159
Javornik Tomaž: 43, 63
Kambič Tim: 155
Kandušer Maša: 415
Karakatič Sašo: 404
Karsten Rottwitt: 71
Kerbler Ana: 483
Kirn Nejc: 328
Kitanovski Andrej: 265
Klančar Gregor: 190
Klinar Katja: 265
Kodrič Domen: 249
Koplan Matej: 85
Kordež Jaka: 348
Kos Andrej: 47
Kos Tomaž: 356
Kosar Tomaž: 360
Kotnik Klemen: 190
Kovačič Jernej: 186
Kozmelj Martin: 120
Kramar Peter: 427
Krč Janez: 182
Kristan Matej: 372, 396
Križaj Janez: 392
Krstic Dragana: 78
Krupić Bilal: 408
Lavrič Henrik: 269, 273, 277

Lenasi Konrad: 233
Lončarević Zvezdan: 139
Lukežič Alan: 372
Magajne Aleš: 423
Makuc Danilo: 269
Markl Matic: 269
Marolt Matija: 335, 384
Marolt Miha: 39
Mavrič Martin: 257
Mavsar Matija: 9
Meden Blaž: 400
Meh Peer Jaša Vid: 253
Meh Peer Nataša: 253
Mernik Marjan: 360
Mihelj Matjaž: 151, 155
Miklavčič Damijan: 427
Miklavčič Peter: 63
Milenkovic Nadica: 352
Miljavec Damijan: 257
Mirčeta Kristijan: 384
Mlakar Boštjan: 483
Mlinar Tomi: 55, 59
Moravec Matej: 360
Munih Marko: 151, 155, 163
Musil Bojan: 419
Nemec Bojan: 139
Nemec Mitja: 201, 205, 217, 281
Nika Hrušovar: 190
Nikolic Petar: 78
Nikolovska Kristina: 190
Nogić Matej: 13
Nosan Simon: 265
Oblak Tim: 400
Ograjenšek Irena: 190
Ožbot Miha: 190
Pahič Rok: 139
Panič Damjan: 155
Papa Gregor: 43
Papić Marko: 120
Pavlin Gal: 459
Pavlin Mojca: 415
Pečnik Klemen: 85, 116
Peer Peter: 400
Perko Jernej: 131
Pernuš Martin: 190, 396
Perš Janez: 392, 396, 408
Petek Matjaž: 143
Petkovšek Marko: 217, 285
Petrič Tadej: 167
Pivec Maja: 4
Podobnik Gašper: 249
Podobnik Janez: 155, 163
Pogačnik Borut: 437
Pogačnik Matevž: 85, 89, 100, 104, 116
Polajžer Boštjan: 261
Polakova Radka: 344
Policardi-Antoncich Franc: 317, 328
Povšič Nejc: 89
Prah Blaž: 221
Pregelj Boris: 441
Prinčič Tadej: 307
Prislan Rok: 321
Puc Jernej: 163
Purg Jakob: 170
Pušnik Igor: 299
Radosavljevic Dragana: 352
Ratkoceri Jakup: 67
Ravnikar Robi: 155
Reberšek Matej: 427
Repas Jernej: 415
Ribič Vid: 104
Rihar Andraž: 217, 221, 281
Ritonja Jožef: 261
Rodič Blaž: 177
Rupnik Urban: 257
Sepesy Maučec Mirjam: 404
Simic Mitar: 303
Skalar Urban: 131
Skočaj Danijel: 380
Skuber Tadej: 17
Slemenšek Jan: 455
Sodnik Jaka: 112, 324, 419
Softic Izudin: 225, 229
Spalevic Petar: 352
Stanić Valentina: 431
Stefanovic Mihajlo: 78
Stojemnova Duh Emilija: 93, 47, 291
Stojmenova Kristina: 324
Strahija Roman: 21
Strahija Tatjana: 21
Stražar Anže: 194
Strban Sebastien: 335
Strmšnik Simon: 253
Stropnik Vid: 93
Sušin Denis: 213
Šafarič Riko: 143
Šehić Zenan: 135
Škrjanc Rebeka: 47
Šlajpah Sebastjan: 151, 155
Štruc Vitomir: 368, 392
Štrumbelj Erik: 376
Štukelj Marina: 299
Šuligoj Veronika: 449
Tavčar Marko: 131
Tement Sara: 419
Todorovic Jelena: 352
Tomažič Tine: 190
Tomc Matej: 423
Tomc Urban: 265
Topič Marko: 182
Topler Marcel: 261
Trbušić Mislav: 233
Trilar Jure: 291
Trost Andrej: 9, 17, 313
Ude Aleš: 139
Ujjwal Maulik: 6
Ulbl Domen: 471
Umberger Mark: 39

Uran Suzana: 143
Uroš Petkovič: 190
Verbič Matic: 190
Vernier P. Thomas: 427
Vodopivec Jan: 151
Vončina Danjel: 217, 221, 265
Vuckovic Dragan: 78
Vuk Baliž Kristjan: 475
Young Kim Min: 384
Zajec Peter: 209, 217, 285
Zajšek Uroš: 483
Zavratnik Veronika: 47
Zdešar Andrej: 131, 190
Zelinac Gordan: 303
Zimšek Danilo: 51
Zoretič Uroš: 108
Žemva Andrej: 17
Žerak Matjaž: 253
Žgank Andrej: 404
Žukovec Martin: 249