Zbornik radova Fakulteta tehničkih nauka, Novi Sad

**UDK: 004.9**
**DOI: https://doi.org/10.24867/03SA01Markovic**

# DETEKCIJA I PREPOZNAVANJE STATIČKIH I DINAMIČKIH GESTOVA RUKE

# DETECTION AND RECOGNITION OF STATIC AND DYNAMIC HAND GESTURES

Zorana Marković, *Fakultet tehničkih nauka, Novi Sad*

**Oblast – RAČUNARSKA GRAFIKA**

**Kratak sadržaj –** *U ovom radu opisan je postupak prepoznavanja gestova ruke na bazi RGBD slika i videa, gde je ideja bila da se istraže mogućnosti detekcije položaja ruke i šake, kao i odredjenih gestova vezanih za njih. Metode koje su razvijene i analizirane u ovom radu su bazirane na mašinskom učenju, ali su pored njih analizirane i neke metode koje se ne baziraju na mašinskom učenju, radi poređenja rezultata. Prvo je opisana metoda za prepoznavanje statističkih gestova šake, korišćenjem postojećih jednostavnih metoda iz OpenCV biblioteke. Potom je razvijena metoda detekcije položaja statičkih gestova šake korišćenjem konvolucionih mreža. Na kraju, kao glavni cilj rada, je implementirana metoda detekcije dinamickih gestova šake i ruke [1], korišćenjem mašinskog učenja na bazi 3D konolucionih mreža. Ta metoda je dalje adaptirana radi optimizacije rezultata na selektovanim podacima. U radu su dati eksperimentalni rezultati, gde su analizirane performanse implementiranih algoritama.*

**Ključne reči:** *Prepoznavanje gestova ruke, mašinsko učenje, konvolucione neuronske mreže*

**Abstract** – *In this paper, the process of recognizing hand gestures based on RGB images and videos is described, where the idea was to explore possibilities of detection arm and hand position and hand gestures. Methods that were developed and analyzed in this paper are based on machine learning, but also along them, some methods that aren't based on machine learning were also analyzed, to compere results. First, the method of static hand gestures recognition using already existing OpenCV library methods were described. Next, the method for detecting static hand gestures using convolutional neural networks was developed. Finally, as the main focus of the paper, the method of detecting dynamic hand gestures was implemented [1], using machine learning based on 3D convolutional neural networks. This method was further adapted to optimize the results of a selected data set. Likewise, experimental results are given, where the performances of the implemented algorithm was analyzed.*

**Keywords:** *Handgesture recognition, machine learning, convolutional neural networks*

**NAPOMENA:**
**Ovaj rad proistekao je iz master rada čiji mentor je bio dr Vladimir Zlokolica, doc.**

## 1. UVOD

U poslednjih nekoliko godina, detekcija gestova ruke je postala važan segment koji ima široku primenu u svim oblicima implementacije savremenih tehnologija.

Detekcija pokreta postaje imperativ dizajna kompjuterskog interfejsa medicinskih uređaja, mašinskih i električnih sistema [1,2], a primenjiva je i na polju industrije video igrica [3] i znakovnog jezika [4]. Pomoću ovakvog interfejsa korisnik ne mora da fizički pritisne dugme ili ima interakciju sa uređajima.

U automobilskoj industriji, najveća beneficija jeste povećanje sigurnosti tokom vožnje [5]. Poslednjih godina došlo je do velikog interesovanja stručnjaka i nastanka mnogih radova koji su vezani upavo za automobilsku industriju, a koji su doprineli razvoju tehnologije koja omogućava detekciju pokreta.

Pojedini radovi [1,2,6] koriste mašinsko učenje, konkretno veštačke neuronske mreže, za rešavanje tehnoloških problema.

Suština ovog sistema jeste postojanje dovoljno velike baze podataka, koja poseduje mnogo varijacija, i koja omogućuje uspostavljanje generalizacije nad gestovima i korisnicima [1]. Takođe faktor osvetljenja utiče na tačnost detektovanja pokreta [2]. Iz toga sledi da se tačnost detekcije pokreta može menjati u zavisnosti od doba dana i osvetljenja kabine automobila. Postoje i problemi vezani za tehnologiju i kvalitet samih ulaznih informacije. Ako su ulazne infromacije loše (sadrže dosta artefakata), mogućnost detekcije odgovarajućeg gesta se znatno može smanjiti.

Za rešavanje ovih problema pojedini radovi koriste kombinaciju RGBD za povećavanje tačnosti detektovanja gestova kao i za generisanje više podataka [1][2]. Za rešavanje navedenih problema, mnogi su koristili različite tipove optimizacija i funkcija grešaka [7] u polju mašinskog učenja (konkretno neuronskih mreža) da bi postigli bolje rezultate. Detekcija gestova se takođe može postići korišćenjem metoda koje nisu bazirane na mašinskom učenju [8].

U Poglavlju 2. će se opisati proces prepoznavanja statičkih gestova ruke bez mašinskog učenja.

U Poglavlju 3. će se opisati proces prepoznavanja statičkih gestova sa mašinskim učenjem.

U Poglavlju 4. će se opisati pustupak prepoznavanja dinamičkih gestova ruke.

Na kraju, u Poglavlju 5. dati su eksperimentalni rezultati primene navedenih metoda.

## 2. PREPOZNAVANJE STATIČKIH GESTOVA RUKE BEZ MAŠINSKOG UČENJA

Za izradu koda, korišćena je *C++* i *OpenCV* biblioteka.

Set podataka koji je korišćen dobijen je pomoću mikrosoft uređaja Kinekt v1. Za detekciju gestova korišćeni su samo podaci kanala koji sadrže vrednosti dubine, rezolucije 640x480. Naime, u ranim fazama ekperimentisanja se uočilo da RGB (HSV) vrednosti daju loše rezultate prilikom pokušaja uklanjanja pozadine. Gestovi koji su namenjeni za prepoznavanje u ovom delu projekta, su gestovi prikazivanja brojeva od 1. do 5.

Da bi se detektovao gest, prvo je potrebno odrediti najveći region na frejmu. Da bi se to postiglo, ekstraktovani su pikseli na frejmu koji imaju vrednost od 245. do 255. (odnosno oni koji su najbliži uređaju) tako što je svim ostalim pikselima, koji nemaju vrednosti u ovom rasponu, dodeljena nova vrednost - 0. Na osnovu preostalih podataka, detektovana je najveća površina na slici. To je urađeno tako što su kreirane konture oko preostalih oblika na sceni pomoću funkcije *findContours*, a zatim je pozvana funkcija *contourArea* koja definiše površinu konture. Ako je površina konture veća od već definisane najveće porvršine, onda ta najveća površina postaje površina konture.

Zatim su kreirani okvir i nedostaci (eng. *convexityDefects*) za određenu površinu kontura. Nedostatak predstavlja oblast koja ne pripada objektu (konturi), ali se nalazi unutar okvira. Potom je iscrtan rezultat, konkretno kontura i primitiva kao što su krug i linija. Ovi primitivi služe za prikazivanje segmenata nedostataka Da bi se detektovao gest, kreirana je *switch* petlja koja uzima broj iscrtanih krugova i na osnovu toga prikazuje koliko selekcija postoji.

## 3. PREPOZNAVANJE STATIČKIH GESTOVA RUKE SA MAŠINSKIM UČENJEM

Za izradu koda koristio se *Python* zajedno sa *Keras* bibliotekom.

### 3.1. Set podataka

Set podataka koji je korišćen za detekciju statičkih gestova je LaRED (eng. *Large RGB-D Extensible Hand Gesture Dataset*) set podataka [9]. On se sastoji iz ukupno 243,000 RGBD fajlova, zajedno sa binarnim fajlovima maski. Za ovaj rad su korišćeni oni gestovi koji pokazuju brojeve od 1 do 5. Set podataka je podeljen na tri foldera: set za treniranje, set za validaciju i set za testiranje. U svakom od ta 3 foldera se nalaze još po 5 foldera koje predstavljaju klase od 1 do 5. Prilikom treniranja modela, kreirane su tri podgrupe podataka: mala (18,000 slika), srednja (45,000 slika) i velika (90,000 slika).

### 3.2. Preprocesovanje

Na osnovu koda za otvaranje i prikazivanje jednog fajla [9], kreiran je kod koji preuzima sve binarne fajlove kanala koji sadrži vrednost dubine, normalizuje ih i čuva u formatu *PNG*. Svaki binarni fajl se učitava kao 16-bitni niz veličine 320x240. Niz a je zatim prosleđen u funkciju (1) :

$$a[a > 10000] = 0 \qquad (1)$$

tj. da sve vrednosti piksela, koje su veće od deset hiljada, dobiju vrednost 0. Zatim je rezultat normalizovan.

Kreirane su dve podele podataka. Prvi set podataka je podeljen 80%-10%-10%. Drugi fajlovi su raspoređeni tako da samo jedan subjekat čini set za validaciju, a ostali podaci pripadaju setu za treniranje.

### 3.3. Klasifikacija

Na osnovu podele seta podataka, kreirana su dva modela. Za kreiranje oba modela korišćena je konvoluciona neuronska mreža.

Arhitektura prvog modela se sastoji iz četiri 2D konvoluciona sloja praćena aktivacionim slojevima. Za atkivacionu funkciju izabran je *ReLU*. Posle svakog aktivacionog sloja sledi *max pooling*. Navedena arhitektura, do ovog trenutka, je identična za treniranje svih podela podataka (mala, srednja, velika podgrupa). Između trećeg i četvrtog konvolucionog sloja je postavljen *dropout* sloj, vrednosti 0,5. Posle poslednjeg konvolucionog sloja sledi još jedan *dropout* sloj vrednosti 0.2. Rezultat *dropout* sloja je prosleđen u *flatten* sloj. Nakon toga, rezultat je prosleđen u dva skrivena sloja. Tip aktivacione funkcije izlaznog sloja je *softmax*. Za razliku od arhitekture male/srednje podgrupe, valika podgrupa nema *dropout* slojeve. Umesto njih postavljena je normalizacija skupova pre prvog aktivacionog sloja i posle drugog i trećeg aktivacionog sloja.

Arhitektura drugog modela se bazira na primeni NVIDIA arhitekture. Ona se sastoji iz dve podmreže : HRN i LRN. HRN mreža se sastoji iz četiri 2D konvoluciona sloja, praćena aktivacionim slojevima. Za aktivacionu funkciju je ponovo izabran *ReLU*. Sledi *max pooling* posle svakog aktivacionog sloja. Kao i za prvi model, rezultat poslednjeg *max pooling* sloja se prosleđuje u *flatten* sloj. HRN mreža sadrži dva skrivena sloja, između kojih se nalazi *dropout* sloj. Takođe, između poslednjeg skrivenog sloja i izlaznog sloja se nalazi još jedan *dropout*. Izlazni sloj koristi aktivacionu funkciju *softmax*, i ima dimenziju pet. LRN mreža se sastoji iz tri 2D konvoluciona sloja praćena aktivacionim i *max pooling* slojevima. Posle *max pooling* sloja, arhitektura za HRN i LRN je ista. Rezultati izlaznih slojeva HRN i LRN su pomnoženi, odnosno pomnožene su verovatnoće članova klase. Rezultat je prosleđen u konačan izlazni sloj, koji sadrži aktivacionu funkciju *softmax* i ima dimenziju 5.

### 3.4. Treniranje

Za optimizaciju modela korišćen je SGD. SGD su prosleđene vrednosti za brzinu učenja 0,008 i za momentum 0.9. Augmentacija je dobijena pozivanjem funkcije *ImageDataGenerator* i prosleđivanjem vrednosti za *rescale* = 1./255. Tip funkcije greške koje su se koristile je unakrsna entropija i srednja kvadratna

vrednost (isti model se trenirao dva puta sa različitim funkcijama greške).
Za treniranje drugog modela korišćene su iste funkcije. Jedina razlika je ta da je kreirana posebna funkcija koja poziva navedene i rezultat povezuje sa *yield*. Ovo je urađeno zbog spajanja HRN i LRN mreže. Vrednost brzine učenja se menjala u rasponu 0.001 i 0.0001, gde je izabrana poslednja vrednost treniranja 0.0008. Veličina grupe ima vrednost 3, a broj epoha je 100. Treba napomenuti da je za testiranje korišćena samo velika podgrupa. To je urađeno zbog načina testiranja modela (odvajanje jedne grupe/osobe).

## 4. PREPOZNAVANJE DINAMIČKIH GESTOVA RUKE

Za izradu koda koristio se *Python* zajedno sa *Keras* bibliotekom.

### 4.1. Set podataka

Set podataka koji je korišćen za detekciju dinamičkih gestova predstavlja isti set podataka koji je napravljen za VIVA izazov. Set podataka se sastoji iz 1,460 fajlova sa vrednostima kanala dubine i 1,461 fajlova sa vrednostima kanala iluminance koje sadrže prikaz 32 različita gesta ruke. Te gestove su prikazale osam različitih osoba. Format videa je *AVI*. Za projekat je korišćen samo deo seta podataka.
Navedeni deo seta podataka je raspoređen na pet klasa gestova (Gest1, Gest3, Gest6, Gest8 i Gest9) koji su prikazale devet različitih osoba (Osoba 1, 2, 3, 4, 5, 8, 9, 10 i 12). Razlog zbog kojeg je odabran mali broj gestova je nedostatak resursa. Konkretno, prilikom korišćenja svih podataka, došlo je do nedostatka memorije računara. Sa druge strane, razlog zašto nisu izabrani svi subjekti koji su pokazivali gest je taj da set podataka koji je korišćen nije imao isti broj videa za sve subjekte.
Pojedini subjekti nisu posedovali određeni gest koji je postojao kod ostalih.

### 4.2. Preprocesovanje

Pre nego što se set podataka mogao trenirati, morao se augmentovati tako da svaki video sadrži 32 frejma. To je urađeno zbog toga što su svi videi u setu podataka različite dužine. Sve je postignuto pomoću NNI.
Na frejmove, koji sadrže vrednosti kanala iluminance, primenjen je Sobel operator da bi se izdvojile ivice videa i činile sam proces treniranja bržim. Takođe, za grupu podataka koja se odnosi na RGBD, frejmovi kanala sa vrednostima dubine i iluminance su isprepletani. Ukupan broj frejmova za RGBD videe su 64. Svi videi su formata *AVI*.

### 4.3. Klasifikacija

Kreirani su modeli za dva seta podataka, RGB/D i RGBD. Arhitektura oba modela je ista, izuzev ulaznih fajlova koji su različitih dimenzija. Arhitektura se sastoji iz HRN i LRN mreže [1]. HRN se sastoji iz četiri 3D konvoluciona sloja, praćena *aktivacion* slojevima i četiri *maxpooling* sloja. Arhitektura LRN se sastoji iz tri 3D konvoluciona sloja, praćena aktivacionim slojevima i *max pooling* slojevima.
Tip aktivacione funkcije koji je korišćen jeste *ReLU*. Obe arhitekture sadrže *flatten* sloj, dva skrivena sloja između kojih se nalazi *dropout* sloj, vrednosti 0.5, i konačan skriveni sloj, odnosno izlazni sloj.
Izlazni sloj je dimenzije 5 i sadrži *softmax* funkciju aktivacije.Tip funkcije greške koje su se koristile je unakrsna entropija i srednja kvadratna vrednost, radi eksperimenta (isti model se trenirao dva puta sa različitim funkcijama greške).
Rezultati izlaznih slojeva HRN i LRN su pomnoženi. Proizvod je prosleđen u konačan izlazni sloj koji sadrži *softmax* funkciju aktivacije. Dimenzija izlaznog sloja je 5.

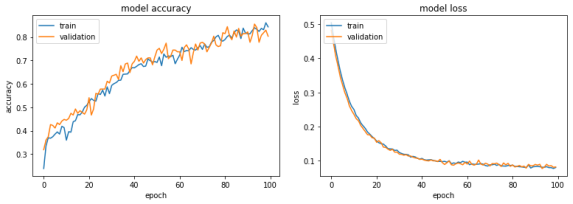
### 4.4. Treniranje

Za optimizaciju je korišćen SGD uz pristustvo momentuma (konkretno uključujući Nesterov ubrzani gradijent). Vrednosti brzine učenja je 0.0008, a vrednosti momentuma je 0.9. Vrednost za veličinu skupa je 6. Da bi se dobio najbolji rezultat, menjale su se vrednosti brzine učenja u rasponu od 0.001 do 0.0001.

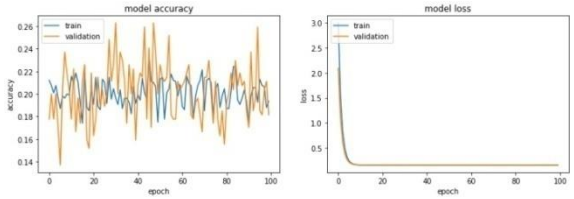
Razlog zbog kojega nisu indetični je nedostatak adekvatnog resursa, konrektno: zbog slabosti grafičke kartice. Takođe, zbog istog razloga, nisu mogle da se iskoriste sve vrste augmentacije podataka. Grafička kartica koja je korišćena za ovaj projekat je NVIDIA GeForce GTX 660.

## 5. REZULTATI

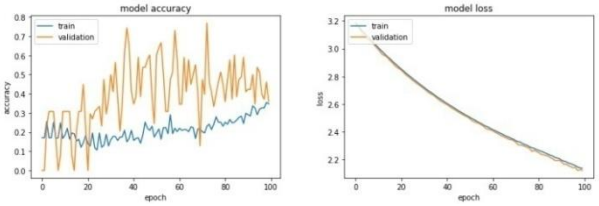
Na sledećim slikama prikazani su najbolji rezultati. Prve dve slike se odnosne na statičke gestove, a druge tri na dinamičke gestove.



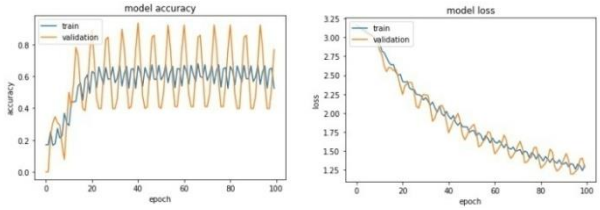

Slika 1. *Rezultati treniranja i testiranja za set podataka (80%-10%-10%) sa velikom podgrupom podataka, koristeći MSE funkciju greške.*



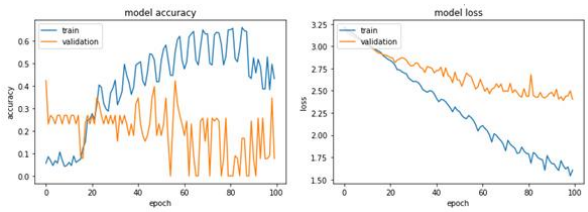

Slika 2. *Slika 4.11 Rezultati treniranja i testiranja za set podataka sa velikom podgrupom podataka (Podela seta podataka : jedan subjekat testiranje, preostali subjekti treniranje), koristeći MSE funkciju greške (100 epoha).*




Slika 3.*Rezultati treniranja i testiranja za set podataka sa kanalom iluminance (sa gradijentom), koristeći unakrsnu entropiju (100 epoha).*




*Slika 4. Rezultati treniranja i testiranja za set podataka sa kanalom dubine, koristeći unakrsnu entropiju (100 epoha).*




*Slika 5.Rezultati treniranja i testiranja za set podataka sa kombinovanim kanalima, koristeći unakrsnu entropiju (100 epoha).*

## 6. ZAKLJUČAK

U radu je opisan proces detekcije statičkih i dinamičkih gestova ruke korišćenjem metoda sa i bez mašinskog učenja. Za detekciju statičkih gestova primenjena je 2D konvoluciona neuronska mreža. Za detekciju dinamičkih gestova primenjena je 3D konvoluciona neuronska mreža i metoda detektovanja gestova na osnovu detektovanja nedostataka. Prikazane su metode primene različitih tipova podataka, njihovo kombinovanje, primenjivanjem HRN i LRN mreže i niz opcija pomoću kojih se mogu postići bolji rezultati. Na osnovu eksperimenata, primećeno je da kombinacija kanala sa vrednostima dubine i iluminance (zajedno sa unakrsnom entropijom kao funkcijom greške) daje najbolje rezultate u odnosu na kanal sa vrednošću dubine i na kanal sa vrednostima ilumiance (sa i bez primena Sobel operatora). Treba napomenuti da se validacija uradila na jednoj osobi (grupi), a treniranje na preostalim osobama (grupama). Na osnovu daljeg testiranja, primećeno je da je dobijen bolji rezultat kada je set podataka podeljen u odnosu 80%-20%, čak i 90%-10% u odnosu na predhodnu metodu. Takođe, tokom testiranja statičkih gestova, primećeno je da su dobijeni bolji rezultati primenom MSE funkcije greške u odnosu na unakrsnu entropiju. Problem koji se može uočiti kod rezultata je taj da su pojedine krive uniformne, odnosno imaju šemu, što umanjuje kvalitet detekcije gesta. Takođe, primećeno je da kada se obavi trening dva puta na istom setu podataka, bez menjanja parametara, dobijaju se dva drastično različita rezultata.

Glavni problem tokom rada su bili nedovoljni resursi, koji su uslovljavali primenu male baza podataka. Takođe navedene metode su relativno nove. Postoji mnoštvo debata o primeni različitih metoda i funkcija, i još uvek ne postoji fiksirana šema koja funkcioniše za bilo koji tip podataka neuronske mreže.

## 7. LITERATURA

[1]P. Molchanov, S. Gupta, K. Kim and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks",*IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1-7., 2015.

[2] E. Ohn-Bar and M. M. Trivedi, "Hand Gesture Recognition in Real Time for Automotive Interfaces: A Multimodal Vision-Based Approach and Evaluations",*IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 6, 2014.

[3] Y. Zhu and B. Yuan, "Real-time hand gesture recognition with Kinect for playing racing video games",*International Joint Conference on Neural Networks (IJCNN)*, pp. 3240-3246., 2014.

[4] T. Starner, J. Weaver and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video",*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371-1375, 1998.

[5] S. Loehmann, M. Knobel, M. Lamara, A. Butz, "Culturally Independent Gestures for in Car Interactionsm",*Kotzé P. et al. (eds) Human-Computer Interaction – INTERACT 2013. Lecture Notes in Computer Science*, vol 8119., 2013.

[6]P. Molchanov, S. Gupta, K. Kim and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition",*11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1-8.,2015.

[7]J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai, T. Chen,"Recent Advances in Convolutional Neural", arXiv:1512.07108, 2017.

[8] S. U. Rahman, Z. Afroze, M. Tareq, "Hand Gesture Recognition Techniques For Human Computer Interaction Using OpenCv", International Journal of Scientific and Research Publications, vol. 4, no. 12, 2014.

[9]Y. Hsiao, J. Sanchez-Riera, T. Lim, K. Hua, W. Cheng, "LaRED: A Large RGB-D Extensible Hand Gesture Dataset", *The 2014 ACM Multimedia Systems Conference*, 2014.

**Kratka biografija:**

**Zorana Marković je** rođena u Novom Sadu 1994.god. Diplomski rad na Fakultetu tehničkih nauka iz oblasti Računarske grafike – Animacija u inženjerstvu odbranila je 2017.god.