Predlog diplomskog rada

1. **Definicija problema/cilja projekta.**

Detekcija i praćenje šake na videu, iscrtavanje landmark-ova i prepoznavanje gesta (konkretno: thumbs up, thumbs down, fist, stop, rock, smile, live long, peace, call me, ok)

1. **Motivacija problema rešavanog u projektu.**

Jedna od mogućih primjena detekcije gesta bi bila u medicini, konkretno u fizikalnoj terapiji pacijenta, oporavak od operacije nekog od ligamenta na šaci je dosta složen proces i zahtijeva svakodnevne vježbe standardnih pokreta šakom kako bi se povratila potpuna mobilnost, pacijentu bi dosta pomoglo ukoliko bi mogao da uz pomoć računara vidi da li vježbe izvodi pravilno. Sa druge strane, praćenje položaja šake se može koristiti u kombinaciji sa drugim VR/AR tehnologijama, kako bi se korisniku pružilo imerzivnije iskustvo, bez držanja kontrolnih palica u rukama.

1. **Relevantna literatura (minimum 3 rada)**

**1.** Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L. and Grundmann, M., 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.

Ovo je konkretan rad koji predstavlja MediaPipe pipeline model za detekciju šake i određivanje landmark-ova na istoj. Pipeline se sastoji iz 2 modela: Palm-Detector model (BlazePalm) i hand landmark model. BlazePalm je SSD model koji kao ulaz prima sliku na kojoj se nalazi šaka i kao izlaz vraća bounding-box iste. Hand landmark model je regresioni model koji predviđa X, Y i Z koordinate za 21 landmark šake a kao ulaz prima blisko isječenu sliku šake (isječeni i skalirani dio slike oivičen bounding-box-om iz prethodnog modela). Skup podataka je bio ručno kreiran i sadržao je slike šaka sa anotiranim landmark-ovima, rađena je podjela na trening/test skup. Model je evaluiran koristeci AP i postigli su rezultat od ~95%. Zanimljivo je što ovaj pipeline može bez problema raditi u realnom vremenu i na slabijem hardveru (kao što je mobilni telefon).

**2.** Agoes, I.M.H.A.S., 2021. Applying Hand Gesture Recognition for User Guide Application Using MediaPipe.

Zadatak ovog rada je bilo prepoznavanje gesta šake sa slike korišćenjem MediaPipe framework-a. Sama detekcija landmark-ova rađena je uz pomoc MediaPipe pipeline-a dok je prepoznavanje gesta rađeno preko unaprijed definisanih pravila gdje se na osnovu međusobne udaljenosti landmark-ova kao i ugla koji obrazuju linije skeletnog modela (koji se dobija međusobnim povezivanjem landmark-ova) predviđao položaj dlana i prstiju, samim tim i gest šake. Skup podataka je bio ručno kreiran i sadržao je gestove kojima bi ljudi mogli upravljati kinect platformom (open menu, close menu, menu1, ... , menu 8, machine controll, speed controll i power controll) i rađena je podjela na trening/test skup. Rezultati prepoznavanja su evaluirani upotrebom preciznosti, odziva i F-mjere. Dobijene F-mjere za svaki od gestova pojedinačno nisu bile manje od 91% dok je ukupna F-mjera za klasifikaciju bila 95%.

**3.** Oudah, M., Al-Naji, A. and Chahl, J., 2020. Hand gesture recognition based on computer vision: a review of techniques. *journal of Imaging*, *6*(8), p.73.

U ovom radu, rađen je pregled raznih tehnika za prepoznavanje gestikulacije šake, konkretno: skin-color-based recognition, appearance-based recognition, motion-based recognition, skeleton-based recognition, depth-based recognition, 3D-model-based recognition i deep-learning pristup. Autori nisu vršili sami eksperimente već su samo radili pregled svih metoda, te su svi rezultati preuzeti iz drugih radova. Za mjeru performanse se koristila tačnost (accuracy). Prikazane su dobre i loše strane svakog od ovih algoritama, na primjer: deep-learning pristup radi jako dobro ali mu je potreban ogroman skup podataka za obučavanje dok skin-color-based pristup ne zahtjeva nikakvo obučavanje, radi jako brzo ali ima problem prilikom detekcije u kompleksnoj/loše osvijetljenoj sceni...

1. **Skup podataka.**

S obzirom da mi je potreban relativno mali skup podataka (samo za testiranje modela), koristiće se ručno anotiran skup podataka koji će se sastojati od slika sa Google Images pretraživača kombinovano sa ručno kreiranim fotografijama sa telefona.

1. **Metodologija**.

Ulaz u sistem predstavljaće real-time video sa web kamere. Za detekciju i lokalizaciju šake kao i određivanja koordinata landmark-ova na istoj koristiće se već implementiran Google MediaPipe Hands pipeline iz MediaPipe framework-a. Za klasifikaciju gesta šake koristiće se potpuno povezana neuronska mreža u MediaPipe Hand Gesture arhitekturi (5 potpuno povezanih slojeva sa 64, 128, 512, 64, 32 neurona respektivno). Kao ulaz, primaće X i Y koordinate svakog od 21 landmark-a dok će izlaz predstavljati jedna od 10 klasa gestova (konkretno: thumbs up, thumbs down, fist, stop, rock, smile, live long, peace, call me, ok). Mreža za klasifikaciju će biti obučavana korišćenjem transfer learning pristupa gdje će se potrebne težine preuzeti sa MediaPipe sajta.

1. **Metod evaluacije.**

S obzirom da je MediaPipe Hands pipeline već implementiran a za neuronsku mrežu planiram da iskoristim prethodno obučene težine, kompletan gore pomenuti test skup će se koristiti za evaluaciju modela.

Kao mjeru performanse za klasifikaciju gesta koristiće se accuracy, precision, recall i F-mjera.

Kao mjera performanse za rad u realnim uslovima koristiće se odnos prosječnog FPS-a (koliko frejmova po sekundi model može da obradi) i F-mjere.

1. **Softver**

Za implementaciju MediaPipe Hands pipeline-a koristiće se MediaPipe framework. Više o framework-u možete pogledati na njihovom [sajtu](https://google.github.io/mediapipe/). Što se tiče implementacije neuronske mreže koristiće se Keras framework. Za rad sa real-time videom sa web kamere koristiće se OpenCV biblioteka.