Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

David Dukić

Računalno prepoznavanje izraza lica u stvarnom vremenu metodama dubokog učenja s primjenom na snimke eksperimenta u razrednoj okolini

Zagreb, 2021.

Ovaj rad izrađen je na Zavodu za elektroničke sustave i obradbu informacija pod vodstvom doc. dr. sc. Ane Sović Kržić i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2020./2021.

Sadržaj

1 Uvod

Prepoznavanje izraza lica (engl. *facial expression recognition, FER*), skraćeno PIL, odnosno prepoznavanje emocija, moguće je promatrati kao klasifikacijski zadatak gdje se od sudionika eksperimenta traži da svrstaju izraze lica drugih ljudi u jednu od N predefiniranih klasa emocija. Problem prepoznavanja emocija pokazao se zahtjevnim, čak i za ljude. Ljudi često pokazuju više emocija istovremeno. Na primjer, osoba u strahu će nerijetko biti istovremeno i tužna i iznenađena. Iz tog razloga moguće je vrlo lako pogriješiti pri previđanju emocionalnog stanja drugih ljudi koristeći samo vizualne informacije. Stoga, ljudi ne sude o tuđim emocijama samo na temelju izraza lica, već koriste i govor tijela, audio signale kao što je glas etc.

S razvojem algoritama strojnog učenja (engl. *machine learning*) i odgovarajućeg hardvera, računalom podržano PIL (engl. *computer aided FER*) postalo je moguće. Budući da je PIL i ljudima težak zadatak, još je izazovnije automatizirati ga te omogućiti predviđanje emocija sa slika u stvarnom vremenu (engl. *real-time*).

Većina trenutno dostupnih rješenja za računalno PIL u domeni računalnog vida (engl. *computer vision*) koristi neznatno modificirani osnovni model emocija koji je predstavio američki psiholog Paul Ekman. Ekmanova teorija razmatra $N = 6$ osnovnih emocija: ljutnju, strah, tugu, užitak (engl. *enjoyment*), gađenje i iznenađenje [1]. Moderni algoritmi strojnog učenja obično vrše klasifikaciju na temelju skupova podataka (engl. *data set*) koji su labelirani neznatno modificiranom verzijom Ekmanovog modela emocija. Jedina razlika leži u tome što je emocija *užitak* zamijenjena emocijom *sreća*, unatoč tome što je užitak općenitija emocija od sreće. Uz to, neki skupovi podataka sadrže i emociju *neutralno* kao jednu od postojećih klasa koja ne ukazuje na pojavu bilo kakve emocije.

U ovom radu upotrijebljen je sličan model emocija, kompatibilan sa zadanim istraživačkim ciljem. Emocije koje su korištene kao klase kroz istraživanje bile su ljutnja, tuga, sreća, iznenađenje, neutralno i ostalo. Klasa *ostalo* predstavlja strah, gađenje te sve ostale emocije koje nisu obuhvaćene osnovnim modelom emocija kao što je recimo prijezir (engl. *contempt*). Cilj istraživanja bio je razviti modele za računalno PIL koji mogu predviđati emocije iz videozapisa u stvarnom vremenu ili još brže. Štoviše, zamisao je bila primijeniti razvijene prediktivne modele na izraze lica sudionika takozvanog "razrednog eksperimenta", dobiti što preciznije predikcije emocija te statistički analizirati prikupljene podatke. S obzirom na činjenicu da modeli PIL još nisu bili razvijeni u trenutku kada je održan eksperiment, izraze lica bilo je potrebno zabilježiti kamerom. Razredni eksperiment uključivao je rješavanje jednostavnih zadataka vizualnog programiranja te pokretanje programskog koda na robotu Lego Mindstorms

[1]. Snimanje videozapisa izvedeno je kamerom postavljenom ispred računala. U svrhu ovog rada, eksperiment je nazvan **razredni eksperiment prepoznavanja izraza lica**. Nadalje, sljedeći istraživački zadatak uključivao je analizu dobivenih predikcija emocija te potragu za odgovorom *je li moguće povezati emocije s aktivnostima koje su sudionici izvodili u trenutku pokazivanja emocije?*. Naravno, ideja je bila dobiti što točnije predikcije emocija, a ne pojednostaviti zadatak razvojem modela koji detektiraju samo promjene emocija. U tu svrhu, značajan napor u ovom radu bio je usmjeren na razvoj odgovarajućih modela PIL. Kako bi se pronašao prediktivni model koji najbolje odgovara uvjetima razrednog eksperimenta, izvršen je opsežan pregled područja te postojećih rješenja što je opisano u odjeljku 2.

Heuristika (engl. *heuristic*) za odabir klase *other* bila je da sudionici eksperimenta, s velikom vjerojatnošću, neće pokazivati osjećaje straha i gađenja u učionici prilikom rješavanja programskih zadataka te rada s Lego robotom. Time je zadatak razvoja modela olakšan i pojednostavljen. O navedenoj odluci te ostalim odlukama povezanim sa skupom podataka raspravlja se u odjeljku 3. U istom odjeljku predstavljen je i opis isprobanih modela PIL te su izneseni argumenti o tome koji se razvijeni prediktivni model najbolje uklapa u predviđanje nad snimkama razrednog eksperimenta. Konačno, taj isti odjeljak opisuje okolnosti u kojima je izveden razredni eksperiment.

Odjeljak 4 prikazuje rezultate evaluacije razvijenih modela PIL. Detaljan postupak evaluacije služio je kako bi se odredilo koji je od prediktivnih modela najprikladniji za primjenu u razrednom eksperimentu. Također, opisan je i postupak odabira modela, odnosno optimizacija hiperparametara korištenih algoritama strojnog učenja. Nadalje, u istom odjeljku, predstavljeni su rezultati i zaključci iz eksplorativne analize podataka (engl. *exploratory data analysis*) nad predviđenim emocijama te takozvanim *tipovima aktivnosti*. Definirane su četiri glavne aktivnosti koje su sudionici eksperimenta obavljali prilikom rješavanja zadataka: *programiranje*, *robot*, *pomoć* i *ostalo*. Konačno, rad završava odjeljkom 5 koji iznosi sveukupni pregled provedenih eksperimenata te prijedlog potencijalnih ideja za proširenje provedenog istraživanja u budućnosti.

[1] `https://www.lego.com/en-gb/themes/mindstorms`

Glavni doprinosi rada su sljedeći:

1. Iscrpno istraživanje postojećih rješenja u području strojnog učenja za PIL te sličnih razrednih eksperimenata gdje su se pratile emocije sudionika kroz vrijeme
2. Razvoj četiri različita modela PIL te detaljna statistička usporedba njihovog uspjeha na standardnim skupovima podataka
3. Eksplorativna analiza nad predviđenim emocijama u razrednom eksperimentu s uvidom u povezanosti i korelacije između emocija, zadataka, spola te tipova aktivnosti.

2 Pregled područja

Pregled područja je analiziran kroz tri različite perspektive: pregled skupova podataka, pregled modela te pregled sličnih eksperimenata.

2.1 Pregled skupova podataka

Tri najčešće korištena skupa podataka za PIL su "The Extended Cohn-Kanade" skup podataka (CK +) [2], "Facial Expression Recognition 2013" skup podataka (FER-2013) [3] te "Static Facial Expressions in the Wild" skup podataka (SFEW) [4, 5]. U nastavku su kratko opisani radovi koji su postigli *state-of-the-art* rezultate nad ta tri skupa podataka. Inforamcije o najboljim rezultatima preuzete su sa stranice *Papers with code*[2].

Neuronske mreže (engl. *neural networks*) pod imenom "Frame Attention Networks" postigle su *state-of-the-art* rezultate na CK+ skupu podataka [6]. U ovom istraživanju usvojen je dio baze CK+ u svrhu stvaranja skupa podataka prikladnog za primjenu na razredni eksperiment.

Skup podataka FER-2013 sastavljen je i predstavljen za natjecanje organizirano od strane Goodfellow et al. [3]. U svom radu autori iznose najbolje postignute rezultate na FER-2013 skupu podataka, dok se model koji je postigao najveću točnost na testnom skupu još uvijek smatra *state-of-the-art* na FER-2013 skupu podataka. Autori također tvrde da bi skup podataka mogao imati pogreške u označavanju. Stoga, u ovom radu nisu korištene sve dostupne slike iz tog skupa podataka. Preuzet je podskup slika za koji je detaljnim pregledom utvrđeno da ne sadrži pogreške u označavanju.

State-of-the-art rezultati na SFEW skupu podataka postignuti su pomoću modela pod imenom "Novel Region Attention Network" autora Wang et al. [7]. SFEW skup

[2] `https://paperswithcode.com/`

podataka sastoji se od slika iz stvarnih situacija s okluzijama i varijacijama poza lica s označenim emocijama. Kako bi se poboljšala uspješnost predviđanja modela za PIL "u divljini" (engl. *FER in the wild*), što odgovara **razrednom eksperimentu**, dodane su i slike iz SFEW skupa za učenje u konačni skup podatka. Više objašnjenja o konačnom skupu podataka izneseno je u odjeljku 3.

2.2 Pregled modela

Modeli koji se koriste za rješavanje zadatka PIL mogu se podijeliti u tri kategorije ovisno o vrsti primijenjenih algoritama strojnog učenja: tradicionalni modeli, modeli dubokog učenja te hibridi stvoreni kombinacijom oba pristupa. Što se tiče modela emocija, oni se mogu podijeliti na dvije najčešće varijante: kategoričku (engl. *categorical*) i dimenzionalnu (engl. *dimensional*). Kategorički modeli emocija koriste fiksan broj emocija i gledaju na zadatak PIL iz klasifikacijske perspektive što znači da je cilj modela predvidjeti diskretnu emociju. Dimenzionalni modeli emocija predviđaju vrijednost u dvodimenzionalnom prostoru definiranom mjerama valentnosti (engl. *valence*) i uzbuđenja (engl. *arousal*). S obzirom na to da se kod dimenzionalnih modela predviđa kontinuirana brojčana vrijednost, takve modele može se smatrati regresijskim.

2.2.1 Kategorički modeli

Wang et al. [8] usmjerili su svoje istraživanje prema rješavanju zadatka prepoznavanja suptilnih izraza lica. Autori su predložili mehanizam zasnovan na pažnji (engl. *attention-based mechanism*) s naglaskom na glavna lokalna područja izraza lica te njihove korelacije. Eksperimenti su vršeni nad skupovima podataka pod nazivom "Large-scale Subtle Emotions and Mental States in the Wild" (LSEMSW) te CK+.

Autori rada [9] razvili su neuronsku mrežu pod nazivom "Oriented Attention Pseudo-siamese Network" (OAENet) za rješavanje zadatka PIL te tvrde da njihov model OAENet koristi i lokalne i globalne informacije s lica. Pozornost lokalnih regija poboljšava se korištenjem težinske maske koja kombinira karakteristične točke na licu (engl. *facial landmarks*) i koeficijente korelacije. Razvijene mreže testirane su nad sljedećim skupovima podataka: "Real-world Affective Faces Database" (RAF-DB), AffectNet, CK+ te MMI.

Rješavanje PIL "u divljini" nužno je za predviđanje emocija u stvarnim životnim situacijama. Viswanatha Reddy et al. [10] riješili su taj zadatak kombinirajući karakteristične točke lica i mrežu XceptionNet na AffectNet skupu podataka.

Li et al. [11] predstavili su vjerojatnosni model baziran na dinamičnoj Bayesovskoj mreži (engl. *dynamic Bayesian network*) koja koristi tri razine aktivnosti na licu.

Autori su ekperimentirali na CK+ te MMI skupovima podataka.

Garcia et al. [12] pojednostavili su zadatak PIL klasifikacijom izraza lica u samo dvije klase: sretnu i tužnu. Zadatak je riješen koristeći modele dubokog učenja. Također, stvorili su i skup podataka sa slikama ljudi iz različitih okruženja, izraza lica, poza, dobi, etničke pripadnosti etc. Konačno, prikazali su implementaciju aplikacije za prepoznavanje izraza lica u stvarnom vremenu.

Neki su se autori usredotočili na izdvajanje vremenskih i prostornih obilježja za modele koje su razvili. Autori rada [13] razvili su mrežu "Part-based Hierarchical Bidirectional Recurrent Neural Network" (PHRNN) koju su iskombinirali s mrežom pod nazivom "Multi-signal Convolutional Neural Network" (MSCNN) u duboku evolucijsku prostorno-vremensku mrežu. Prva je korištena za lociranje vremenskih obilježja na temelju karakterističnih točaka lica iz uzastopnih slika. Druga je korištena za izdvajanje prostornih obilježja iz mirnih slika. Opisana metoda postigla je *state-of-the-art* rezultate na CK+, Oulu-CASIA te MMI skupovima podataka. Zhang et al. [14] su koristili kombinaciju dubokog i tradicionalnog učenja za rješavanje problema PIL u video sekvencama. Svoju metodu nazivaju hibridnim modelom dubokog učenja. Prvo su iskorištene dvije konvolucijske neuronske mreže (engl. *convolutional neural network, CNN*) za dobivanje prostornih i vremenskih značajki. Zatim su te značajke integrirane u duboku mrežu vjerovanja (engl. *deep belief network*). Na kraju, izvedeno je prosječno sažimanje (engl. *average pooling*) kako bi se omogućilo konačno klasificiranje pomoću linearnog stroja potpornih vektora (engl. *support vector machine, SVM*). Eksperimenti su provedeni nad skupovima podataka BAUM-1, RML i MMI. Autori tvrde da su nadmašili *state-of-the-art* rezultate na tim skupovima podataka.

Postoje i potpuno tradicionalni pristupi rješavanju problema PIL. Primjer vrijedan spominjanja jest rad autora Yeasin et al. [15] koji su koristili skrivene Markovljeve modele (engl. *hidden Markov models, HMM*) za izdvajanje temeljnog modela za svaki izraz lica te su evaluirali svoje rezultate na CK skupu podataka.

Perveen et al. [16] predložili su dinamičku reprezentaciju zasnovanu na jezgri u videozapisima za rješavanje zadatka PIL s modelom Gaussove mješavine (engl. *Gaussian mixture model, GMM*). Koristili su tri različite jezgre: jezgru temeljenu na mapiranju, vjerojatnosti te podudaranju. Rezultati su evaluirani na skupovima podataka MMI, "Acted Facial Expressions in the Wild" (AFEW) te BP4D.

Konačno, Kabakus [17] je predstavio jedinstvenu metodu za rješavanje zadatka PIL s novom arhitekturom konvolucijske neuronske mreže koju naziva PyFER. Mreža je evaluirana na standardnom CK+ skupu podataka i pokazala je visoke rezultate.

2.2.2 Dimenzionalni modeli

Kim et al. [18] proveli su zanimljivo istraživanje o prepoznavanju emocija. Autori nisu koristili samo lice osobe kako bi predvidjeli emocije u prostoru valentnosti i uzbuđenja (engl. *valence-arousal space*). Njihovo istraživanje otkrilo je da pozadina slike može poslužiti kao dobra značajka za predviđanje konkretne emocije prikazane na slici. Model koji su koristili bio je duboka neuronska mreža. Lee et al. [19] stvorili su multimodalne povratne mreže pažnje (engl. *multi-modal recurrent attention networks*) za predviđanje emocija na temelju boje slike, dubine slike te termalnih videozapisa koji se koriste kao multimodalni ulaz. Autori tvrde da njihova metoda može postići *state-of-the-art* rezultate u dimenzionalnom PIL na skupovima podataka RECOLA, SEWA te AFEW.

2.3 Pregled sličnih eksperimenata i obrazaca uporabe

Autori sličnih radova u ovom pododjeljku proveli su srodne eksperimente prateći emocije ispitanika u učionici i izvlačeći zaključke o zabilježenim emocijama. Štoviše, neki od njih razvili su modele PIL u svrhu njihove upotrebe za poboljšanje obrazovnog sustava.

Yang et al. [20] predložili su metodu PIL s ciljem razumijevanja učenja u virtualnom okruženju. Koristili su skup podataka "Japanese Female Facial Expression" (JAFFE) koji se sastoji samo od lica japanskih žena. Izvađene značajke predane su neuronskoj mreži koja je bila naučena da predviđa jednu od šest predefiniranih emocija. Autori tvrde da bi njihova metoda bila korisna za uporabu u obrazovanju na daljinu.

Još jedan istraživački eksperiment vrijedan spomena proveli su Seng et al. [21] gdje su autori previđali emocije kako bi procijenili zadovoljstvo kupaca iz videozapisa. Ovo istraživanje nije bilo orijentirano na primjenu u okruženju kakvo se može naći u učionici, ali ima sličnu ideju bilježenja emocija te njihove obrade radi izvlačenja zaključaka. Autori su koristili i audio i video podatke kako bi odredili ocjene zadovoljstva kupaca na temelju predviđenih emocija.

Neki autori su stvorili vlastite skupove podataka i na njima naučili modele u svrhu predviđanja emocija u obrazovnom sustavu. Jedan primjer dolazi iz rada [22] gdje je svrha istraživanja bila omogućavanje PIL u *online* učenju. Autori su u svom kategoričkom modelu koristili pet emocija. Različite konvolucijske neuronske mreže bile su naučene, uspoređene te temeljito evaluirane koristeći novostvoreni skup podataka.

Slično, Li et al. [23] razvili su samokodirajuću cikličku mrežu dubokog učenja čija je namjena bila rješavanje zadatka PIL u učionici. Stoga su autori prikupljali slike s javnih video tečajeva te snimki predavanja u učionicama kako bi omogućili učenje

modela. Zadatak je bio pojednostavljen te su modeli naučeni samo za predviđanje tri klase na sakupljenom skupu podataka: negativnu, neutralnu i pozitivnu. Njihov model koristio je ideje konvolucijskog samokodiranja i cikličke konzistencije.

Tonguç et al. [24] proveli su istraživanje najsličnije **razrednom eksperimentu prepoznavanja izraza lica** iz ovog rada. Tijekom predavanja, autori su pratili promjene u emocijama učenika. Emocije su praćene pomoću "Microsoft Emotion Recognition" aplikacijskog programskog sučelja (engl. *application programming interface, API*). Predavanje se sastojalo od tri dijela: uvod, aktivnosti i zatvaranje. Zabilježene su promjene u emocijama i provedena je statistička analiza kako bi se utvrdilo jesu li promjene značajne s obzirom na spol, sat predavanja etc. Njihov eksperiment pokazao je da su se povećale emocije prijezira, ljutnje i straha, dok su se smanjile emocije sreće, tuge i gađenja u uvodnom dijelu predavanja. Tijekom zatvaranja predavanja sreća se povećala, dok su se sve ostale emocije smanjile. Dio predavanja u kojem su studenti obavljali aktivnosti bio je manje pokriven njihovim radom i za taj dio samo izvještavaju kako se tuga malo povećala, ali onda se sreća povećala kako su se smanjivale sve druge negativne emocije što je bilo uzrokovano aktivnostima koje je instruktor dodijelio učenicima. Podaci su prikupljani svakih 10 sekundi s kamera postavljenih ispred učenika na njihovim računalima.

Razredni ekperiment koji je proveden u sklopu ovog rada sličan je dijelu "aktivnosti" prethodno opisane studije [24] jer su praćene emocije dok su učenici rješavali programske zadatke. Izrazi lica bilježeni su pomoću kamera tableta postavljenih ispred računala. Nakon toga, na snimljene video podatke primijenjeni su razvijeni modeli PIL za predviđanje emocija svake sekunde. Razvijeni modeli učeni su i na primjerima lica "u divljini" čime je postignuta relativna robusnost kod predviđanja emocija na izrazito nagnutim licima. Ideja da se audio značajke koriste za unaprjeđenje predviđanja razvijenih modela bila je vrlo interesantna, ali neizvediva u slučaju ovog istraživanja jer su studenti tijekom postupka rješavanja programskih zadataka bili tihi većinu vremena.

3 Materijali i metode

Kao što je prezentirano u prethodnom odjeljku, većina istraživanja u području računalnog PIL provedena je pomoću algoritama dubokog učenja koristeći kategoričke modele emocija. Ovaj rad razmatra i tradicionalne i duboke algoritme učenja u kombinaciji s kategoričkim modelom emocija.

3.1 Skup podataka

Korištene su slike iz tri standardna skupa podataka za prepoznavanje emocija: CK+, FER-2013 i SFEW. Brojčane informacije o skupu podataka dane su tablicom 1.

CK+ skup podataka sadrži skup slika za svakog ispitanika i svaku od sljedećih emocija: ljutnja, gađenje, strah, sreća, tuga, iznenađenje i prijezir. Za specifičnu emociju, slike su snimane postupno, počevši od neutralnog izraza lica pa sve do ciljane emocije koju je sudionik eksperimenta trebao pokazati. Stoga su u ovom radu korištene samo prva i zadnja slika od svakog ispitanika te svake emocije. Prva slika prikupljena je samo jednom za svakog ispitanika kako bi se izbjegla neuravnoteženost klasa (engl. *class imbalance*) zbog najvećeg broja slika neutralne klase. Klasa *prijezir* nije korištena.

Iz skupa FER-2013 ručno su odabrane slike za svaki razred koje su bile označene ispravnom oznakom. U konačnom skupu podataka, pripremljenom za ovaj rad, najviše primjera su upravo slike iz FER-2013 skupa podataka. Slike iz FER-2013 skupa podataka labelirane su sljedećim emocijama: ljutnja, gađenje, strah, sreća, tuga, iznenađenje te neutralno.

CK+ i FER-2013 su skupovi podataka koji sadrže slike snimljene u uglavnom statičnom okruženju, a ne "u divljini". Budući da je prepoznavanje emocija u dinamičnom okruženju vjerojatnije u stvarnosti, konačni skup podataka dopunjen je slikama iz SFEW skupa podataka koji je jedan od najpopularnijih skupova podataka sa slikama "u divljini". Točnije, dodane su sve slike iz skupa za učenje koji je bio označen istim oznakama emocija kao i skup podataka FER-2013.

Kako bi se skup podataka prilagodio očekivanom okruženju PIL za razredni eksperiment, razred *ostalo* stvoren je spajanjem svih slika označenih kao strah ili gađenje u jednu klasu, pod pretpostavkom da učenici neće pokazivati osjećaje gađenja ili straha tijekom programiranja i rada s robotom. Nadalje, važno je napomenuti da su odbačene slike iz skupova podataka s dječjim licima zbog činjenice da su svi sudionici razrednog eksperimenta bile odrasle osobe.

U konačnici, skup podataka brojao je 3929 slika. Neke su slike iz skupa podataka bile već unaprijed izrezane kako bi se prikazalo samo lice ispitanika (uglavnom

Tablica 1: Sažetak skupa podataka.

Klasa	Početno (3929 ukupno)			Konačno (3950 ukupno)		
	CK+	FER-2013	SFEW	Učenje	Validacija	Testiranje
Ljutnja	57	414	178	417	104	131
Sreća	84	417	198	453	114	142
Neutralno	105	395	150	418	104	130
Ostalo	126	374	164	426	106	133
Tuga	70	430	172	429	108	135
Iznenađenje	87	412	96	385	96	119
Ukupno	529	2442	958	2528	632	790

za skup podataka FER-2013), dok su druge zahtijevale ručno otkrivanje i izdvajanje lica. Također, SFEW skup podataka ponekad je imao više lica na jednoj slici koja su pokazivala istu emociju. Stoga su takve slike razdvojene u više slika gdje je na svakoj slici izdvojeno zasebno lice s početne slike. Za detekciju lica korišten je model po nazivom "Single Shot MultiBox Detector" (SSD) [25]. Ako je model imao pouzdanost veću od 50%, pretpostavka je bila da je detektirani objekt lice. Posljedično, broj konačnih slika u skupu podataka (3950) veći je od broja početnih slika skupa podataka (3929). Konačno, slike su modificirane tako da odgovaraju dimenzijama $300{\times}300$ piksela, pretvorene su u sive tonove (engl. *grayscale*) te je primijenjena metoda pod nazivom "Contrast Limited Adaptive Histogram Equalization" (CLAHE).

Podjela skupa podataka na skup za učenje, validaciju i testiranje napravljena je stratificirano kako bi se izbjegao problem disbalansa klasa. Nadalje, podjela je napravljena s istim slučajnim stanjem (engl. *random state*) određenim unaprijed kako bi se omogućila usporedivost razvijenih modela. Ideja je bila trenirati sve modele na istom skupu podataka za učenje, podesiti hiperparametre na istom skupu za validaciju te usporediti uspjeh predviđanja na istim neviđenim podacima iz testnog skupa.

3.2 Modeli strojnog učenja

Implementirane su i testirane metode tradicionalnog i dubokog učenja u programskom jeziku Python koristeći specijalizirane biblioteke strojnog učenja kao što su scikit-learn[3] i Pytorch[4]. Eksperimenti s modelima dubokog učenja rađeni su uz pomoć platforme Google Colab[5].

[3] `https://scikit-learn.org/stable/`
[4] `https://pytorch.org/`
[5] `https://colab.research.google.com/`

3.2.1 Tradicionalan pristup

Početna ideja za rješavanje problema PIL koristila je karakteristične točke lica za kreiranje značajki koje su potom predane tradicionalnom modelu strojnog učenja. Karakteristične točke su locirane koristeći C++ biblioteku dlib [26] u kombinaciji s programskim jezikom Python. Uz pomoć te biblioteke moguće je dobiti točno 68 karakterističnih točaka na ljudskom licu. Karakteristične točke formiraju skup *početnih točaka* $P = \{(x_i, y_i)\}_{i=1}^{68}$.

S obzirom na to da ljudi mogu pokazivati istu emociju u različitim dijelovima slike zbog pokreta glave, oblika lica etc., provedeno je skaliranje (engl. *scaling*) početnih točaka kako bi se taj efekt ublažio:

$$P_{scaled} = \left\{ \left(\frac{x_i - \bar{x}}{\max |P|}, \frac{y_i - \bar{y}}{\max |P|} \right) \right\}_{i=1}^{68}$$

gdje $(\bar{x}, \bar{y})$ označava srednju vrijednost skupa P, odnosno centroid na licu. Nazivnik je najveća apsolutna vrijednost između svih x i y vrijednosti iz skupa P. Nadalje, kreirane su i dodatne značajke koristeći udaljenosti od svake karakteristične točke do centroida svih karakterističnih točaka. Značajke udaljenosti (engl. *distance features*) izračunate su primjenom L_2-norme na karakteristične točke iz standardiziranog skupa točaka P_{std} te njihov centroid $(\bar{P}_{std,x}, \bar{P}_{std,y})$:

$$P_{dist} = \left\{ \sqrt{(P_{std,i,x} - \bar{P}_{std,x})^2 + (P_{std,i,y} - \bar{P}_{std,y})^2} \right\}_{i=1}^{68}$$

gdje je standardizirani skup karakterističnih točaka bio definiran kao:

$$P_{std} = \left\{ \left(\frac{x_i - \bar{x}}{\hat{\sigma_x}}, \frac{y_i - \bar{y}}{\hat{\sigma_y}} \right) \right\}_{i=1}^{68}$$

a $\hat{\sigma_x}$, $\hat{\sigma_y}$ su označavali procjene standardne devijacije x i y koordinata iz skupa početnih točaka P, respektivno.

Konačno, značajke P_{scaled} i P_{dist} spojene su u konačni vektor značajki s ukupnim brojem elemenata jednakim 204. Vektor značajki obuhvaćao je i x i y vrijednosti skupa P_{scaled} te sve vrijednosti udaljenosti iz skupa P_{dist}. Nadalje, u svrhu filtriranja redundantnih značajki, primijenjena je metoda univarijantnog odabira značajki. Ovisno o f-vrijednosti, koja je dobivena metodom analize varijance (engl. *analysis of variance, ANOVA*), odabrano je 120 značajki koje su pokazale statistički značajnu razliku kroz 6 različitih emocija.

Brojne tradicionalne metode su isprobane nad izvučenim značajkama: SVM, logistička regresija, algoritam k-najbližih susjeda, klasifikator slučajna šuma te klasifikator XGBoost. Odabran je SVM jer je generalizirao najbolje nad izrađenim značajkama.

3.2.2 Duboki pristup

Predstavljeni SVM model imao je dovoljno dobre performanse u statičnom slučaju, ali je generalizirao loše u situacijama kada sudionik eksperimenta nije gledao izravno u kameru zbog rotacije glave. To bi se zasigurno moglo objasniti nemogućnošću prediktora dlib-ovih karakterističnih točaka da pravilno postavi karakteristične točke na nagnuto lice. Stoga je započet razvoj dubokih modela koristeći jednu vlastitu konvolucijsku neuronsku mrežu te nekoliko predtreniranih.

Prvo je razvijena konvolucijska neuronska mreža s vlastitom arhitekturom (EM-CNN) za PIL koja je prikazana u tablici 2. Ulaz u mrežu EM-CNN bila je slika u sivim tonovima veličine 300×300 piksela s 3 ulazna kanala. Zbog toga što slike u sivim tonovima imaju samo jedan kanal, tri kanala su simulirana kopiranjem prvog kanala dva puta. Dodana je i augmentacija slike (engl. *image augmentation*) u skup za učenje uključujući slučajni vodoravni okret (engl. *random horizontal flip*) te slučajnu rotaciju slike gdje je kut rotacije biran slučajno iz intervala $[-45, 45]$. Na kraju, slike iz skupova za učenje, validaciju te testiranje bile su standardizirane koristeći srednju vrijednost od 0.516 i standardnu devijaciju od 0.247. Vrijednosti aritmetičke sredine i standardne devijacije bile su izračunate *a priori* na čitavom skupu podataka. Predstavljena mreža radila je bolje nego SVM model, što će biti konkretno pokazano u idućem odjeljku.

Međutim, skup podataka od skoro 4000 slika pokazao se nedovoljnim za treniranje EM-CNN mreže od nule kako bi se dobilo značajno poboljšanje nad SVM modelom. Iz tog razloga, primijenjena je metoda finog podešavanja (engl. *fine tuning*) na dvije popularne predtrenirane konvolucijske neuronske mreže: ResNet-34 [27] i Inception-v3 [28]. Na kraj ovih mreža nadodan je potpuno povezani sloj kako bi se omogućila klasifikacija u 6 klasa. Kao što će odjeljak 4 demonstrirati, Inception-v3 i Resnet-34 imali su slična generalizacijska svojstva na kreiranom skupu podataka. Zato su iskombinirane predikcije oba modela za primjenu u razrednom ekperimentu PIL.

U fazi učenja, Inception-v3 koristi dva izlazna sloja. Prvi izlaz je linearni sloj na kraju mreže, dok je drugi izlaz pomoćni. U fazi testiranja jedino se primarni izlaz koristi. Ulazna slika za Inception-v3 mora biti dimenzija 299×299 pa su slike iz skupa podataka bile malo smanjene prije nego su poslane na ulaz mreže kako bi bili zadovoljeni njezini zahtjevi.

ResNet-34 je duboka rezidualna konvolucijska neuronska mreža s 34 sloja. Dala je bolje rezultate nego ResNet-18 i ResNet-152, koji su također isprobani. Dimenzije ulazne slike za ResNet-34 iznose 224×224 piksela.

Sljedeći koraci predobrade podataka bili su isti za Inception-v3 i ResNet-34. Obje mreže očekuju tri kanala na ulazu, pa su slike u sivim tonovima iz skupa podataka duplicirane na dva preostala kanala. Nadalje, primijenjen je isti cjevovod predobrade

Tablica 2: Arhitektura konvolucijske neuronske mreže EM-CNN.

Tip sloja	Broj filtera	Veličina filtera	Korak (engl. *stride*)	Nadopunjavanje (engl. *padding*)	Dimenzije izlaza
Ulazni sloj	-	-	-	-	300×300×3
Konvolucijski sloj	16	11	1	1	292×292×16
Sloj grupne normalizacije	-	-	-	-	292×292×16
ReLU	-	-	-	-	292×292×16
Sloj sažimanja max. vrijednosti	1	2	2	0	146×146×16
Konvolucijski sloj	32	7	1	1	142×142×32
Sloj grupne normalizacije	-	-	-	-	142×142×32
ReLU	-	-	-	-	142×142×32
Sloj sažimanja max. vrijednosti	1	2	2	0	71×71×32
Konvolucijski sloj	64	6	1	1	68×68×64
Sloj grupne normalizacije	-	-	-	-	68×68×64
ReLU	-	-	-	-	68×68×64
Sloj sažimanja max. vrijednosti	1	2	2	0	34×34×64
Konvolucijski sloj	128	5	1	1	32×32×128
Sloj grupne normalizacije	-	-	-	-	32×32×128
ReLU	-	-	-	-	32×32×128
Sloj sažimanja max. vrijednosti	1	2	2	0	16×16×128
Konvolucijski sloj	256	3	1	1	16×16×256
Sloj grupne normalizacije	-	-	-	-	16×16×256
ReLU	-	-	-	-	16×16×256
Sloj sažimanja max. vrijednosti	1	2	2	0	8×8×256
Potpuno povezani sloj	-	-	-	-	16384

kao i za EM-CNN: slučajno horizontalno okretanje, slučajna rotacija te standardizacija skupa za učenje. Na validacijski i testni skup primijenjena je samo standardizacija.

3.3 Razredni eksperiment prepoznavanja izraza lica

Razredni eksperiment PIL brojao je ukupno 40 sudionika koji su trebali riješiti 8 zadataka vizualnog programiranja pomoću edukacijskog robota Lego Mindstorms EV3. Svaki zadatak mogao je biti riješen u dvije do pet minuta. Najduže vrijeme koje je bilo koji sudionik eksperimenta proveo rješavajući neki zadatak bilo je pola sata.

Lica sudionika snimana su tabletom smještenim ispred računala. Budući da je tablet bio malo ispod zaslona računala, snimljena lica su na snimkama bila blago nakošena. Snimani su i zasloni računala kako bi se sačuvali pokreti miša te pritisci tipaka na tipkovnici. Zato je bilo moguće točno znati u kojem je trenutku koji zadatak počeo i završio. S obzirom na to da su satovi na tabletu i računalu bili sinkronizirani, moglo se povezati trenutke kada je svaki ispitanik započeo i dovršio neki zadatak na računalu i videu. Videozapisi su podijeljeni po zadacima kako bi se omogućila analiza zavisnosti između emocija i zadataka, što su prve dvije varijable za analizu.

Nekoliko sudionika nije riješilo većinu zadataka. Zato su njihovi videozapisi iz-

bačeni te su nakon toga preostali videozapisi 32 sudionika (19 muških i 13 ženskih). Dodatno, neki sudionici nisu riješili zadnji, osmi zadatak pa za taj zadatak postoji samo 29 videozapisa.

Treća varijabla koja je uključena u eksplorativnu analizu podataka bila je spol. Htjelo se vidjeli postoji li razlika između predviđenih emocija muških i ženskih sudionika eksperimenta. Posljednja varijabla koja je uvedena je *tip aktivnosti*. Identificirana su četiri tipa aktivnosti koja su se javila tijekom eksperimenta:

1. Programiranje (engl. *programming, P*) - sudionik gleda u ekran, dodiruje miš ili tipkovnicu

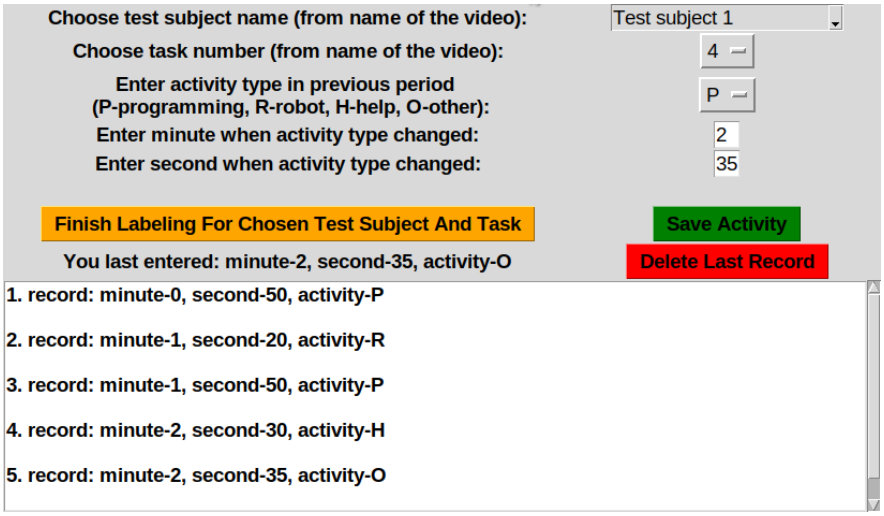
2. Robot (R) - sudionik gleda u robota, isprobava program koji je napisao za pokretanje na robotu i promatra ponašanje robota

3. Pomoć (engl. *help, H*) - postojale su situacije u kojima su sudionici zapeli s rješavanjem zadataka i trebali su pomoć s programskim dijelom ili pokretanjem robota pa ovaj se ovaj tip aktivnosti odnosi na slučaj u kojem su organizatori eksperimenta davali upute i navodili sudionike do rješenja

4. Ostalo (engl. *other, 0*) - sve što su sudionici eksperimenta radili, a da nije pokriveno s prve tri kategorije uključujući pričanje s drugim sudionicima, gledanje uokolo po učionici, hodanje, konzumiranje hrane i pića etc.

Naravno, bilo je potrebno nekako dobiti labelirane videozapise navedenim tipovima aktivnosti. U tu svrhu razvijena je jednostavna aplikacija pod nazivom "Video Labeler". Koristeći funkcionalnosti aplikacije, 14 anotatora labeliralo je videozapise svih 32 sudionika (gotovo 23 sata video materijala). Anotatori su bili obučeni da unose oznake samo kada bi došlo do promjene tipa aktivnosti. S obzirom na to da je ovaj zadatak označavanja bio prilično jednostavan, svaki video je označavao samo jedan anotator. Zbog toga nije moguće priopćiti mjeru slaganja između anotatora (engl. *inter-annotator agreement*). Primjer rada aplikacije prikazan je na slici 1.

Kao što je prethodno spomenuto, predviđanje emocija obavljeno je nakon razdvajanja videozapisa u zadatke. Svake sekunde videozapisa lice je detektirano pomoću SSD [25] modela. Budući da je bilo potrebno eliminirati detektirana lica koja su bila izuzetno nagnuta, odbačena su sva detektirana lica za koja je razina pouzdanosti detekcije bila ispod 95%. Kako bi se dobile još robusnije predikcije, iskorištene su predikcije od fino podešenih Inception-v3 i ResNet-34 modela. Emocija je u nekoj sekundi zabilježena samo ako su predviđanja oba modela bila jednaka. Na ovaj način dobiveno je 39144 predviđenih emocija nad otprilike 23 sata videozapisa. To znači da je opisanim



Slika 1: Primjer označavanja videozapisa uz pomoć aplikacije “Video Labeler”.

postupkom 48% sličica (engl. *frame*) iz videozapisa označeno emocijama na razini sekunde. Važno je napomenuti da je bilo moguće da se ponekad ispred kamere nađe više od jednog lica. Tipična situacija kada se ovo događalo je kada bi sudionik tražio pomoć. U tim situacijama, gdje je bilo više detektiranih lica, zadržano je ono s najvećom detektiranom površinom graničnog pravokutnika (engl. *bounding box*). Nad tim licem je onda obavljena predikcija.

4 Eksperimentalni rezultati

Eksperimenti su rađeni nad modelima te nad predikcijama emocija s odabranim najboljim modelima. Modeli su evaluirani kroz standardne metrike i uspoređeni uz pomoć statističkih testova. Predviđene emocije su statistički analizirane kroz postupak eksplorativne analize podataka.

4.1 Rezultati evaluacije modela

Kroz rezultate evaluacije uspoređuju se četiri modela: SVM s ručno izrađenim značajkama, EM-CNN, fino podešena mreža Inception-v3 i fino podešena mreža ResNet-34. Svi modeli su učeni, validirani i testirani na istim podjelama originalnog skupa podataka.

4.1.1 Detalji odabira modela

Za SVM model korištena je jezgra radijalna bazna funkcija (engl. *radial basis function kernel*). Naštimavani su regularizacijski parametar C te jezgreni koeficijent γ koristeći

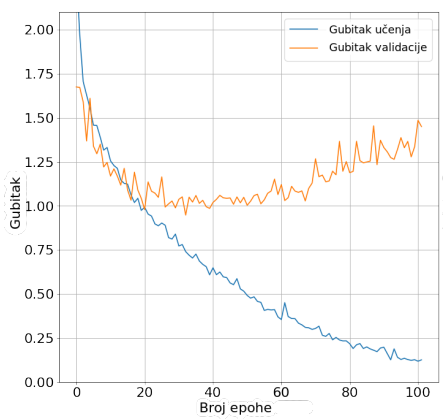
metodu pretraživanja po rešetki (engl. *grid search*) s 5-strukom unakrsnom validacijom (engl. *5-fold cross validation*). Za oba parametra testirane su sve vrijednosti iz skupa $\{2^{-15}, 2^{-14}, ..., 2^{9}, 2^{10}\}$. Hiperparametri koji su dali najvišu točnost na validacijskom skupu bili su: $C = 16$, $\gamma = 0.25$.

Konvolucijske neuronske mreže su imale različite hiperparametre za optimizaciju s obzirom na SVM, ali iste međusobno. Najbolji hiperparametri za tri razvijene mreže predstavljeni su u tablici 3. Isprobani su stohastički gradijentni spust (engl. *stochastic gradient descent, SGD*) te Adam kao optimizatori. Adam je dao bolje rezultate. Optimalna arhitektura za EM-CNN je ona iz tablice 2. Za sve modele, najbolji raspoređivač stope učenja (engl. *learning rate scheduler*) bio je multiplikativni. Taj raspoređivač množi stopu učenja svake skupine parametara s istim multiplikativnim faktorom u svakoj epohi. Najbolja generalizacija na validacijskom skupu postignuta je kada je stopa učenja bila reda veličine 10^{-4}. Hiperparametri neuronskih mreža optimirani su slučajnom pretragom (engl. *randomized search*). Implementirana je i metoda ranog zaustavljanja (engl. *early stopping*) koja je pratila promjene u točnosti na validacijskom skupu. Ako nije bilo poboljšanja u točnosti na validacijskom skupu u uzastopnih 50 epoha, učenje je zaustavljeno. Za svaku mrežu parametri koji su izabrani kao optimalni bili su oni koji su postigli najvišu točnost na validacijskom skupu neovisno o metodi ranog zaustavljanja.

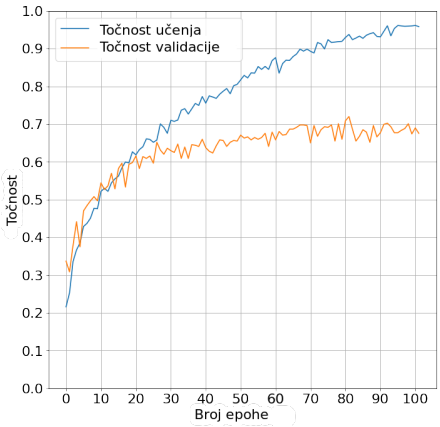
Tablica 3: Hiperparametri koji su dali najbolje rezultate na validacijskom skupu za konvolucijske neuronske mreže (fino podešen model Inception-v3 – **Inception-v3**; fino podešen model ResNet-34 – **ResNet-34**; potpuno povezani – **PP**).

Mreža	Arhitektura	Optimizator	Najbolja stopa učenja	Raspoređivač stope učenja	Rano zaustavljanje
EM-CNN	vidi tablicu 2	Adam	0.0005	multiplikativni (faktor=0.99)	da
Inception-v3	ista kao u [28] s PP slojem na vrhu	Adam	0.0005	multiplikativni (faktor=0.95)	da
ResNet-34	ista kao u [27] s PP slojem na vrhu	Adam	0.0006	multiplikativni (faktor=0.99)	da

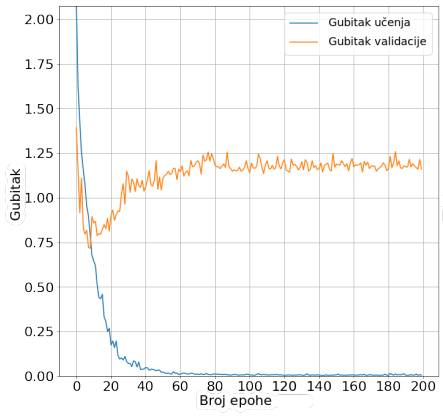
Grafovi gubitaka unakrsne entropije (engl. *cross-entropy losses*) te točnosti kroz epohe dani su slikom 2. Maksimalan broj epoha nije postavljen unaprijed jer je korištena metoda ranog zaustavljanja. Posljedično, svaka mreža imala je različit konačni broj epoha. Generalno govoreći, grafovi gubitaka i točnosti za sve mreže pratili su slične trendove. Gubitak na skupu za učenje padao je gotovo monotono, dok je gubitak na validacijskom skupu padao do neke točke gdje je počeo ponovno rasti. Točnosti na skupu za učenje gravitirale su prema vrijednosti 1.0, dok su točnosti na validacijskom skupu rasle do vrijednosti između 0.6 i 0.8 te su oscilirale oko tih vrijednosti dok metoda ranog zaustavljanja nije prekinula učenje.



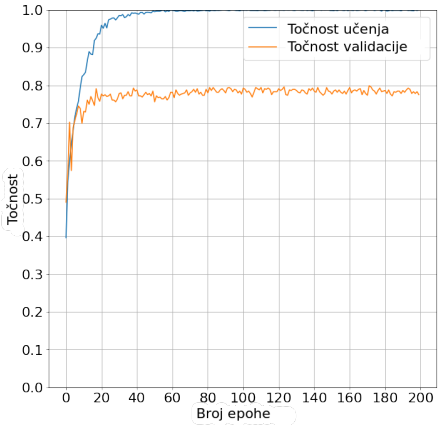
(a) Gubici mreže EM-CNN.



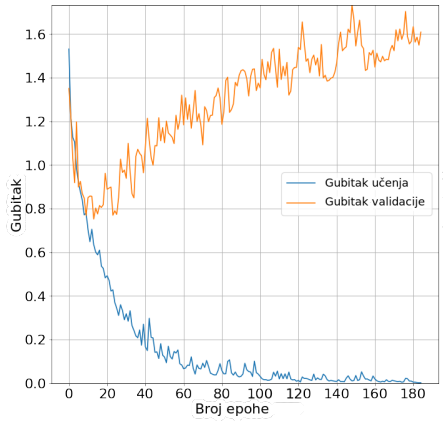
(b) Točnosti mreže EM-CNN.



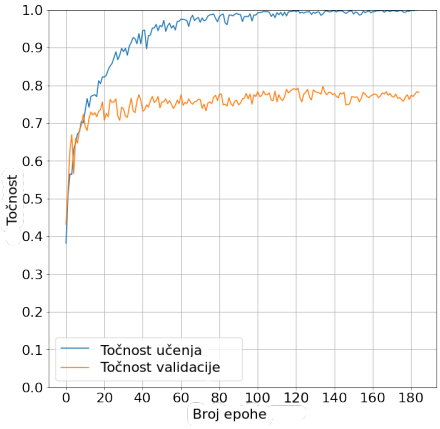
(c) Gubici mreže Inception-v3.



(d) Točnosti mreže Inception-v3.



(e) Gubici mreže ResNet-34.



(f) Točnosti mreže ResNet-34.

Slika 2: Gubici unakrsne entropije i točnosti kroz epohe za vrijeme faza učenja i validacije konvolucijskih neuronskih mreža. Gubici su prikazani u lijevom, dok su točnosti dane u desnom stupcu slike. Grafovi su izloženi za tri konvolucijske neuronske mreže: **EM-CNN**, **Inception-v3** (fino podešen model) te **ResNet-34** (fino podešen model), respektivno prema recima slike od vrha prema dnu.

4.1.2 Evaluacija

Tablica 4 sadrži evaluacijske rezultate na skupu za testiranje za četiri prethodno opisana modela. Korištene su sljedeće standardne metrike za evaluaciju: makro preciznost (engl. *precision, P*), makro odziv (engl. *recall, R*), makro mjera F1 (engl. *F1-score, F1*) te točnost (engl. *accuracy, Acc*). Fino podešena mreža Inception-v3 postigla je najbolje rezultate na skupu za testiranje kroz sve evaluacijske metrike. ResNet-34 prati njezine rezultate s malom razlikom. EM-CNN te SVM imaju veće padove u svim evaluacijskim metrikama. S obzirom na to da vjerojatnost slučajnog pogotka ispravne emocije za 6 definiranih klasa iznosi 16.67%, svaki od modela postigao je prilično dobre rezultate.

Tablica 4: Rezultati evaluacije korištenih modela kroz preciznost, odziv, mjeru F1 te točnost na testnom skupu dani u postotcima. Podebljane brojke ističu najbolje rezultate za svaku metriku.

Model	Metrika			
	P	R	F1	Acc
SVM	58.617	59.328	57.978	59.367
EM-CNN	65.958	65.77	65.622	65.57
Fino podešen Inception-v3	**79.033**	**78.994**	**78.811**	**78.987**
Fino podešen ResNet-34	77.002	76.417	76.259	76.456

4.1.3 Statistička usporedba modela

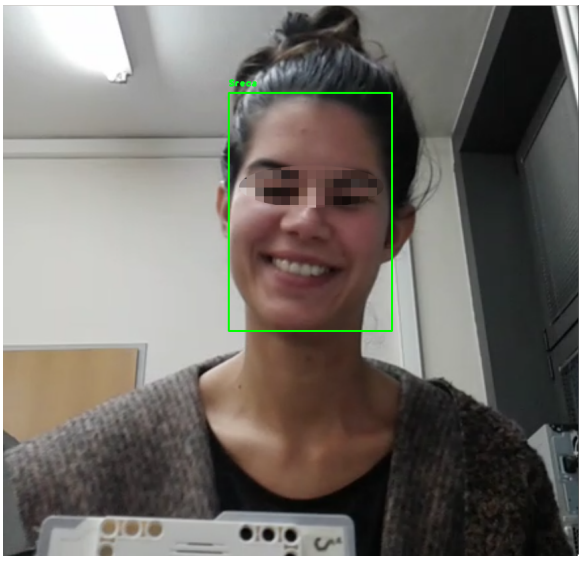
Dvostrani permutacijski testovi (engl. *permutation test*) provedeni su nad predikcijama modela na skupu za testiranje. Rezultati su prikazani u tablici 5. Testovi su rađeni nad makro F1 metrikom. P-vrijednosti pokazuju kako su svi rezultati statistički značajni uz razinu značajnosti $\alpha_b = 0.00167$ dobivenu primjenom Bonferroni korekcije na početnu razinu značajnosti $\alpha = 0.01$, osim rezultata testa između fino podešenih modela **Inception** i **ResNet**. Stoga, nije ispravno tvrditi da postoji statistički značajna razlika u generalizaciji između ta dva modela. Međutim, može se tvrditi da je SVM dokazano imao najlošije rezultate, a drugi najlošiji model bila je mreža EM-CNN. S obzirom na to da nije statistički potvrđeno koji je od fino podešenih modela bolji, odlučeno je da će se njihove predikcije kombinirati pri primjeni na videozapise razrednog eksperimenta.

Tablica 5: P-vrijednosti dvostranih permutacijskih testova za parove modela (fino podešen model Inception-v3 - **Inception**; fino podešen model ResNet-34 - **ResNet**).

	SVM	**EM-CNN**	**Inception**	**ResNet**
SVM	-	$\mathbf{0.0003}$	$\mathbf{< 10^{-5}}$	$\mathbf{< 10^{-5}}$
EM-CNN	-	-	$\mathbf{< 10^{-5}}$	$\mathbf{< 10^{-5}}$
Inception	-	-	-	0.0542
ResNet	-	-	-	-

4.2 Eksplorativna analiza podataka

Nad videozapisima je obavljena predikcija emocija u vremenu, a predviđene emocije zabilježene su samo u slučaju da su oba odabrana prediktivna modela (Inception-v3 i ResNet-34) dala jednako predviđanje. To je jedan od razloga zašto postoje dijelovi videozapisa za koje nema zapisanih predikcija. No, to ne predstavlja problem s obzirom na to da nije bitno u kojoj točno sekundi se neka emocija javila, već je bitno u kojem se kontekstu emocija pojavila u odnosu na rješavani zadatak, spol sudionika eksperimenta te tip aktivnosti koji je obavljan u trenutku pojave emocije. Primjer predikcije nad isječkom iz videozapisa za vrijeme aktivnosti *robot* prikazan je na slici 3.



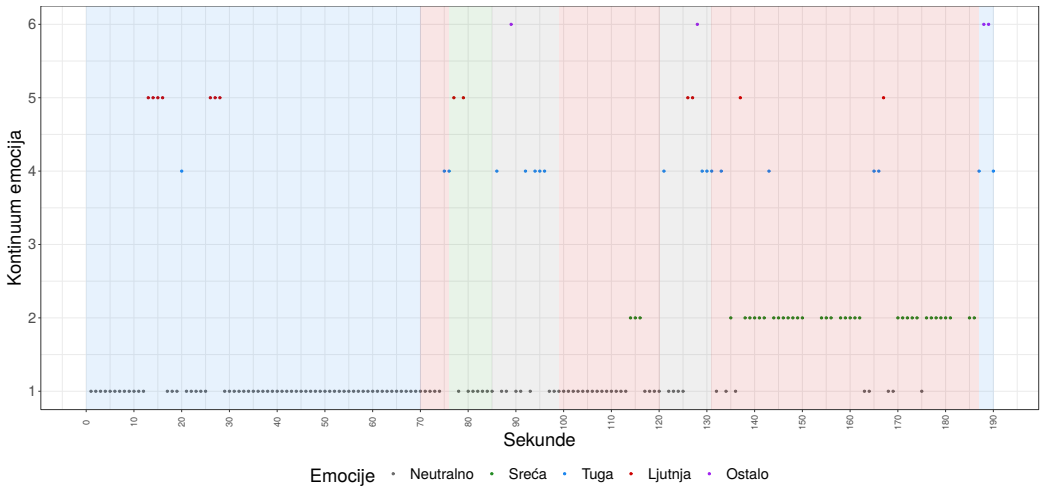
Slika 3: Primjer isječka iz videozapisa razrednog eksperimenta. Na slici se može uočiti situacija u kojoj sudionica eksperimenta promatra kretanje robota nakon uspješnog prebacivanja i pokretanja napisanog programa. Prediktivni modeli Inception-v3 te ResNet-34 složili su se kako se u ovom kadru iz videozapisa nalazi lice koje pokazuje emociju sreće.

Ukupne brojke predviđenih emocija u svim videozapisima za muškarce i žene odvojeno te zajedno izložene su u tablici 6. Predviđanje nad 23 sata videozapisa trajalo je 4 sata i 30 minuta na procesoru Intel® Core™ i7-6700HQ i računalu s 8 gigabajta RAM-a. To znači da se za jednu sekundu videozapisa predikcija uz pomoć dva modela obavi u roku od 0.2 sekunde. Može se uočiti kako su najčešće predviđene emocije *neutralno* i *tuga* neovisno o spolu. Ovako visoke brojke za neutralnu emociju bile su očekivane jer su većinu trajanja razrednog eksperimenta sudionici programirali i nisu pokazivali emocije. S druge strane, visoke brojke za emociju tuge bile su neočekivan ishod primjene prediktivnih modela na videozapise. No, postoji jednostavno objašnjenje. Iz razloga što su kamere bile postavljene tako da snimaju lica blago od dolje, neutralna lica gledana odozdo mogu izgledati tužno. Kako bi se analize radile nad što nepristranijim podacima, emocije *neutralno*, *tuga* i *ostalo* nisu korištene u nastavku.

Tablica 6: Pobrojane predviđene emocije nad videozapisima za muške i ženske sudionike razrednog eksperimenta. Frekvencije su sortirane silazno po broju zajedničkih predviđenih emocija (neovisno o spolu).

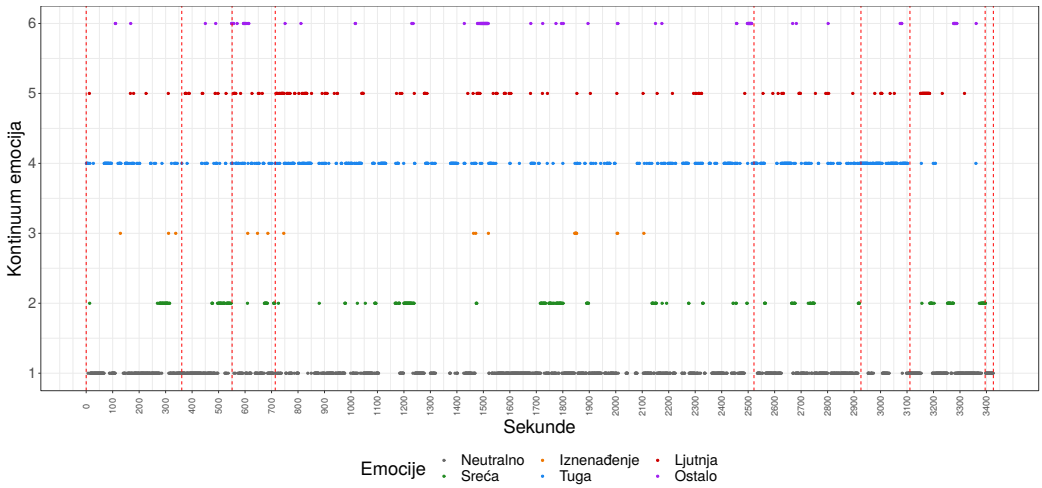
Emocija	Muškarci	Žene	Zajedno
Neutralno	9634	7185	16819
Tuga	8080	8090	16170
Sreća	1106	1902	3008
Ljutnja	410	966	1376
Ostalo	282	776	1058
Iznenađenje	321	392	713

Zanimljivo je pogledati kako izgleda vremenski prikaz predviđenih emocija za jednog konkretnog sudionika i zadatak iz eksperimenta. Slika 4 prikazuje upravo to. Na y-osi nalazi se takozvani kontinuum emocija gdje su emocije prikazane stupnjevito po brojevima (1-neutralno, 2-sreća, 3-iznenađenje, 4-tuga, 5-ljutnja, 6-ostalo). Ovaj konkretan primjer sadrži primjere svih emocija osim emocije *iznenađenja* te primjere svih tipova aktivnosti čije je trajanje prikazano obojanim pravokutnicima. Iz slike se može zaključiti kako je najčešće pokazivana emocija bila *neutralno* te kako je najduži dio vremena u danom zadatku sudionik eksperimenta potrošio na programiranje (plavi pravokutnik). Interesantno je primijetiti kako se ovdje emocija ljutnje javila najviše puta za vrijeme programiranja, a emocija sreće za vrijeme eksperimentiranja s robotom (crveni pravokutnik) dok se od prisutnih emocija klasa *ostalo* najrjeđe pojavila.



Slika 4: Prikaz predviđenih emocija u vremenu za jednog sudionika eksperimenta i njegov drugi zadatak. Ovdje su predikcije napravljene koristeći samo fino podešeni Inception-v3 model. Plavi pravokutnici označavaju aktivnost *programiranje*, crveni aktivnost *robot*, zeleni aktivnost *pomoć*, a sivi aktivnost *ostalo*.

Slika 5 donosi prikaz predviđenih emocija u vremenu nad svih osam zadataka istog sudionika eksperimenta kao sa slike 4. S obzirom da je ovakav prikaz vrlo nepregledan, teško je analizirati emocije i tipove aktivnosti u odnosu na apsolutno vrijeme. Jedino što je sa slike jasno vidljivo je kako najviše dolaze do izražaja emocije *neutralno* i *tuga* zbog već opisanih razloga. Kod ovakvog prikaza postoji još jedan problem, a to je da nisu svi sudionici eksperimenta utrošili niti jednaku niti sličnu količinu vremena na rješavanje zadataka. Zato je donesena odluka o analizi emocija u odnosu na tipove aktivnosti, spol i zadatke kroz vjerojatnosne distribucije i statističke testove.

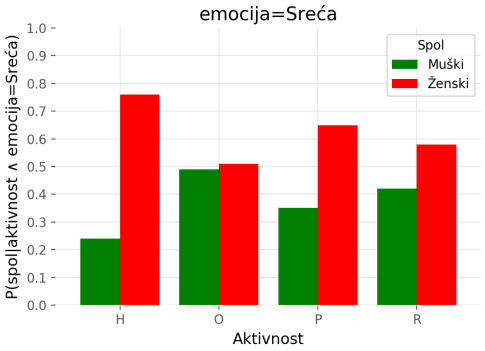


Slika 5: Prikaz predviđenih emocija u vremenu za jednog sudionika eksperimenta i svih osam zadataka koje je riješio. Ovdje su predikcije napravljene koristeći samo fino podešeni Inception-v3 model. Okomite crvene linije predstavljaju graničnike između zadataka. Aktivnosti nisu obilježene radi preglednosti.

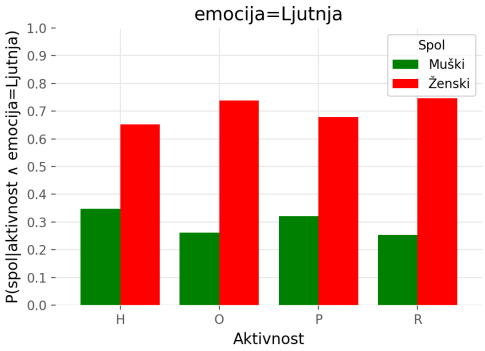
Za eksplorativnu analizu povezanosti između emocija, aktivnosti i spola, korištene su vjerojatnosne distribucije čiji su parametri procijenjeni koristeći procjenitelj najveće izglednosti (engl. *maximum likelihood estimator, MLE*). Vjerojatnosti su procijenjene uz pomoć izraza:

$$P(varijabla_1 = v_1 | varijabla_2 = v_2 \wedge varijabla_3 = v_3) = \frac{C(v_1, v_2, v_3)}{C(v_2, v_3)}$$

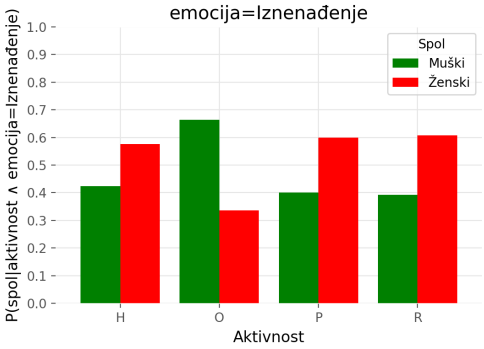
gdje C označava frekvenciju, a svaka od varijabli $varijabla_1, varijabla_2, varijabla_3$ može opisivati emociju, aktivnost ili spol. Na primjer, za vjerojatnosne distribucije sa slike 6 vrijedi $varijabla_1 = spol, varijabla_2 = aktivnost, varijabla_3 = emocija$. Vjerojatnosti se sumiraju u 1.0 po varijabli spol. To znači da su stupčasti prikazi usporedivi samo unutar varijable spol. Grafovi su razdvojeni po emocijama radi bolje preglednosti te su prikazani za tri odabrane emocije: sreća, ljutnja i iznenađenje. Slika 6a pokazuje kako su ženski sudionici eksperimenta s većom vjerojatnošću pokazivali emociju sreće kroz sve aktivnosti. Isto vrijedi i za sliku 6b te emociju ljutnje, ali ne i za sliku 6c gdje su muški sudionici za vrijeme aktivnosti *ostalo* s većom vjerojatnošću pokazivali emociju iznenađenja. Zanimljivo je primijetiti i kako je kod ženskih sudionika došlo do vrlo visoke vjerojatnosti pokazivanja emocije sreće za vrijeme aktivnosti *pomoć* te vrlo visoke vjerojatnosti pokazivanja emocije ljutnje za vrijeme aktivnosti *robot*.



(a) Rezultati za emociju sreće.



(b) Rezultati za emociju ljutnje.

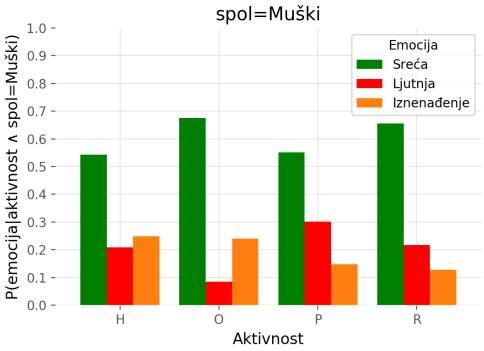


(c) Rezultati za emociju iznenađenja.

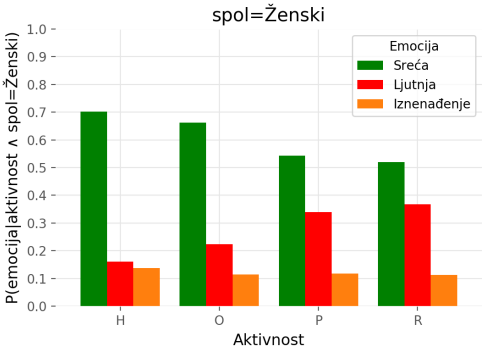
Slika 6: Procijenjene vjerojatnosne distribucije $P(spol|aktivnost \wedge emocija)$ za tri odabrane emocije. Vjerojatnosti su procijenjene koristeći MLE. Aktivnosti su prikazane sažeto: H-pomoć, O-ostalo, P-programiranje, R-robot.

Na sličan način kao i kod prethodne analize, metodom MLE procijenjene su vjerojatnosne distribucije $P(emocija|aktivnost \wedge spol)$ što znači da vrijedi $varijabla_1 = emocija, varijabla_2 = aktivnost, varijabla_3 = spol$. Vjerojatnosti se sumiraju u 1.0 unutar varijable *emocija*. Rezultati su prikazani na slici 7. Prvi zaključak koji se nameće je da neovisno o tome je li spol muški ili ženski, emocija sreće je najvjerojatnija[6] unutar svih tipova aktivnosti. Drugi zanimljiv zaključak je da je ljutnja kao emocija druga najvjerojatnija unutar gotovo svih tipova aktivnosti neovisno o spolu. Ne uzimajući u obzir par protuprimjera, emocija iznenađenja imala je najmanju vjerojatnost pojavljivanja od sve tri emocije.

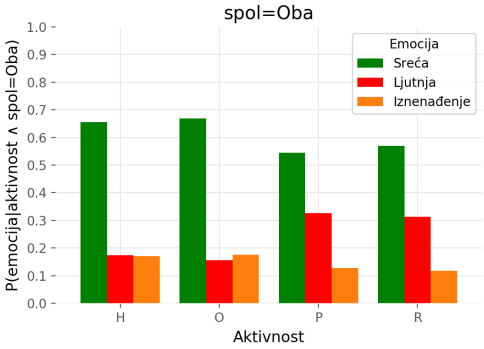
[6]u ovom pojednostavljenom scenariju gdje su razmatrane samo emocije sreće, ljutnje i iznenađenja



(a) Rezultati za muški spol.



(b) Rezultati za ženski spol.

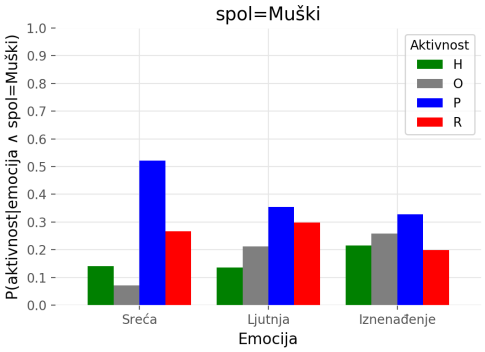


(c) Rezultati neovisno o spolu.

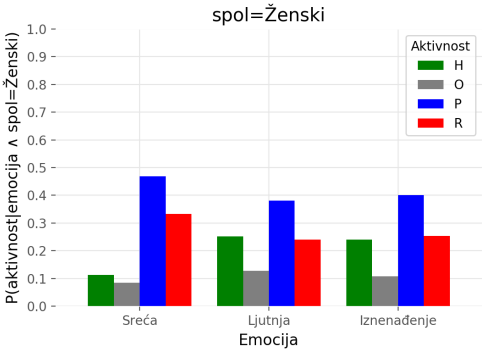
Slika 7: Procijenjene vjerojatnosne distribucije $P(emocija|aktivnost \wedge spol)$ za tri odabrane emocije. Vjerojatnosti su procijenjene koristeći MLE. Aktivnosti su prikazane sažeto: H-pomoć, O-ostalo, P-programiranje, R-robot.

Posljednja analiza kroz vjerojatnosne distribucije opisuje procijenjene vjerojatnosti $P(aktivnost|emocija \wedge spol)$ metodom MLE dok za varijable vrijedi $varijabla_1 = aktivnost, varijabla_2 = emocija, varijabla_3 = spol$. Vjerojatnosti se sumiraju u 1.0 unutar varijable *aktivnost*. Ovakav prikaz omogućuje usporedivost vjerojatnosti pojavljivanja pojedine emocije kroz tipove aktivnosti. Interesantno je uočiti kako su oblici distribucija za sudionike eksperimenta ženskog spola sa slike 8b odredili oblike distribucija neovisno o spolu prikazane na slici 8c. Gledajući grafove, nameće se zaključak kako je, neovisno o spolu, najveća vjerojatnost da će se neka od emocija sreće, ljutnje ili iznenađenja javiti upravo za vrijeme aktivnosti *programiranje*. Takva raspodjela može se objasniti time što je aktivnost programiranja trajala najduže od svih ostalih aktivnosti neovisno o zadatku ili sudioniku eksperimenta pa je i veća vjerojatnost da će se neka emocija javiti za vrijeme te aktivnosti. Nadalje, odmah nakon programiranja, najvjerojatnije je javljanje bilo koje od emocija za vrijeme aktivnosti *robot* ako se promatra situacija sa slike 8b premda se pojava emocije iznenađenja unutar aktivnosti

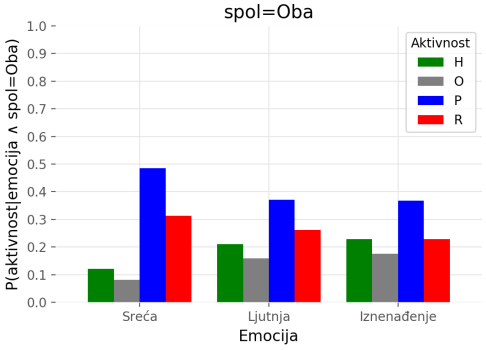
pomoć i robot čini jednako vjerojatna.



(a) Rezultati za muški spol.



(b) Rezultati za ženski spol.



(c) Rezultati neovisno o spolu.

Slika 8: Procijenjene vjerojatnosne distribucije $P(aktivnost|emocija \wedge spol)$ za tri odabrane emocije. Vjerojatnosti su procijenjene koristeći MLE. Aktivnosti su prikazane sažeto: H-pomoć, O-ostalo, P-programiranje, R-robot.

U svrhu analize zavisnosti između emocija i aktivnosti po zadacima, proveden je znatan broj χ^2 testova nezavisnosti. Rezultati su dani tablicom 7. Uz odabranu razinu značajnosti od $\alpha = 0.01$, dokazano je da postoji zavisnost između tri analizirane emocije i četiri tipa aktivnosti gotovo u svim zadacima. Kod muškaraca nije dokazana zavisnost za zadatke 2, 5 i 6 dok kod žena nije dokazana zavisnost za zadatke 3 i 5. Međutim, ako se promatraju svi zadaci zajedno, ovisno i neovisno o spolu sudionika eksperimenta, postoji zavisnost između pokazivanja emocija sreće, ljutnje i iznenađenja te tipa aktivnosti koji je u trenutku pokazivanja emocije obavljao sudionik eksperimenta.

Tablica 7: P-vrijednosti nakon provedbe χ^2 testa nezavisnosti između kategoričkih varijabli *emocija* i *aktivnost*. Testovi su provedeni ovisno i neovisno o spolu te zadacima. Znak $\perp\!\!\!\perp$ označava postojanje zavisnosti između varijabli. Podebljane su p-vrijednosti za rezultate testova gdje je dokazana zavisnost između emocija i tipova aktivnosti.

Zadatak	$H_0 : emocija \not\perp\!\!\!\perp aktivnost$ $H_1 : emocija \perp\!\!\!\perp aktivnost$		
	Muškarci	Žene	Zajedno
1	$\mathbf{7.7758^{-9}}$	$\mathbf{0.0011}$	$\mathbf{6.9398^{-14}}$
2	0.3476	$\mathbf{8.4836^{-7}}$	$\mathbf{4.3635^{-6}}$
3	$\mathbf{2.0015^{-6}}$	0.1992	$\mathbf{5.8443^{-6}}$
4	$\mathbf{4.7162^{-11}}$	$\mathbf{3.0712^{-8}}$	$\mathbf{0.0016}$
5	0.4471	$\mathbf{1.2940^{-6}}$	$\mathbf{2.0383^{-5}}$
6	0.0469	0.0215	0.0425
7	$\mathbf{0.0007}$	$\mathbf{4.4209^{-11}}$	$\mathbf{2.1151^{-8}}$
8	$\mathbf{3.2560^{-7}}$	$\mathbf{1.4299^{-10}}$	$\mathbf{3.8455^{-16}}$
Svi zadaci	$\mathbf{1.4910^{-16}}$	$\mathbf{1.5996^{-20}}$	$\mathbf{2.8511^{-28}}$

5 Zaključak

Problem računalnog prepoznavanja izraza lica u stvarnom vremenu pokazao se izrazito zahtjevnim za rješavanje, ali i korisnim za primjenu u sklopu razrednog eksperimenta. Kroz ovaj rad dan je detaljan pregled postojećih rješenja te je sakupljen skup podataka sastavljen od slika iz tri poznata skupa podataka za prepoznavanje emocija: CK+, FER-2013 i SFEW. Zatim su nad skupom podataka razvijena četiri modela: jedan s metodama tradicionalnog strojnog učenja te tri s metodama dubokog učenja. Modeli su temeljito uspoređeni po uspjehu na skupu za testiranje, a od svih modela odabrana su dva za predviđanje emocija na videozapisima sudionika eksperimenta koji su snimani dok su rješavali zadatke vizualnog programiranja i ekperimentirali s edukacijskim robotom Lego Mindstorms EV3. Najbolji modeli bili su Inception-v3 i ResNet-34. Njihovom primjenom na videozapise dobivene su predikcije emocija u vremenu po zadacima za 32 sudionika eksperimenta. Time je omogućena provedba eksploративne analize podataka koja je dala zanimljive uvide u odnose između emocija, spola, zadataka i tipova aktivnosti. Aktivnosti su bile manualno označene pomoću aplikacije "Video Labeler". Iz svega navedenog, može se zaključiti kako je doprinos ovog rada znanstvenoj zajednici višestruk. Evenutalna poboljšanja, te ideje za budući rad, mogla bi uključivati prikupljanje većeg broja podataka za učenje, ponovno provedbu razrednog eksperimenta s boljim pozicioniranjem kamere te strože uvjete na ponašanja sudionika u razrednom ekperimentu kako bi se eliminirale aktivnosti koje su spadale u tip *ostalo*.

Literatura

[1] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

[3] I. J. Goodfellow, D. Erhan, P. Luc Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59 – 63, 2015, special Issue on "Deep Learning of Representations". [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014002159

[4] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, 2012.

[5] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 461–466. [Online]. Available: https://doi.org/10.1145/2663204.2666275

[6] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3866–3870.

[7] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[8] S. Wang, Y. Yuan, X. Zheng, and X. Lu, "Local and correlation attention learning for subtle facial expression recognition," *Neurocomputing*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231220313679

[9] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "Oaenet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognition*, p. 107694, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320320304970

[10] G. Viswanatha Reddy, C. Dharma Savarni, and S. Mukherjee, "Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features," *Cognitive Systems Research*, vol. 62, pp. 23 – 34, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389041720300206

[11] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2559–2573, 2013.

[12] M. Garcia Villanueva and S. Ramirez Zavala, "Deep neural network architecture: Application for facial expression recognition," *IEEE Latin America Transactions*, vol. 18, no. 07, pp. 1311–1319, 2020.

[13] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[14] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning affective video features for facial expression recognition via hybrid deep learning," *IEEE Access*, vol. 7, pp. 32 297–32 304, 2019.

[15] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 500–508, 2006.

[16] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial expression recognition in videos using dynamic kernels," *IEEE Transactions on Image Processing*, vol. 29, pp. 8316–8325, 2020.

[17] A. T. Kabakus, "Pyfer: A facial expression recognizer based on convolutional neural networks," *IEEE Access*, vol. 8, pp. 142 243–142 249, 2020.

[18] H. Kim, Y. Kim, S. J. Kim, and I. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2980–2992, 2018.

[19] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 6977–6991, 2020.

[20] D. Yang, A. Alsadoon, P. Prasad, A. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Procedia Computer Science*, vol. 125, pp. 2 – 10, 2018, the 6th International Conference on Smart Computing and Communications. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917327679

[21] K. P. Seng and L. Ang, "Video analytics for customer emotion and satisfaction at contact centers," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 3, pp. 266–278, 2018.

[22] C. Bian, Y. Zhang, F. Yang, W. Bi, and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning," *IET Computer Vision*, vol. 13, no. 3, pp. 329–337, 2019.

[23] S. LI, Y. Gao, F. Wang, and T. SHI, "Research on classroom expression recognition based on deep circular convolution self-encoding network," in *2020 15th International Conference on Computer Science Education (ICCSE)*, 2020, pp. 523–528.

[24] G. Tonguç and B. Ozaydın Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture," *Computers & Education*, vol. 148, p. 103797, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131519303471

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *Lecture Notes in Computer Science*, p. 21–37, 2016. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_2

[26] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

David Dukić

Računalno prepoznavanje izraza lica u stvarnom vremenu metodama dubokog učenja s primjenom na snimke eksperimenta u razrednoj okolini

Sažetak

Prepoznavanje izraza lica može se promatrati kao klasifikacijski zadatak. Ovaj rad donosi rješenje tog zadatka u četiri koraka. Prvo je provedeno iscrpno istraživanje sličnih radova odnosno korištenih skupova podataka te modela strojnog učenja. Zatim su sakupljene slike iz tri standardna skupa podataka za prepoznavanje emocija. Nakon toga, kako bi se omogućila automatska klasifikacija izraza lica u šest predefiniranih klasa emocija, razvijena su i evaluirana četiri modela strojnog učenja. Konačno, modeli koji su postigli najveću makro F1 točnost na skupu za testiranje bili su kombinirani s ciljem prepoznavanja emocija u stvarnom vremenu u razrednoj okolini. Najveće točnosti predviđanja od 78.811% te 76.259% postignute su uz pomoć fino podešenih popularnih predtreniranih konvolucijskih neuronskih mreža Inception-v3 te ResNet-34. 32 osobe su sudjelovale u razrednom eksperimentu prepoznavanja izraza lica. Izrazi lica sudionika snimani su koristeći kameru tableta za vrijeme rješavanja osam zadataka vizualnog programiranja i eksperimentiranja s Lego Mindstorms EV3 edukacijskim robotom. Nadalje, najbolji modeli su primijenjeni kako bi se dobile predikcije emocija na razini svake sekunde snimljenih videa. Predikcije su onda statistički analizirane kako bi se uočilo postoji li korelacija između emocija, zadatka koji je rješavan, spola i tipa aktivnosti. Tipovi aktivnosti su proizašli iz promatranja sudionika za vrijeme eksperimenta. Donesena je odluka o četiri tipa aktivnosti: *programiranje*, *robot*, *pomoć* i *ostalo*. Aktivnosti su bile ručno labelirane uz pomoć anotatora za svaki video. Rezultati statističke analize pokazali su kako ženski sudionici eksperimenta sa znatno većom vjerojatnošću pokazuju emocije sreće, ljutnje i iznenađenja kroz gotovo sve tipove aktivnosti. Nadalje, unutar svih tipova aktivnosti, emocija sreće se najvjerojatnije pojavljivala neovisno o spolu. Konačno, tipovi aktivnosti unutar kojih su se najčešće javljale emocije sreće, ljutnje i iznenađenja bili su *programiranje* i *robot*. Analiza zavisnosti emocija i aktivnosti kroz zadatke pokazala je kako zavisnost postoji za većinu zadataka neovisno o spolu.

Ključne riječi: Inception-v3, konvolucijska neuronska mreža, prepoznavanje izraza lica, ResNet-34, tip aktivnosti

David Dukić

Real-Time Facial Expression Recognition using Deep Learning with Application to Recordings of the Experiment in the Classroom Environment

Summary

Facial expression recognition can be viewed as a classification task. This work tackled it in four steps. First, exhaustive related work research of previously used data sets and machine learning models was conducted. Second, images from three standard emotion recognition data sets were collected. Third, in order to enable automatic classification of facial expressions into six predefined emotion classes, four machine learning models were developed and evaluated. Finally, the models that achieved the highest macro F1-score on the test set were combined with the goal of real-time emotion prediction in a classroom environment. The highest prediction scores of 78.811% and 76.259% were accomplished by fine-tuning popular pre-trained convolutional neural networks Inception-v3 and ResNet-34, respectively. 32 participants took part in the facial expression recognition classroom experiment. Their facial expressions were recorded using a tablet camera during the process of solving eight visual programming tasks and experimenting with Lego Mindstorms EV3 educational robot. Afterwards, top-performing models were applied to retrieve emotion predictions each second of the recorded videos. Predictions were then statistically analyzed to observe whether there exists any correlation between emotions, the task at hand, gender and activity type. Activity types arose from observing participants during the experiment. The four activity types were decided upon: *programming*, *robot*, *help*, and *other*. They were manually labeled by annotators for every video. The results of the statistical analysis showed that female participants in the experiment were much more likely to show emotions of happiness, anger, and surprise through almost all activity types. Furthermore, within all types of activities, the emotion of happiness most likely occurred regardless of gender. Finally, the types of activities within which emotions of happiness, anger, and surprise were most common were *programming* and *robot*. Analysis of the dependence of emotions and activities through tasks showed that dependence exists for most tasks regardless of gender.

Keywords: Inception-v3, convolutional neural network, facial expression recognition, ResNet-34, activity type