Zaznavanje gest na vgrajeni napravi s prvoosebnim pogledom

Blaž Rolih and Luka Čehovin Zajc

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-pošta: br9136@student.uni-lj.si, luka.cehovin@fri.uni-lj.si

Abstract

We are presenting a prototype of a wearable computing visual gesture recognition system that works on an embedded device. It uses a series of neural networks to detect a hand in first-person view, track its position and classify gestures into a number of categories. All this is done almost entirely on a dedicated embedded device, leaving resources of the main computer free for the actual application. As a part of our work we have implemented gesture control for a music player. Experimental evaluation shows that our system is able to detect a number of gestures reliably. It is able to learn to recognize a number of gestures using a fairly limited training dataset.

1 Uvod

Z razvojem računalniških sistemov ter njihovim prodorom v vse pore našega življenja so se razvijali tudi načini interakcije z njimi. Rokovanje z digitalnimi napravami, s katerimi se srečujemo tekom dneva, postaja vedno bolj heterogeno in prilagojeno namenu. Kljub temu, da se za zahtevno računalniško delo še vedno v glavnem uporabljata miška in tipkovnica, se s telefoni že desetletja upravlja z zasloni na dotik, popularni pa postajajo tudi govorni vmesniki ter ročne geste.

Zanimiva platforma za nove načine interakcije je nosljivo računalništvo (ang. wearable computing), ki obsega naprave, ki jih nosimo na telesu. Razni gumbi in stikala so lahko zasnovani intuitivno, vendar pa niso uporabni v vsakem kontekstu, marsikdaj bi lahko določene akcije sprožili s telesnimi gestami. Tak način rokovanja postaja zanimiv v kontekstu obogatene resničnosti (ang. augmented reality, AR), pri čemer bi lahko za zaznavanje uporabili tudi kamero, ki jo sistem uporablja za zajemanje okolja. V primeru nosljivih naprav pa mora biti zaznava gest prilagojena zaznavanju s senzorji na telesu, v primeru kamere to pomeni zaznavanje s prvoosebnim pogledom, kot je prikazano na Sliki 1.

V članku je predstavljen prototip za brezstično upravljanje z gestami roke na podlagi računalniškega vida v kontekstu prvoosebnega pogleda. Predstavljena študija problem obravnava v dveh medsebojno povezanih smereh. Zanima nas kakšen algoritem za razpoznavo lahko uporabimo, da bo razpoznava hitra in robustna tudi v prvoosebnem pogledu, obenem pa primerna za uporabo na vgrajenih senzorjih.

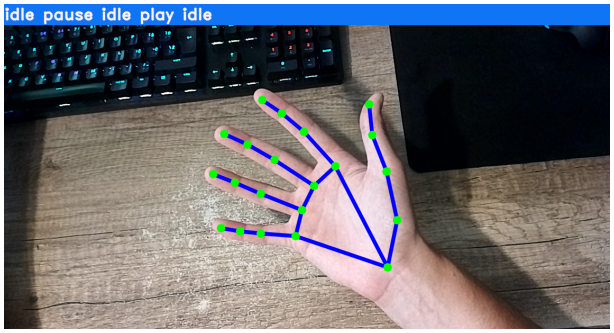
Slika 1: Ilustracija prototipa za zaznavanje gest z napravo OAK-D v prvoosebnem pogledu.

2 Sorodna dela

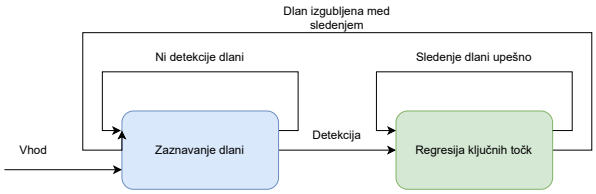
Prepoznavanje gest je kljub dolgi zgodovini [1] še vedno aktualno raziskovalno področje, razvoj algoritmov je povezan z razvojem strojne opreme, ki omogoča vedno bolj ambiciozne scenarije razpoznave. Najbolj znan sistem za razpoznavanje gest v zadnjem desetletju je verjetno še vedno Microsoft Kinect, z uporabo katerega je bilo vse od njegove prve verzije implementiranih več sistemov za prepoznavanje gest [2, 3] ter tudi bolj fine geste znakovnega jezika [4]. Kinect s pomočjo infrardeče svetlobe določa globinsko sliko, ki se lahko uporabi za določitev telesne poze uporabnika [5] ali njegove roke, zato pa potrebuje ustrezno nadzorovano okolje ter močan gostiteljski sistem za obdelavo podatkov od globinske slike naprej.

Zaradi omejitev globinskih kamer novejši pristopi preizkušajo tudi druge senzorje, raziskovalci so dosegli dobre rezultate z uporabo radarskih valov [6, 7]. Interakcija s pomočjo radarja je primerna tudi za upravljanje v avtomobilu [8]. Poleg tega se je razširila uporaba globokega učenja in nevronskih mrež, ki dobro obravnavajo razlike v izvajanju gest med ljudmi [9] z ustrezno zastavljeno učno množico. Z uporabo te metodologije so bili doseženi novi preboji pri razpoznavanju gest [10] pa tudi v kontekstu znakovnega jezika [11].

Glavni fokus obstoječih raziskav na temo razpoznavanja gest gredo torej v smer izboljšave natančnosti z



Slika 2: Zaznane ključne točke na roki v kontrolnem prikazu prototipa.



Slika 3: Delovanje algoritma za spremljanje položaja roke.

metodološkimi spremembami in izboljšavo in večanjem količine učnih podatkov. V primerjavi z njimi je naš cilj bolj sistemski, preveriti hočemo možnosti za učinkovito implementacijo lahkega in hitrega, obenem pa dostopnega senzorskega sistema za razpoznavo gest z novo generacijo vgrajenih naprav, v našem primeru gre za platformo DepthAI, ki jo razvija podjetje Luxonis[1]. Kljub kompaktnosti in nizki porabi izbrana naprava omogoča izvajanje kompleksnih globokih modelov v realnem času. Drugi sistemi, ki temeljijo na cenovno dostopnih napravah [12] so še vedno odvisni od močnega gostiteljskega računalnika, zato bi lahka in prenosna platforma lahko podprla študije na področju uporabniških vmesnikov z uporabniki izven laboratorijskega okolja.

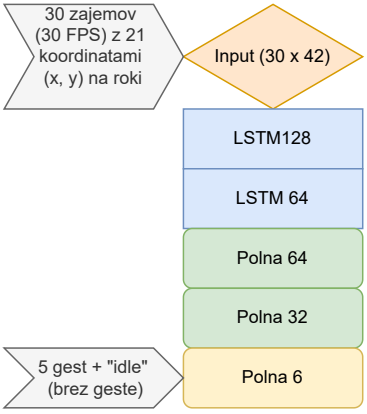
3 Metodologija

Kot smo že omenili, je naš sistem zasnovan pragmatično, združuje pred-naučen globoki model za določitev poze roke ter preprosto mrežo za razpoznavanje gest. Taka kombinacija se je že izkazala za robustno in učinkovito v drugih raziskavah [7]. Na ta način pridobimo fleksibilen in robusten sistem, saj lahko oba modula zamenjamo ali doučimo neodvisno enega od drugega.

Prvi del algoritma temelji na dveh konvolucijskih nevronskih mrežah ter vmesnih operacijah, kot je to opisano v [13] in prikazano v diagramu na Sliki 3. Za zaznavanje dlani se uporablja hitri detektor rok, ki temelji na arhitekturi SSD [14]. To omogoča dovolj hitro zaznavanje rok pri različnih velikostih. Druga mreža pa opravi regresijo 21 ključnih točk roke, kot je to prikazano na Sliki 2. Za pohitritev se detekcija ne izvaja v vsakem koraku, po začetni zaznavi roke se za določitev pozicije opravi sledenje na podlagi regresije ključnih točk [13].

Drugi del algoritma temelji na procesiranju zaznanega stanja roke v večih zaporednih časovnih korakih čemur

[1]https://docs.luxonis.com/en/latest/



Slika 4: Večnivojska LSTM arhitektura, ki jo uporabljamo za zaznavo gest.

algoritem pripiše ustrezno gesto. Za to nalogo smo definirali preprosto večnivojsko arhitekturo s kombinacijo plasti LSTM [15] in polno povezanih plasti, prikazana na Sliki 4. Mreža prejme zaporedje pozicij posameznih točk roke zajetih v časovnem oknu 30 zajemov (kar pri normalnem delovanju ustreza eni sekundi).

4 Sistem

Sistem za zaznavanje gest je bil implementiran na vgrajeni napravi OAK-D[2]. Platforma je namenjena realnočasni uporabi metod računalniškega vida in strojnega učenja na zmogljivih in učinkovitih vgrajenih sistemih. Jedro naprave je namenski čip za procesiranje slikovnih podatkov Movidius MyriadX podjetja Intel. Sama naprava je v veliki meri tudi odprtokodna in je dostopna velikemu krogu uporabnikov.

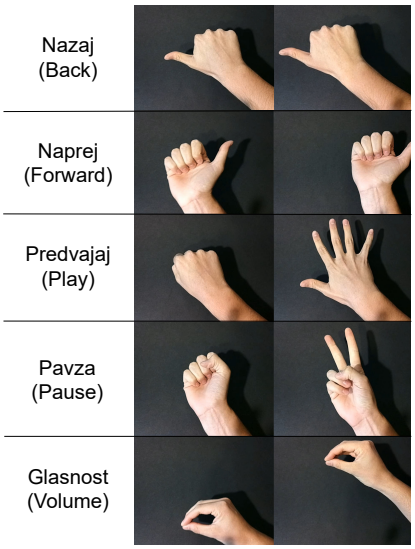
V naših scenarijih uporabe predpostavljamo, da bi bil sistem, podoben platformi DepthAI vgrajen v očala za obogateno resničnost. To smo simulirali tako, da je imel uporabnik kamero nameščeno na glavo s pomočjo naglavnega traku, kot je prikazano na Sliki 1. OAK-D podpira sočasno izvajanje več nevronskih mrež ter ostalih operacij, postopek obdelave se opiše s cevovodom, ki se izvaja neposredno na napravi, brez vmesnega nadzora gostiteljskega sistema.

5 Evalvacija

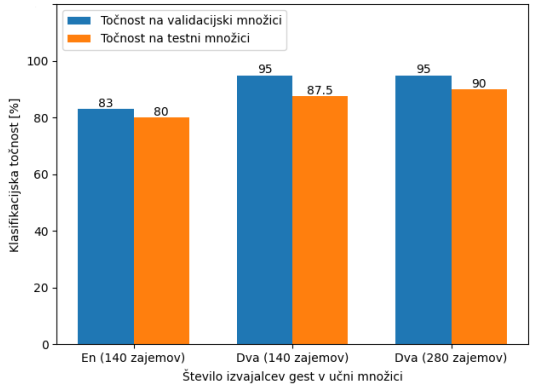
Prototip za zaznavanje gest smo preizkusili na scenariju upravljanja predvajalnika glasbe. Za upravljanje smo določili pet gest, ki so ilustrirane na Sliki 5. Zaradi zasnove sistema, predvsem zaradi robustne predstavitve rok, sistem že z manjšo učno množico doseže visoko točnost. Uporabljamo pred-naučena modela za detekcijo dlani in regresijo ključnih točk roke [13]. Modela sta naučena na veliki količini različnih podatkov, kar zagotavlja robustno zaznavanje rok.

Model za razpoznavanje gest v času smo naučili na podatkih, ki smo jih zajeli sami. Za eksperiment v obsegu tega članka smo zajeli množico posnetkov gest, ki so jih izvajale tri osebe. Vsaka oseba je za vsako od

[2]https://store.opencv.ai/products/oak-d



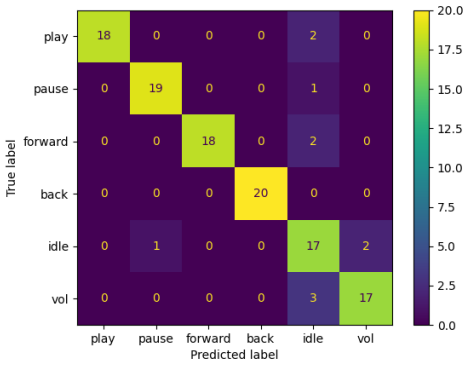
Slika 5: Prikaz petih gest v sistemu, geste so ilustrirane z začetno in končno pozicijo roke.



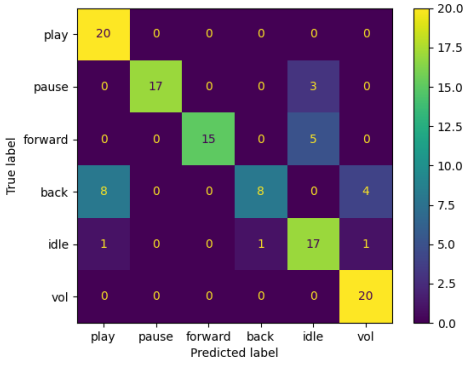
Slika 6: Primerjava obsega učne množice, uporaba gest ene osebe napram uporabi gest dveh oseb v enakem in dvojnem obsegu.

petih gest izvedla 20 ponovitev. Poleg tega smo za ustrezno obravnavo zaznave gest v obliki klasifikacije zajeli tudi 40 primerov gibanja brez namenske geste, kar označimo z razredom *idle*. Skupno je bilo zajetih 400 kratkih posnetkov, kar je sicer malo za globoko metodo, a ker učimo samo zadnji del algoritma z manj parametri, že to zadostuje za dobre rezultate. Model za razpoznavo gest smo učili na posnetkih ene ali dveh oseb, testiranje pa izvedli na posnetkih tretje. Same učne podatke smo v nadaljevanju razdelili v razmerju $70/30$ še na del za validacijo. Pri učenju smo uporabili algoritem Adam s stopnjo učenja 0.001 in velikost paketa 32. Kot funkcijo izgube smo uporabili križno entropijo. Zaradi preprostosti modela je učenje že na CPE trajalo le nekaj minut, kar je uporabno za hitro inkrementalno izboljšavo in učenje novih gest.

Na diagramu 6 vidimo točnost na testni množici, ki v primeru modela, naučenega na gestah dveh oseb dosega 90%. Model torej dokaj dobro posplošuje tudi na še ne videnih podatkih. Ob pregledu matrike zamenjav na Sliki 7 pa opazimo, da model dobro zaznava posamezne geste, ter jih uspešno loči, težave pa ima pri razlikovanju dejanske geste in odsotnosti le-te, (razred *idle*). V primeru uporabe primerov zgolj ene osebe pa natančnost pade, iz matrike zamenjav, prikazane na Sliki 8 pa vidimo, da do tega pride tudi na račun zamenjav med gestami. Po pričakovanjih model torej bolje posplošuje, če geste v učni množici izvaja več kot ena oseba, tudi, če gre za enako število učnih primerov. Predpostavljamo, da bi lahko z dodatnimi primeri večih izvajalcev gest robustnost še izboljšali, prav tako pa ocenjujemo, da bi se lahko sistem z gestami enega izvajalca v uporabi prilagodil na specifike njegovih gest.

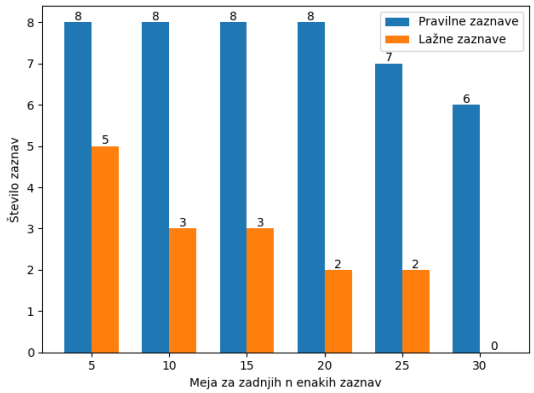


Slika 7: Matrika zamenjav modela naučenega na gestah dveh izvajalcev.



Slika 8: Matrika zamenjav modela naučenega na gestah enega izvajalca.

V praksi je pri zaznavi gest pomembna točnost klasifikacije, obenem pa želimo zmanjšati verjetnost lažne zaznave. V toku slik je obravnavanih veliko časovnih oken, pri tem pa se hitro zgodi, da pride do šumnega primera, ki povzroči napako. V praksi lahko take zaznave preprečimo z robustnim pristopom. V vsaki sekundi namreč dobimo 30 zaznav, gesta pa je potrjena le, če je N zadnjih zaznav napovedalo isto kategorijo. Za testiranje zanesljivosti daljše uporabe sistema je zajeta sekvenca osmih gest, med katerimi nastopa tudi naključno premikanje rok, ki ne predstavlja dejanskih gest s pomenom. Na diagramu na Sliki 9 vidimo vpliv števila potrebnih zaporednih zaznav geste. Vidimo, da s povečevanjem velikosti okna pada število lažnih zaznav, vendar pa od meje 20 pada tudi število pravilnih zaznav, verjetno tudi



Slika 9: Primerjava rezultatov za mejo, ki določa koliko istih zaznav pomeni dejansko zaznavo geste.

zaradi sorazmerne kratkosti obravnavanih gest. Potrebnih 20 zaznav tudi ne vpliva opazno na zamik zaznav. Za dodatno izboljšanje bi lahko model inkrementalno doučili z množico primerov premika roke, kjer prihaja do lažnih zaznav (ang. hard positive/negative mining).

Prototip smo ovrednotili tudi s testnimi uporabniki. Uporabnikom so bile predstavljene geste, sledila je prosta interakcija ter kratek intervju. Sistem se je uporabnikom zdel uporaben ter večina odzivov je bila pozitivnih. Same geste so bile vsem uporabnikom smiselne, saj so bile zasnovane tako, da so semantično povezane s samo akcijo. Uporabniki gest ne bi spremenili, bi pa nekateri združili gesti za predvajanje in ustavljanje glasbe, ki sta dejansko negacija ena druge. Tudi v sklopu teh testov so glavno težavo povzročile lažne zaznave, ki so se pojavile pri potezah, ki spominjajo na dejanske geste.

6 Zaključek

V članku smo predstavili prototip sistema za zaznavanje gest na podlagi kamere z globokimi nevronskimi mrežami. Posebnost sistema je, da deluje v kontekstu nosljivega računalništva in torej obdeluje slike, zajete v prvoosebnem pogledu, hkrati pa deluje skoraj v celoti na dostopni vgrajeni napravi in je kot tak lahko razširljiv in nadgradljiv. V okviru študije smo preverili tudi nekatere lastnosti sistema na scenariju brezstičnega upravljanja z glasbenim predvajalnikom in ugotovili, da se kljub majhnemu številu učnih primerov model dobro nauči posploševanja gest, z dodajanjem primerov večih uporabnikov pa se robustnost še poveča. V nadaljnjem delu imamo namen izvesti širšo študijo z več udeleženci, obenem pa sistem razširiti na več aplikacij in na ta način preveriti meje sistema glede razlikovanja gest ter robustnost na variacije izvedbe le-teh.

Zahvala: Raziskava je bila delno financirana v okviru ARRS programa P2-0214.

References

[1] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, pages 379–385, 1992.

[2] Yi Li. Hand gesture recognition using kinect. In *2012 IEEE International Conference on Computer Science and Automation Engineering*, pages 196–199, 2012.

[3] Yanmei Chen, Bing Luo, Yen-Lun Chen, Guoyuan Liang, and Xinyu Wu. A real-time dynamic hand gesture recognition system using kinect sensor. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2026–2030, 2015.

[4] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American sign language recognition with the kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 279–286, New York, NY, USA, 2011. Association for Computing Machinery.

[5] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013.

[6] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. page 851–860. Association for Computing Machinery, 2016.

[7] Nanziba Basnin, Lutfun Nahar, and Mohammad Hossain. An integrated cnn-lstm model for micro hand gesture recognition. pages 379–392, 02 2021.

[8] Karly A. Smith, Clément Csech, David Murdoch, and George Shaker. Gesture recognition using mm-wave sensor for human-car interface. *IEEE Sensors Letters*, 2(2):1–4, 2018.

[9] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, 2007.

[10] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *ICCV Workshops*, pages 3120–3128, 2017.

[11] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3413–3423, June 2021.

[12] Cloe Huesser, Simon Schubiger, and Arzu Coltekin. *Gesture Interaction in Virtual Reality: A Low-Cost Machine Learning System and a Qualitative Assessment of Effectiveness of Selected Gestures vs. Gaze and Controller Interaction*, pages 151–160. 08 2021.

[13] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *CoRR*, abs/2006.10214, 2020.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.