SVEUČILIŠTE U SPLITU
FAKULTET ELEKTROTEHNIKE, STROJARSTVA I BRODOGRADNJE

POSLIJEDIPLOMSKI DOKTORSKI STUDIJ ELEKTROTEHNIKE I INFORMACIJSKIH TEHNOLOGIJA

KVALIFIKACIJSKI ISPIT

NAPREDNA SUČELJA ZA INTERAKCIJU ČOVJEKA I RAČUNALA

Luka Kraljević

Split, ožujak 2019.

Sadržaj

1. Uvod 3

2. Modeli dubokog učenja 5

2.1. Umjetne neuronske mreže 5

2.2. Konvolucijske neuronske mreže 6

2.3. Povratne neuronske mreže 9

3. Sučelje za prepoznavanje gesti 11

3.1. Stereo računalni vid 12

3.2. Arhitekture dubokog učenja kod prepoznavanje gesti 14

4. Sučelje mozak-računalo 17

4.1. Prepoznavanje emocija 18

5. Glasovno sučelje 21

5.1. Automatsko prepoznavanje govora 21

5.2. Arhitekture dubokog učenja za automatsko prepoznavanje govora 23

5.3. Udaljeno prepoznavanje govora 24

6. Zaključak 27

LITERATURA 28

1. Uvod

Pojam sučelje (*eng. Interface*) odnosi se na programsku podršku i/ili sklopovlje preko kojeg je korisniku omogućena interakcija sa različitim vrstama uređaja. Svrha sučelja je pružiti korisniku jednostavnu i učinkovitu kontrolu sa ciljem postizanja što veće točnosti i preciznosti.

U današnje vrijeme svjedočimo osjetnom razvoju senzora koji postaju sve manji, jeftiniji, pouzdaniji i učinkovitiji a omogućavaju i realizaciju sposobnosti komunikacije. Ovi ključni faktori, kao i dostupnost drugih tehnologija u koje se isti mogu integrirati, pridonose pojavi novih oblika potrošačke elektronike koje sačinjavaju dosadašnji svakodnevni predmeti sa ugrađenom dovoljnom razinom inteligencije da postaju inteligentni uređaji. Jedan od najznačajnijih oblika takovih uređaja koji u posljednje vrijeme imaju osjetan stupanj proliferacije su i nosivi elektronički uređaji (*eng. wearble devices - wearbles*). Unutar navedene kategorije spadaju i brojni drugi uređaji koji otvaraju nove aspekte primjene tehnologije kroz široki raspon aplikacija kao što su, među ostalima, pametne kuće, inteligentni transportni sustavi – pametna vozila, itd. Sve ove okolnosti pridonijele su situaciji gdje se ljudi trenutno nalaze u novoj vrsti okoline odnosno pametnim okolinama. Razvoj pametnih okolina iziskuju razradu novih koncepata i pristupa u interakciji koje se protežu izvan okvira današnjih standardnih oblika a nude naprednije, intuitivnije i prirodnije načine interakcije.

Dosadašnja standardna ulazna sučelja poput miša i tipkovnice postaju nepraktična a trenutni smjer tehnološke evolucije korisničkih sučelja orijentiran je prema oblicima interakcije urođenim u ljudsku biologiju odnosno na temelju koje nastaje tehnološko područje prirodnih korisničkih sučelja (*eng. Natural User Interfaces- NUI*). Prirodno korisničko sučelje je sustav za interakciju sa uređajima gdje je korisniku omogućeno upravljanje kroz intuitivne radnje povezane s svakodnevnim ljudskim ponašanjem. Definicija NUI-a odnosi se na senzorske ulaze kao što su geste i govor ali i na opise sustava koji su tehnološki napredniji i inteligentniji - svjesni konteksta, uz sposobnost prepoznavanja lica, okoline i emocija. Prava revolucija u metodama interakcije doći će kada prijeđemo sa grafičkog sučelja na prirodna korisnička sučelja, od miša i tipkovnice do govora, gesti pa čak i misli.

Napretkom računalne snage, mrežnih infrastruktura, računanja u oblaku te prikupljanjem velikih količina potrebnih podataka uređajima je omogućeno razumijevanje i tumačenje prirodne ljudske interakcije. Dolaskom dubokog učenja inspiriranog radom ljudskog mozga mnoga istraživačka polja poput računalnog vida i prepoznavanja govora su postigla značajan napredak. Kroz različite eksperimente pokazano je da računalo može obavljati zadatke prepoznavanje puno bolje nego čovjek, međutim efikasnost sustava temeljenih na dubokom učenju često degradira u stvarnim praktičnim - realnim uvjetima. Sve dok performanse ovakvih sustava ne dostignu ljudsku razinu u realnim uvjetima, za zaključiti je da će istraživanja u ovim poljima biti i dalje biti inspirirana ljudskim sustavom.

U komunikaciji između ljudi, vrsta i omjer neverbalnih i verbalnih faktora čini mjeru kvalitete prijenosa informacije. Upravo ovi faktori sačinjavaju prirodnu i intuitivnu vrstu izražavanja. U skupini verbalnih faktora nalazi se govor i sluh te sposobnost ispravnog, točnog i preciznog izražavanja te, sa druge strane, i ispravnog, točnog i preciznog shvaćanja i interpretiranja prenošene informacije. Bez obzira na logičan zaključak da su verbalni faktori oni koji određuju mjeru kvalitete komuniciranja i prijenosa informacija, dominantan faktori i oni koji konačno upotpunjuju kontekst prenošene informacije su upravo neverbalni faktori. U ovoj skupini dominantan udio ima gestikulacija odnosno izražavanje gestama, pokretima ruku, glave i tijela, odnosno mimikom. Sljedeći faktor također spada u kategoriju neverbalnih faktora a kritičan je za definiranja konteksta komunikacije, a to je upravo emocionalno stanje i kontekst unutar komunikacije. Iz navedenoga, kao dominantni i kritični faktori komunikacije, ističu se govor i prepoznavanje govora, korištenje gesta i elemenata gestikulacije te konačno, prepoznavanje emocionalnog konteksta. Prirodna i intuitivna komunikacija mora se temeljiti na upravo ovim dominantnim faktorima.
Kako bi komunikacija između čovjeka i uređaja dosegla istu razinu intuitivnosti i prirodnosti, ista se također mora prvotno i dominantno temeljiti na faktorima govora i prepoznavanja govora, prepoznavanja i ispravne interpretacije gestikulacije te prepoznavanja emocionalnog konteksta odnosno prepoznavanja emocija čovjeka. Govor, gestikulacija i emocije čovjekovo su svakodnevno i prirodno okruženje.

Ovaj rad svoju će pozornost obratiti upravo na faktore govora, gestikulacije i emocije odnosno sposobnosti pametne okoline za prepoznavanje govora, gesti i emocija. Rad će pružati pregled područja za napredna sučelja odnosno sučelja prepoznavanja gesti, sučelja sa sposobnosti integracije afektivnog računanja te sa posebnim naglaskom na sučelja za prepoznavanje govora.

U drugom poglavlju dan je pregled metoda dubokog učenja koje se koriste kod navedenih područja. U trećem poglavlju je objašnjeno upravljanje bez-kontaktnim gestama (eng. Touchless user interface (TUI)). Objašnjeni su osnovni principi računalnog vida korišteni za percepciju dubine, te dan je pregled algoritama dubokog učenja korištenih kod prepoznavanja gesti. U poglavlju četiri opisano je sučelje za upravljanje mislima (eng. Brain computer interface - BCI) te je dan pregled najnovijih istraživanja povezanih s detekcijom emocionalnih stanja na temelju EEG signala. U poglavlju pet će biti objašnjeno automatsko prepoznavanje govora, kao i osnovna arhitektura dubokog učenja za pretvorbu govora u tekst. Iznijeti će se pregled područja iz prepoznavanja govora temeljen na dubokom učenju te će se naglasak staviti na metode za postizanje robusnosti govornog sučelja (Voice user interface VUI) kao najprirodnije vrste ljudske interakcije.

2. Modeli dubokog učenja

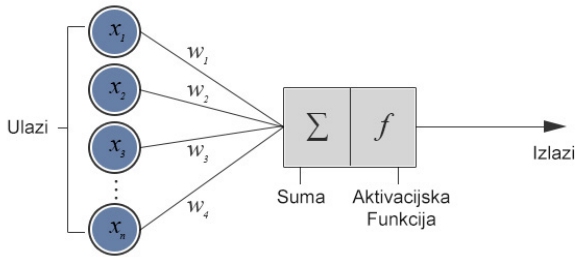
Tijekom zadnjih godina duboko učenje, grana umjetne inteligencije inspirirana strukturom ljudskog mozga napravila je ogromni napredak u davanju strojevima sposobnosti razumijevanja fizičkog svijeta. Povijesno gledano, računala obavljaju zadatke koristeći deterministične algoritme koji detaljno specificiraju sve potrebne korake. Iako ovakav pristup daje zadovoljavajuće rezultate u mnogim situacijama, pružanje eksplicitnog algoritma nije uvijek moguće kao što je u slučajevima kod prepoznavanja govora, lica, emocija ili gesti. Pokušavajući pristupiti tim izazovima kroz paradigmu ručnog kodiranja, osim intenzivnog rada potrebnog za definiranje velikog broja parametara poput atributa koji opisuju lice ili fonema, pokazalo se da i strojevi nisu mogli obraditi podatke koji se ne uklapaju u eksplicitno definirane parametre. Nasuprot tome sustavi dubokog učenja imaju mogućnost razumijevanja podatka bez potrebe za eksplicitnim algoritmom. Postoje različiti modeli dubokog učenja koji su pokazali uspjeh u specifičnim domenama. U nastavku ćemo objasniti varijante modela koji tvore osnovu za realizaciju prirodnih korisničkih sučelja – prepoznavanje gesti, emocija i govora.

2.1. Umjetne neuronske mreže

Umjetne neuronske mreže (*engl. Artifical Neural Networks - ANN*) su model strojnog učenja inspiriran načinom funkcioniranja ljudskog mozga. ANN se sastoji od umjetnih neurona, koji su međusobno povezani vezama, gdje je svaka veza karakterizirana svojom težinom, koja se u procesu učenja mijenja na način da se prilagođava željenom izlazu. Kao što je već spomenuto neuron je osnovni građevni blok neuronske mreže, a predstavljen je matematičkom funkcijom (2.1) koja transformira ulaz u izlaz. Izlazne vrijednosti su određene vrijednostima s ulaza te težinama veza na koju se primjenjuje aktivacijska funkcija.

$$y = h_w(x) = \frac{1}{1 + e^{-wTx}} \tag{2.1}$$

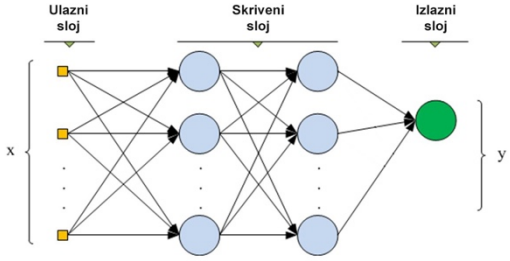
gdje je x ulazni vektor, w vektor težina a y izlazni signal. Na slici 1.1 je prikazan neuron sa n ulaza



Slika 2.1. Model neurona. Izvor [1]

Unaprijedna mreža (*engl. feedforward network*) ili višeslojni perceptron (*engl. multi-layer perceptron - MLP*) je mreža koja preslikava skup ulaznih podataka u skup prikladnih izlaza. Arhitektura MLP je takva da su neuroni međusobno povezani tako da izlaz iz jednog neurona

može biti ulaz u drugi. MLP se sastoji od slojeva gdje se krajnje lijevi naziva ulazni sloj (*engl. input layer*) kranje desni izlazni sloj (*engl. output layer*) a slojevi u sredini se nazivaju skriveni slojevi (*engl. hidden layer*). Na slici 2.2 je prikazan MLP sa *n* ulaza, dva skrivena sloja i jednim izlazom.



Slika 2.2. Višeslojni perceptron. Izvor [2]

Da bi mreža davala točne rezultate potrebno je pronaći skup parametara *w* koji daju najbolju aproksimaciju izlaza *y*, odnosno potrebno je mrežu istrenirati. Treniranje neuronskih mreža postiže se korištenjem *Back-propagation* algoritma. Algoritam se sastoji od unaprijedne propagacije (*engl. forward propagation*) gdje se na ulaz dovodi poznati skup uzoraka te se računa odziv mreže. Nakon toga se računa pogreška kao razlika odziva i očekivanog izlaza te se ta pogreška potom propagira unazad. Zatim se primjenjuje neka od metoda optimizacije tako da se mijenjaju vrijednosti težina veza među neuronima, parametri modela bi se trebali istrenirati tako da se smanji funkcija koštanja.

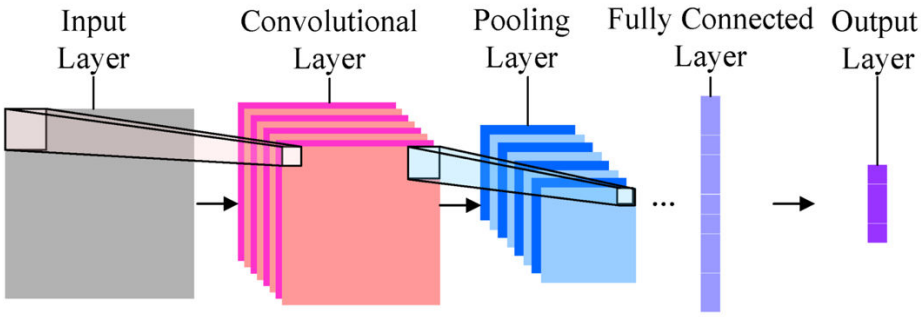
Većina modernih modela dubokog učenja (*engl. Deep Neural Network - DNN*) se zasniva na umjetnim neuronskim mrežama, a značenje riječi duboka mreža se odnosi na broj skrivenih slojeva koji se koriste u obradi podataka. Dodani slojevi omogućavaju modeliranje kompleksnih nelinearnih veza među ulaznim podacima tako da svaki sloj transformira ulazne podatke u nove značajke. U dubokom učenju više slojeva mreže označava višu razinu naučenih značajki.

2.2. Konvolucijske neuronske mreže

Konvolucijske neuronske mreže (*engl. Convolutional Neural Networks - CNN*) su jako slične klasičnim DNN mrežama uz razliku što CNN arhitektura posjeduje posebne slojeve za izvlačenje značajki. Ovi slojevi generiraju značajke za slijedeće slojeve te nema potrebe za prethodno „ručno izrađenim“ značajkama kakve se koriste kod DNN.

Arhitektura CNN je inspirirana organizacijom ljudskog vizualnog korteksa gdje individualni neuroni mozga reagiraju na podražaje samo u ograničenom području vizualnog polja poznatom kao receptivno polje. Receptivna polja se preklapaju da bi se pokrilo cijelo vizualno područje. Konvolucijske neuronske mreže se temelje na tri osnovne ideje: lokalna receptivna polja, dijeljenje težina i sažimanje.

Na slici 2.3 prikazana je opća struktura konvolucijskih neuronskih mreža.



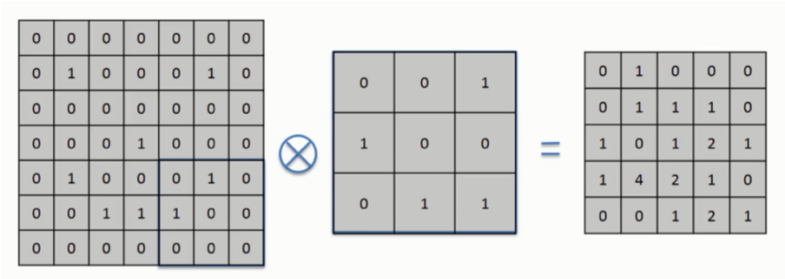
Slika 2.3. Struktura konvolucijski neuronskih mreža. Izvor [3]

Arhitektura CNN se sastoji od tri osnova tipa sloja: na ulazu mreže su konvoluciski slojevi i slojeva sažimanja (engl. *pooling*) nakon čega slijedi nekoliko potpuno povezanih slojeva istih kao kod običnih ANN.

Klasični DNN na svome ulazu očekuje jednodimenzionalne podatke, koji do izlaza prolaze niz transformacija kroz skrivene slojeva. Kod ove arhitekture u svakom sloju neuroni su potpuno vezani sa neuronima slijedećeg sloja, dok neuroni pojedinog sloja međusobno ne dijele nikakve veze. Ovakav model dubokog učenja pretpostavlja da su značajke nezavisne, što često nije slučaj.

S druge strane CNN na svome ulazu očekuje podatke određene strukture, uglavnom matrice. Umjesto povezivanja pojedinog podatka sa ulaznim neuronom, CNN svaki ulazni neuron povezuje s malim, lokaliziranim regijama podataka koji se nazivaju receptivna polja. Arhitektura CNN se zasniva na pretpostavci o međusobnoj ovisnosti ulaznih podatka.

Konvolucijski sloj je osnovni građevni blok CNN mreže. U njemu se obavlja operacija konvolucije nad ulaznom matricom koristeći matricu koja se ovisno o literaturi naziva kernel ili filter. Primjer konvolucije prikazan je na slici 2.4.

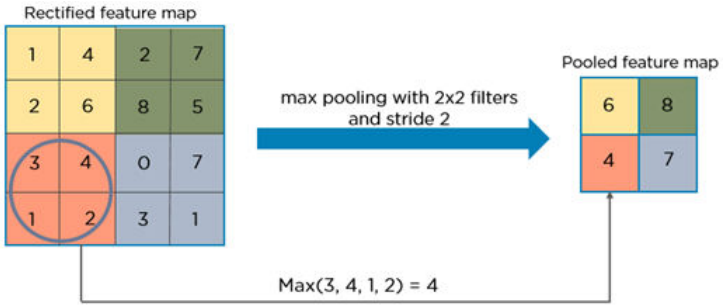


Slika 2.4. Operacija konvolucije. Izvor[4]

Princip rada konvoluciskog sloja zasniva se na dijeljenju težina. Svi neuroni prvog skrivenog sloja imaju jednak faktor pristranosti (*engl. bias*) kao i istu matricu težina koja definira vezu s odgovarajućim receptivnim poljem. Ovo znači da svi neuroni prvog skrivenog sloja detektiraju istu vrstu značajki unutar pripadajućeg receptivnog polja. Zbog toga se mapiranje podataka s ulaznog sloja prema skrivenom sloju naziva mapa značajki (engl. *feature map*).

Načelno, neuroni u konvolucijskom sloju raspoređeni su u tri dimenzije gdje treća dimenzija definira broj mapa značajki. Za pojedinu mapu značajki u operaciji konvolucije koristi se jedna jezgra (engl. *kernel*) definirana zajedničkim težinama i faktorom pristranosti.

Osim upravo opisanih konvolucijskih slojeva, CNN također sadrži i slojeve sažimanja (engl. *pooling*). Uloga ovog sloja je postupno smanjenje prostorne veličine značajki radi smanjenja količine parametara i računanja a samim time se vrši kontrola od pretreniranja (engl. *overfitt*). Slojevi sažimanja se obično koriste odmah nakon konvolucijskih slojeva kako bi pojednostavnili izlazne informacije mape značajki. U praksi se najčešće koristi sloj sažimanja kod kojeg se na mapu značajki primjenjuje filter veličine 2x2 s pomakom od 2 duž cijele širine i visine čime se postiže odbacivanje 75% aktivacija. Kao funkcija filtra najčešće se koriste maksimalna vrijednost i usrednjavanje. Na slici 2.5 prikazano je sažimanja maksimalnom vrijednošću.



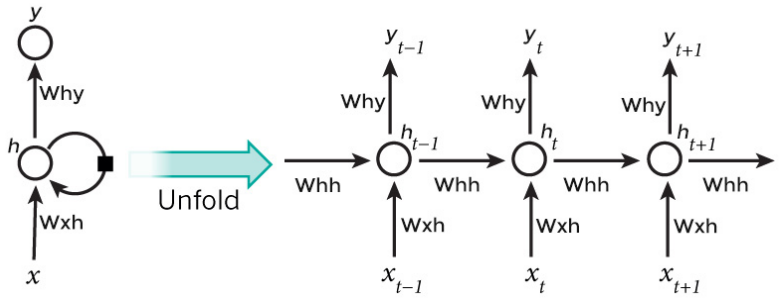
Slika 2.5.: Sažimanje značajki maksimalnom vrijednošću. Izvor:[5]

Arhitektura konvolucijskih neuralnih mreža je dizajnirana po uzoru na tehniku ljudskog mozga za percepciju slika i kao takva najveći uspjeh je doživjela u području računalnog vida. Ovaj tip neuralnih mreža radi uz pretpostavku strukturne ovisnosti značajki kao što je riječ kod slika, bliski pikseli vjerojatno pripadaju istom objektu ili vizualnoj strukturi. Primjena konvolucijskih neuralnih mreža radi dobre sposobnosti samostalnog učenja značajki dovela je do velikog napretka u područjima prepoznavanja, klasifikacije i segmentacije slika [6, 7, 8] te se kao takva uobičajeno koristi i kod detekcije i prepoznavanja gesti [9]-[21]. Konvolucijske neuronske mreže su također pokazale dobre rezultate i u drugim područjima poput afektivnog računanja gdje su autori kao ulazne podatke koristili slike transformiranog EGG signala [22, 23, 24, 25] ili čisti signal [26]. Postoje i brojna istraživanja u području prepoznavanja govora [27, 29, 29, 30, 31] koja ukazuju na povećanje performansi kod akustičnog modela kada se DNN ili Gaussian mixture models (GMMs) zamjene sa CNN arhitekturom. U [32] je demostriran *„end-to-end*“ sustav za prepoznavanje govora koji se sastoji samo od konvolucijske mreže i CTC algoritma, te je pokazano da takav sustav može ostvariti *state-of-the-art* rezultate. Također pokazalo se da konvolucijske neuronske mreže mogu uvesti poboljšanje u smislu robusnosti na buku [33, 34].

2.3. Povratne neuronske mreže

Povratna neuronska mreža (*engl. Recurrent Neural Networks - RNN*) je model umjetne neuronske mreže gdje veze među neuronima čine usmjereni ciklus (*engl. directed cycle*). Takav ciklus ili povratno djelovanje je povezano sa operacijama koje imaju vremensko kašnjenje. Ideja iza RNN-ova jest korištenje sekvencijalnih informacija. Kod tradicionalnih ANN pretpostavlja se da su svi ulazi međusobno neovisni, što nije povoljno za obavljanje određenih zadataka npr. ako želimo predvidjeti slijedeću riječ u rečenici potrebno je znati prethodne riječi. RNN se zovu povratne jer ove mreže obavljaju isti zadatak za svaki element sekvence gdje izlaz ovisi i o prethodnom izračunu. Ovo ukazuje da RNN ima svojstvo pamćenja (memorije) informacija o prethodnim koracima koji su predstavljeni kroz unutrašnje stanje mreže. U teoriji RNN može raditi sa proizvoljno dugom sekvencom, ali u praksi ograničeni su samo na nekoliko vremenskih koraka. Iako je moguće slaganje više RNN jednu povrh druge, sama RNN mreža je duboki model jer se vremenskim odmatanjem RNN stvaraju duboki slojevi mreže. (duljina rečenice). Vrijeme u povratnim neuronskim mrežama označava redoslijed kojim se podaci šalju na ulaz.

Pod odmatanjem mreže podrazumijevamo da je mreža napisana za kompletnu sekvencu. Na primjer ako se rečenica (sekvenca) koju promatramo sastoji od 10 riječi to znači da mrežu moramo odmotati u 10 slojeva, svaki sloj za jednu riječ. Slijedeći dijagram 2.6 prikazuje vremensko odmotavanje RNN mreže.



Slika 2.6. Povratna neuronska mreža razvijena u punu mrežu. Izvor[35]

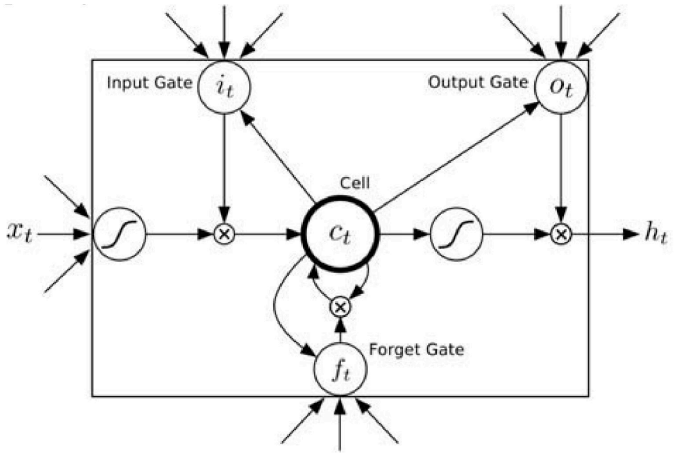
RNN je bitno drukčija od unaprijednih neuronskih mreža iz razloga što RNN ne djeluje samo na temelju ulaznih signala već i na temelju unutrašnjih stanja. Unutrašnja stanja šifriraju informaciju iz prošlosti(vremenskoj sekvenci) koja je već obrađena od strane RNN. U ovom smislu RNN možemo promatrati kao dinamični sistem koji obavlja transformaciju ulaza na izlaz. Korištenje unutrašnjih stanja (memorije) RNN omogućuje reprezentaciju i učenje serijski proširenih ovisnosti tokom dugog perioda, bar u teoriji. Matematički opis jednostavne RNN mreže s jednim sakrivenim slojem, koja se često koristi u područjima obrade signala u smislu nelinearnog unutrašnjeg modela stanja dan je slijedećim izrazim. U svakom trenutku t, neka je $\boldsymbol{x}_t$ ulazni vektor dimenzija Kx1, $\boldsymbol{h}_t$ Nx1 vektor vrijednosti skrivenih stanja a $\boldsymbol{y}_t$ Lx1 vektor izlaznih vrijednosti. Tada se jednostavna RNN može opisati kao:

$$\boldsymbol{h}_t = f(\boldsymbol{W}_{xh}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1}) \tag{2.2}$$

$$\boldsymbol{y}_t = g(\boldsymbol{W}_{hy}\boldsymbol{h}_t) \tag{2.3}$$

Gdje je $\boldsymbol{W}_{hy}$ LxN matrica težina od N skrivenih jedinica do L izlaza, $\boldsymbol{W}_{xh}$ je NxK matrica koja spaja K ulaza sa N skrivenih stanja i $\mathbf{W}_{hh}$ je NxN matrica težina koja povezuje N skrivenih stanja iz trenutka $t-1$ u trenutak t. $\mathbf{u}_t = \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}$ je N x 1 je vektor potencijala skrivenog sloja, $\mathbf{v}_t = \mathbf{W}_{hy}\mathbf{h}_t$ je Lx1 vektor potencijala izlazong sloja, $f(\mathbf{u}_t)$ je aktivacijska funkcija skrivenog sloja a $g(\mathbf{v}_t)$ je izlazna aktivacijska funkcija. Aktivacijska funkcija skrivenog sloja je najčešće *sigmoid*, *tanh* ili *RELU*, dok izlazni sloj najčeće koristi *softmax* funkciju.

Neuralna mreža s povratnom vezom u svom osnovnom obliku ne može modelirati kompleksniju vremensku dinamiku, i u praksi se pokazalo da je jako ograničena kad je u pitanju veličina ulazne sekvence (ne može gledati daleko u prošlost). Jedno od rješenja je ugrađivanje memorije u RNN, takozvanih ćelija s dugoročnom memorijom (engl. *long-short-term memory - LSTM*) Osnovna idea LSTM ćelija u RNN mreži je korištenje raznih vrsta vrata (*eng. gate*) radi bolje kontrole toka informacija kroz mrežu. Struktura LSTM ćelije prikazana je na slici 2.7



Slika 2.7 Arhitektura ćelije s dugoročnom memorijom. Izvor:[35]

Korištenjem LSTM ćelija u svrhu zamjene osnovnog mapiranja ulaza na skriveni sloj, te u svrhu zamjene tranzicija među skrivenim slojevima rješavamo problem eksplodirajućeg gradijenta.

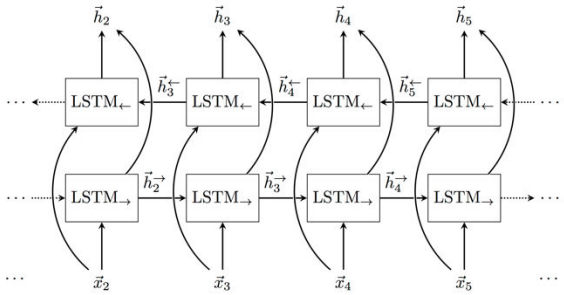
Konvencionalne povratne neuronske mreže obrađuju podatke samo u jednom smjeru, međutim u određenim područjima kao što je prirodna obrada jezika i prepoznavanje govora poželjno bilo koristiti informacije iz prošlosti i budućnosti [37]. Dvosjmerna povratna neuronska mreža (Bidirectional RNNs - BRNN) proširuje arhitekturu RNN uvođenjem dodatnog skrivenog sloja kod kojeg podaci putuju u suprotnom, negativnom vremenskom smjeru. Vremensko-prostorna formulacija dana je izrazom:

$$\overrightarrow{\mathbf{h}_t} = f(\mathbf{W}_{x\vec{h}}\mathbf{x}_t + \mathbf{W}_{\vec{h}\vec{h}}\overrightarrow{\boldsymbol{h}}_{t-1} + \mathbf{b}_{\vec{h}}) \tag{2.4}$$

$$\overleftarrow{\mathbf{h}_t} = f(\mathbf{W}_{x\overleftarrow{h}}\mathbf{x}_t + \mathbf{W}_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{\boldsymbol{h}}_{t+1} + \mathbf{b}_{\overleftarrow{h}}) \tag{2.5}$$

$$\mathbf{y}_t = \mathbf{W}_{\vec{h}y}\overrightarrow{\mathbf{h}_t} + \mathbf{W}_{\overleftarrow{h}y}\overleftarrow{\mathbf{h}_t} + \mathbf{b}_y \tag{2.6}$$

Kombinacijom BRNN sa LSTM ćelijama dobivamo dvosmjernu LSTM mrežu, koja može modelirati kompleksniju vremensku dinamiku u oba smjera kao što je prikazano na slici 2.8.



Slika 2.8. Dvosmjerna LSTM Izvor:[36]

Povratne neuronske mreže predstavljaju snažan mehanizam u modeliranju vremenskog slijeda, te su u novijoj literaturi neizostavni dio procesa kod automatskog prepoznavanja govora [38, 39, 40, 41, 42, 43, 44]. RNN arhitektura se koristi i kod vremenskog modeliranja EEG signala što je dovelo do značajnog poboljšanja u točnosti kod prepoznavanja ljudskih emocija [22, 45, 46, 47]. Problem sekvencijalnog modeliranja pojavljuje se i u području prepoznavanja gesti gdje se RNN arhitektura primijenjena na video signal [48, 49, 50, 51, 52, 53].

3. Sučelje za prepoznavanje gesti

Prepoznavanje gesti je vrsta perceptivnog računalnog korisničkog sučelja koje omogućuje računalima detekciju i tumačenje ljudske geste kao naredbe. Gesta se definira kao bilo koji fizički pokret, veliki ili mali, koji se može protumačiti senzorom pokreta. Neverbalna komunikacija je komunikacija kroz geste ruku, stavove tijela i izraze lica te sačinjava oko dvije trećine ukupne ljudske komunikacije [54]. Geste ruke su jedan od najčešćih oblika jezika tijela koji se koristi za komunikaciju i interakciju. Dok ostatak tijela uglavnom ukazuje na emocionalno stanje, geste ruke mogu posjedovati specifičan jezični sadržaj [55].
Sustavi za bez-kontaktnu interakciju mogu se podijeliti ovisno o načinu dobivanja gesti na ulazu pa razlikujemo sustave koji koriste žičane rukavice, stereo kamere, kamere dubine, termalne kamere i radare. U ovom radu razmatrat će se samo prepoznavanje gesti iz slike/videa odnosno temeljeno na postupcima računalnog vida.

Istraživanje u području prepoznavanja gesti je proteklih godina je dobilo na popularnosti kod istraživača na području računalnog vida i dubokog učenja. Neriješeni izazovi kao što su pouzdana identifikacija faze gestikulacije, osjetljivost na veličinu, oblik i varijacije brzine te problemi zbog okluzije ograničili su upotrebu sustava za prepoznavanje gesti ruke kao pouzdanog modaliteta u dizajnu sučelja.

Prepoznavanje gesti ruke se može kategorizirati na više načina. Podjela gesta može biti vezana za vremensku ovisnost geste pa razlikujemo statičke i dinamičke geste. Statičke geste koje se

još nazivaju pozama su one u kojima se položaj ruke ne mijenja tijekom razdoblja gestikulacije te se uglavnom oslanjaju na oblik, orijentaciju, relativnu poziciju u odnosno na tijelo i kut savijanja prstiju. Kod dinamičkih gesta ruku, položaj ruke se neprestano mijenja s u odnosu na vrijeme. Dinamičke geste obično imaju tri faze: priprema, pomak i retrakcija (povlačenje)[56]. Informacija o gesti je sadržana u vremenskoj sekvenci faze pomaka. Dinamičke geste se pored oblika i kutova savijanja prstiju oslanjaju i na putanju i orijentacije ruke. Podjela gesti može biti i na temelju interpretiranog značenja. Na primjer, amblemi i ilustratori su tipične klase koje opisuju geste kod neverbalne komunikacije. Amblemi su geste koje se mogu zamijeniti za izgovorene riječi (pokazujući palac gore umjesto zamjenjuje verbalnu poruku da je sve uredu). Ilustratori su geste koje se koriste za ilustraciju izgovorenih riječi (davanje uputa usmjeravanjem).

3.1. Stereo računalni vid

Računalni 3D vid je postupak izvlačenja 3D informacija iz digitalne slike. Slično kao kod biološkog procesa binokularnog vida, usporedbom informacija o sceni iz dvije perspektive 3D informacije se mogu izdvojiti ispitivanjem relativnih položaja objekta.
Stereo vid je tehnika dobivanje dubine iz dvije ili više kamera. Kod algoritama za detekciju dubine se koristi pojednostavljeni model kamere tzv. *pinhole* kamera prikazana na slici 3.1.

Slika 3.1 Model pinhole kamere [57]

Kod ovog modela kamere prikaz scene se dobije projekcijom 3D točaka na površinu slike koristeći perspektivnu transformaciju.

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{31} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3.1}$$

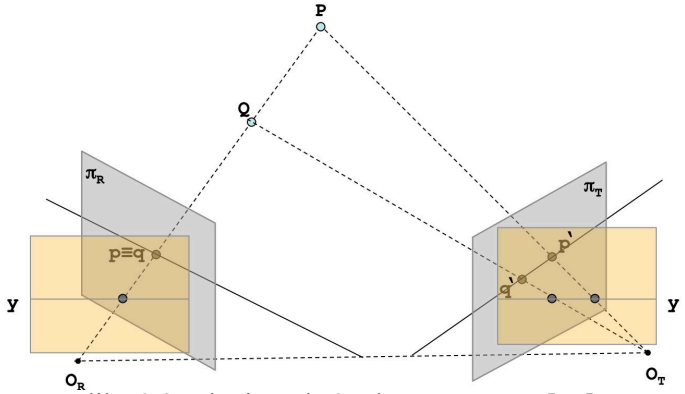
$$s\, m' = A[R|\mathrm{t}]M'$$

X,Y,Z – koridinate 3D točaka u kordinatnom sustavu svijeta
u,v – kordinate projiciranih točaka u pikselima
c_x, c_y – glavna točka, koja se uglavnom nalazi u centru slike
f_x, f_y – fokalne duljine izražene u pikselima

A – matrica intrističnih parametara

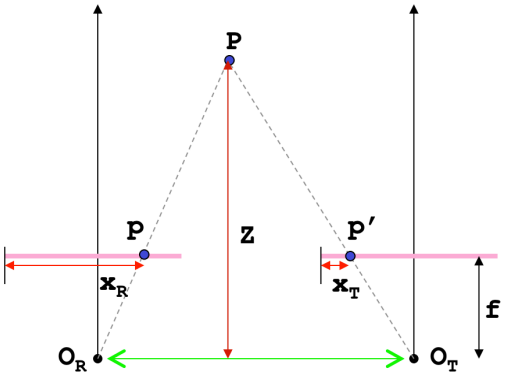
$[R|t]$ – matrica koja definira rotaciju i translaciju, matrica ektrističnih parametara definira gibanje kamere oko statične scene i obratno

Dobivanjem slike iz *pinhole* kamere odnosno pretvorbom iz 3D u 2D gubi se bitna informacija o dubini. Princip rada stereo kamere prikazan je na slici 3.2.



Slika 3.2 Princip rada 3D kamere. Izvor:[57]

Ovaj model se zasniva na epipolarnom ograničenju. Ako promatramo samo lijevu kameru O_R ne možemo odrediti 3D koordinate koje odgovaraju točkama Q i P zato što se sve točke koje se nalaze na pravcu Q_RQ odnosno Q_RP projiciraju u istu točku p≡q ravnine π_R. Ako uzmemo u obzir i desnu kameru O_T vidimo da se sada te iste točke projiciraju u točke q' i p' ravnine π_T. Vidimo dakle da se pomoću dvije ili više kamera može triangulacijom izračunati dubina, uz uvjet mogućnosti pronalaska odgovarajuće zajedničke točke u sceni u dvije slike kamere. Slika x prikazuje dijagram sličnosti trokuta (PQ_RO_T i Ppp') za standardnu konfiguraciju stereo kamere (kamere leže u zajedničkoj ravnini sa fiksnim međusobnim razmakom)

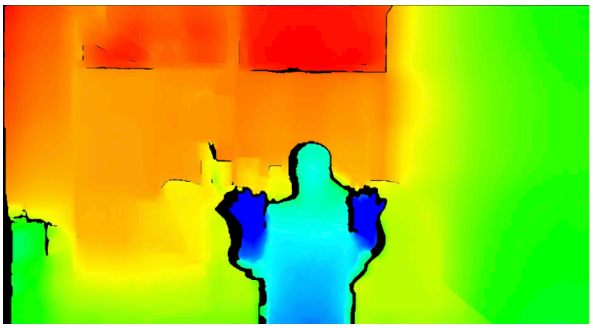


Slika 3.2 Detekcija dubine stereo kamere. Preuzeto iz [57]

Iz navedene slike se jednostavno može izračunati razlika x_R - x_T kao udaljenost između dvije točke na u ravnini slike koja odgovaraju 3D točki promatrane scene.

$$x_R - x_T = \frac{(\overline{O_R O_T}) * f}{Z} \qquad (3.2)$$

Iz gornjeg izraza slijedi da je razlika (eng. disparity) veća što je promatrana točka bliže kameri. U računalnom vidu se dubina prikazuje pomoću dubinske mape tako da se sustav diskretizira u paralelne ravnine, gdje svaka ravnina odgovara određenom intervalu dubine. Na slici 3.3 je prikazana dubinska mapa snimljena ZED kamerom. Algoritam za detekciju dubine je označio plavom bojom objekte bliže kameri a crvenom udaljene. I slike se jasno vidi kako ovakav postupak može dovesti do jednostavnog izdvajanje ruku.

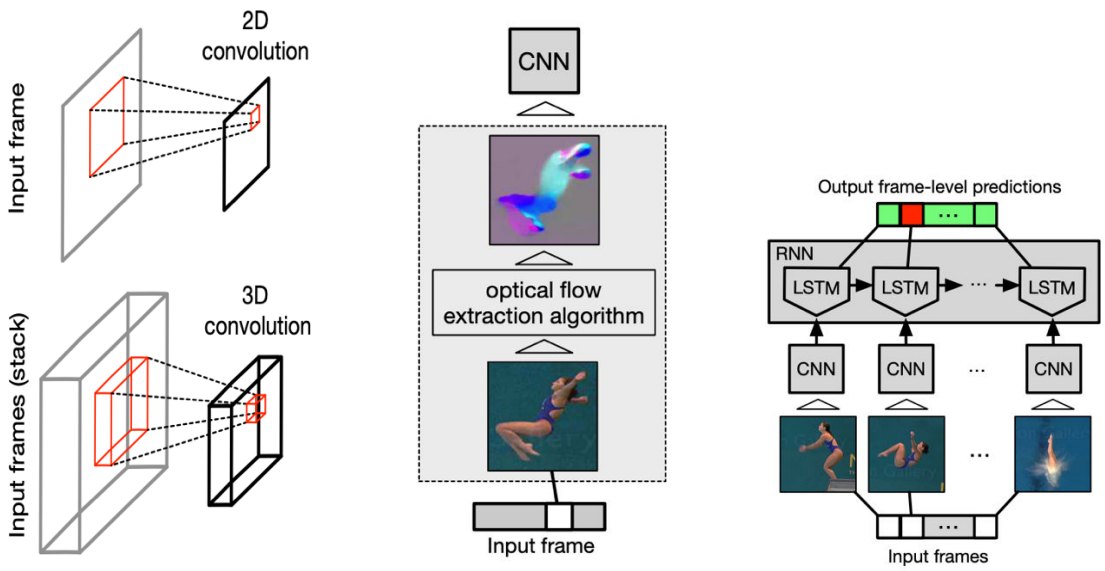


Slika 3.3 Prikaz dubinske mape snimljene ZED kamerom.

3.2. Arhitekture dubokog učenja kod prepoznavanje gesti

U ovom poglavlju ćemo će se iznijeti pregled područja iz prepoznavanja gesti temeljenog na dubokom učenju. Kao što je slučaj i kod većine zadataka računalnog vida poput detekcije objekta i prepoznavanja lica duboko učenje kod prepoznavanja gesti je postiglo izvrsne rezultate nadmašivši dotada najsuvremenije metode. Iako je primjena dubokog učenja na prepoznavanje gesta relativno nova postoji veliki broj istraživanja na ovu temu.

Najvažniji izazov kod sustava za prepoznavanja gesta temeljenih na dubokom učenju je modeliranje vremena. S obzirom na način pristupa vremenskoj dimenzija relevantna istraživanja možemo podijeliti u četiri skupine kao što je ilustrirano na slici 3.5.



Slika 3.4 Ilustrativni primjer različitih arhitektura kod prepoznavanja gesti. Preuzeto iz [58]

Prva skupina temelji se na 2D konvolucijskim neuralnim mrežama, koje u osnovi imaju samo mogućnost korištenja prostore informacije. Primjer ovakve metode za prepoznavanje sekvence slika (videa) je korištenje 2D CNN za pojedinačne slike i nakon toga se usrednjeni rezultat koristiti u klasifikaciji [9].

Jain et al. [10] je predložio CNN arhitekturu za procjenu tjelesne poze te je razvio prostorno-kontekstualni model koji predviđanje temelji na relativnom odnosu zglobova. U radu je korišteno više nezavisnih CNN mreža radi binarne klasifikacije pojedinih dijelova tijela. Ove mreže se koriste kao klizni prozori s prekapanjem nad ulaznim podacima što rezultira manjom mrežom i boljim performansama.

U svome radu, Li et al. [11], koristi CNN mrežu za izvlačenje značajki iz ulazne slike. Autori su predložili korištenje multi-tasking modela gdje se dobivene značajke koriste kod zadatka regresije zglobova i detekcije dijelova tijela. Kang et al. [12] je koristio CNN za izvlačenje značajki iz dubinske mape za prepoznavanje znakovnog jezika.

Druga skupina istraživačkih radova je orijentirana prema ulaznim značajka temeljenim na pokretu. Ove metode prvo izvlače značajke temeljen na pokretu poput protoka svijetlosti, a zatim se te značajke koriste kao ulaz u 2D CNN [14, 15, 16, 17, 18]. Ove metode uzimaju u obzir informaciju o vremenu preko unaprijed izračunatih značajki pokreta. U literaturi su oba pristupa u širokoj primjeni kod prepoznavanja gesti, međutim ako želimo postići bolje rezultate bolje je uključiti vremenske podatke nego prostorne.

Za prepoznavanje stila geste u biometrici, Wu et al [15] je predložio CNN koja radi sa vremenskim i prostornim informacijama. Autor je za dobivanje prostornih informacija koristio dubinsku mapu a za vremensku dimenziju optički tok.

Jain et al. [16] je u svome radu koristio boju i perspektivnu projekciju brzine 3D polja pokretnih površina kao značajke temeljene na pokretu. Korištena je arhitektura CNN mreže za procjenu

lokacije zglobova ljudskog tijela u video signalu, gdje su kao ulazni podaci u mrežu korištene RGB slike i značajke temeljene na pokretu.

Treća skupina temelji se na korištenju 3D filtera u konvolucijskom sloju. 3D konvolucija i 3D sažimanje omogućuju izvlačenje diskriminativnih značajki duž prostorne i vremenske dimenzije uz zadržavanje vremenske strukture što je u suprotnosti s 2D konvolucijskim slojevima. Prostorno-vremenske značajke ekstrahirane ovom vrstom modela su pokazale bolje performanse u odnosu na 2D modele trenirane na istim video uzorcima.
Nekoliko 3D CNN-ova je predloženo za prepoznavanje geste, od kojih su najpoznatiji Molchanov et al. [19]; Huang et al[20] i Molchanov et al. [21].

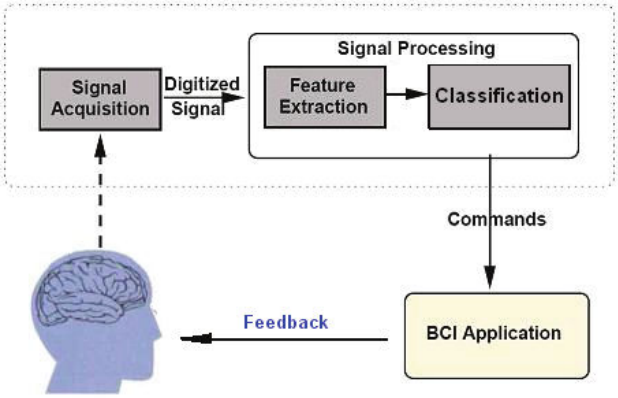
Autori u [21] predlažu 3D CNN za prepoznavanje geste vozačeve ruke iz podataka o dubini i intenzitetu. Autori kombiniraju informacije iz višestrukih prostornih ljestvica za konačno predviđanje. Također koristi se proširenje prostorno-vremenskih podataka za učinkovitije treniranje i smanjenje potencijalne pretreniranosti. Autori u [19] su proširili 3D CNN arhitekturu sa povratnim mehanizmom za detekciju i klasifikaciju dinamičkih gesti. Arhitektura se sastoji od 3D CNN-a koji se koristi za prostorno-vremensko izvlačenje značajki, povratnog sloja za globalno vremensko modeliranje i *softmax* sloja za predviđanje vjerojatnosti geste. 3D CNN za izvlačenje diskriminirajućih prostorno-vremenskih značajke iz čistog video signala za znakovni jezik je predloženo u [20]. Autori su radi poboljšanja performansi koristili više kanalni video (RGB-D i podaci o kosturu), , uključujući informacije o boji i dubini te poziciji zglobova tijela.

U četvrtoj skupini radova se koriste konvolucijske mreže na pojedinom okviru skupa sa modeliranjem vremenske sekvence. U literaturi se za modeliranje vremenske sekvence najčešće koriste povratne neuronske mreže sa LSTM ćelijama.

Neverova et al. [51] su predložili multimodalni (dubina, kostur i govor) sustav za prepoznavanje gesti temeljen na RNN. Svaki modalitet je prvo procesiran u kratkom prostorno-vremenskom bloku gdje se diskriminativne značajke specifična za podatke ručno izdvajaju ili uče.
RNN je korišten za modeliranje dugih vremenskih ovisnosti, fuziju podataka i na kraju za klasifikaciju geste. Slični pristup u detekciji dugih gesti korišten je kod [52]. Autori Wang et al. [53] predlažu korištenje sekvencijalno nadzirane LSTM (SS-LSTM), u kojoj se umjesto dodjeljivanja oznake klase izlaznom sloju RNN-a koristi pomoćno znanje u svakom vremenskom koraku za sekvencijalni nadzor.

4. Sučelje mozak-računalo

Sučelje računalo-mozak (eng. Brain computer interface - BCI), je hardverski i softverski komunikacijski sustav koji omogućuje ljudima interakciju sa okolinom, bez uporabe perifernih živaca i mišića, korištenjem kontrolnih signala generiranih iz elektroencefalografske aktivnosti. Ovi sustavi predstavljaju novi ne-mišićni način prenošenje namjera korisnika na vanjske uređaje kao što su računala, sintetizatori govora i neuronske proteze. BCI je sustav umjetne inteligencije koji ima sposobnost prepoznavanja određenih uzorka signala mozga kroz pet uzastopnih faza: prikupljanje, predobrada i pojačavanje signala, izdvajanje značajki te klasifikacija i kontrolno sučelje [59]. U postupku prikupljanja signala, obično se koriste postupci za smanjenje šuma i odbacivanje neželjenih artefakta. Faza pred-procesiranja priprema signale u prikladniji oblik za daljnju obradu, a faza izdvajanja značajki identificira diskriminirajuće informacije u snimljenim signalima. Moždani signali su pomiješani s drugim signalima koji dolaze iz konačnog skupa moždanih aktivnosti te se preklapaju u vremenu i prostoru. Također signali uglavnom nisu stacionirani i mogu biti iskrivljeni različitim artefaktima kao što su signali elektromiografije (EMG) ili elektrookulografije (EOG). U procesu klasifikacije se klasificira signala opisan vektorom značajki. Zbog toga je nužno pronaći diskriminantne značajke kako bi se postigao dobro prepoznavanje uzoraka a sukladno tome i dobro tumačenje korisnikove namjere. Na kraju razina upravljačkog sučelja prevodi klasificirane signale u smislene naredbe koje se dalje koriste za upravljanje uređajima, kao što su kolica ili računalo. Slika 4.1 prikazuje blok dijagram BCI sučelja.



Slika 4.1 Blok dijagram BCI sučelja. Izvor [60]

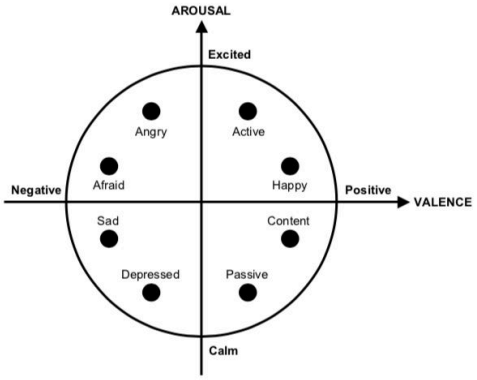
Povijesno gledajući BCI sučelja si bila neprivlačna za ozbiljna znanstvena istraživanja. Dizajn BCI-a smatran je previše složenim zbog ograničene rezolucije i pouzdanosti informacija koje se mogu otkriti u mozgu i kao i njegove velike varijabilnosti. Također BCI sustavi su zahtijevali obradu podatka u stvarnom vremenu i sve do nedavno potrebna tehnologija ili nije postojala ili je bila izuzetno skupa. Napretkom računalne tehnologije algoritmi dubokog učenja su dobili veliku pozornost u razvoj BCI aplikacija, te su pokazali odlične rezultate u rješavanje problema u nekoliko domena, posebno u medicinskim i robotskim poljima. Nedavno se pojavila nova perspektiva u korištenju BCI sučelja za prepoznavanje emocionalnog stanja korisnika.

4.1. Prepoznavanje emocija

Emocije imaju važnu ulogu u ljudskoj komunikaciji i procesu odlučivanja. Kako bi se zaobišla trenutna ograničenja kod interakcije s računalom potreban je pouzdaniji i prirodniji način komunikacije. Osposobljavanje računala da automatski prepozna i reagira na ljudske emocije je ključ uspjeha inteligentne komunikacije između čovjeka i računala. Automatsko prepoznavanje emocija je izazovan zadatak koji privlači interes različitih istraživačkih područja [61]. Radi ispravne reakcije na emocionalno stanje korisnika, računalo mora biti u stanju prikupiti informacije o trenutnoj situaciji te biti opremljeno mjerama za objašnjavanje i uočavanje emocije. U posljednjih nekoliko desetljeća predloženi su mnogi pristupi za procjenu ljudskih emocija. Konvencionalni pristupi usredotočeni su na analizu vizualnih i slušnih signala kao što su govor [62] i izraz lica [63]. Ovakvi pristupi imaju prednost jednostavnog prikupljanja podataka i proučavani su godinama, međutim njihova pouzdanost ne može biti zajamčena, lako lažirati izraz lica ili mijenjati ton glasa radi prekrivanja emocija tijekom socijalne komunikacije. Pouzdaniji način je korištenje fizioloških signala koji uključuju elektroencefalogram EEG, temperaturu T, elektrokardiogram (EKG), Elektromiogram (EMG), galvanski odgovor kože (GSR), disanje (RSP) itd. Stoga, fiziološki signali nude veliki potencijal za nepristrano prepoznavanje emocija. U daljnjem radu ograničit ćemo se na upotrebu EEG signala, odnosno na metode i algoritme za automatsko prepoznavanje emocija korištenjem BCI sučelja.

Riječi koje se koriste u opisivanju emocija su dvosmislene te ne postoji univerzalni način kvantificiranja emocija. Većina radova u području emocionalnog računanja (eng. affecive computing) je bazira na dva popularna modela emocija: kategorički i dimenzionalni

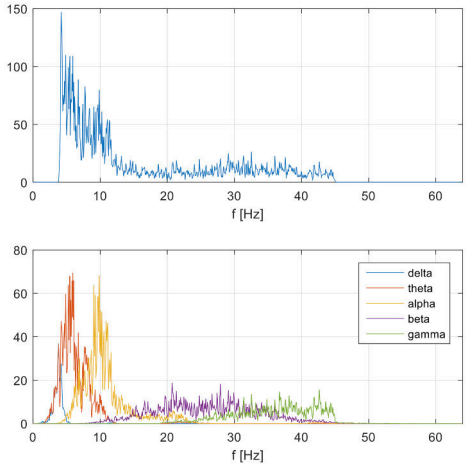
Kategorični model uključuje pronalazak i organizaciju određenog broja univerzalnih kategorija kao što je sreća, tuga, strah, bijes itd. Alternativa kategoričnom pristupu je dimenzionalni model emocija [64] kojeg karakteriziraju dvije emocionalne dimenzije pobuđenost (eng. arousal) i ugoda (eng. valence) kao što je prikazano na slici 4.2



Slika 4.2. Russellov 2-dimenzionalni model. Izvor:[65]

Emocionalno pobuđenost kreće se od mirnog do uzbuđenog, dok ugoda varira od negativne do pozitivne. EEG mjeri električni bio-potencijal koji stvaraju neuroni i neuronske veze u mozgu.

Kao što je već rečeno elektroencefalograf (EEG) je neinvazivna metoda mjerenja električke moždane aktivnosti. Broj EEG signala koji se koristi za detekciju emocija ovisi o broju kanala BCI-a i u literaturi se najčešće koristi 16 ili 32 kanala [BCI_5]. U EEG-u se mogu uočiti karakteristični signali određenih frekvencija: delta δ-valovi (0,5 – 4 Hz), theta θ-valovi (4 – 8 Hz), alfa α-valove (8 – 13 Hz), beta β-valove (13 – 30 Hz), gama γ-valovi (30 – 100 Hz). Kao što je prikazano na slici 3.3.



Slika 3.3. EEG signal u frekvencijskom području i njegova dekompozicija na specifične frekvencijske pojaseve. Izvor: [65]

Iako je korištenje EEG-a za detekciju emocionalnog stanja relativno je novo u usporedbi s audio-vizualnim metodama postoje brojna istraživanja u predviđanju emocija pomoću EEG signala. Iz relevantne literature je vidljivo da postoje dva glavna problema koja su zadržala značajan napredak u afektivnom računalstvu: problem treninga koji se očituje kao problem ručne anotacije podataka te problem inženjeringa pogodnih značajki. Kao i kod svakog problema strojnog učenja performanse najviše ovise o kvaliteti ulaznih podataka, odnosno značajki. Tradicionalne metode kod prepoznavanja emocija iz EEG signala bazirane na strojnom učenju zahtijevaju ručni pronalazak pogodnih značajki za određeni algoritam.

U svome radu Koelstra et al. [66] je predstavio javno dostupnu bazu za analizu emocija iz EEG signala. Autori su prestavili metodu za detekciju emocija gdje je kod klasifikacije korišten naivni Bayesov klasifikator. Predloženo je korištenje spektralne gustoća snage (PSD) i spektralne asimetrije snage (ASM) EEG signala kao značajki za algoritam strojnog učenja.

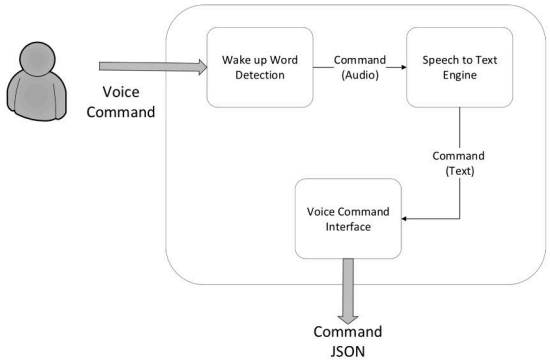
Autori Kraljević et al. [65] su predloži korištenje linearnih prediktivnih koeficijenta (LPC). Spektar EEG signala je prvo rastavljen na pet frekvencijski pojaseva a zatim je svaka spektralana envelopa predstavljena pomoću LPC koeficijentima koji su korištena kao vektor značajki na ulazu u SVM (eng. Support Vector Machine) je definiran.
Wichakam et al. [67] su kao značajke koristi PSD i srednju snagu po frekvencijskom pojasu skupa sa statističkim vrijednostima signala. Autori su u svome radu također eksperimentirali sa brojem kanala.

Navedeni radovi pokazali jako dobre rezultate, međutim kao i u slučaju računalnog vida i prepoznavanja govora, uvođenjem dubokog učenja ostvareni su rezultati koji nadmašuju konvencionalne metode strojnog učenja i ručnog inženjeringa značajki.

Li et al. [22] su u svome radu predstavili hibridni model dubokog učenja koji kombinira CNN i RNN. U predloženoj arhitekturi CNN služi za ekstraktiranje značajki iz skalograma dobivenog valičnom transformacijom EEG signala. Povratna neuralna mreža je korištena za cjelokupnu klasifikaciju emocije korisnika tokom trajanja proces, kao i za klasifikaciju emocija u zadanim vremenskim intervalima. Još jedan sustav sa hibridnim CNN-RNN modelom je predstavljem od Yang et al. [25]. U svome radu autori su predložili metodu pred-procesiranja koja dovodi do povećanja performansi sustava.
Yanagimoto et al. [26] je u svome radu koristio CNN za binarnu klasifikaciju emocionalne ugode. Kao ulaz u mrežu koristi su čisti EEG signal.
Autori Lin et al. [23] su u predloženom sustavu za prepoznavanje emocija koristili AlexNet arhitekturu. EEG signal je prvo rastavljen na šest frekvencijskih pojaseva a zatim je svaki pojas prikazan kao slika u sivim tonovima. Kao ulaz u CNN mrežu koristili su navedene slike u kombinaciji sa ručno izrađenim značajkama koje za prikaz drugi bioloških signala dostupnih u DEEP bazi.

5. Glasovno sučelje

Glasovno korisničko sučelje je tip sučelja koje omogućuje korisniku interakciju sa sustavom koristeći glasovne ili govorne komande. Glavna prednost ovakvog sučelja je u tome što omogućuje „hands-free“ interakciju. Glasovno sučelje je uspješno ugrađeno u automobile, pametne domove, kućne uređaje itd. Ono je primarni način interakcije s virtualnim asistentima na pametnim telefonima i pametnim zvučnicima. Blok dijagram glasovnog sučelja je prikazan na slici.



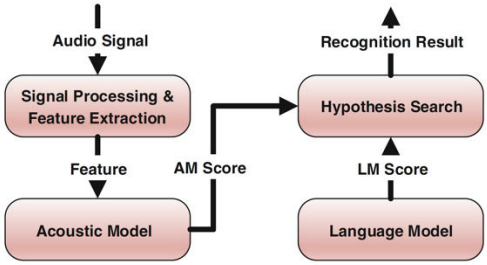
Slika 5.1. Blok dijagram glasovno sučelja. Izvor:[68]

Problematika kvalitetnog glasovnog sučelja uključuje prepoznavanje riječi za aktivacijom sustava (eng. Wake up word WUW), snimanje glasovne naredbe i prevođenje u tekst (Speech To Text -STT, parsiranje i analiza tekstualne komande u svrhu razumijevanja (Natural Language Understanding -NLU) i na kraju povratna informacija korisniku.
Jedna od ključnih komponenti glasovnog korisničkih sučelja je automatsko prepoznavanje govora (ASR) koje omogućuje prevođenje govora u tekst.

5.1. Automatsko prepoznavanje govora

Automatsko prepoznavanje govora (ASR) smatra se poveznicom za bolju i prirodniju komunikaciju između čovjeka i računala. U prošlosti je tehnološka ograničenja limitirala su korištenje govora kao primarnog sredstva za interakciju, pogotovo zato što je u mnogim situacijama alternativna komunikacija kao što su miš i tipkovnica značajno nadmašila govornu učinkovitost i točnost. Tijekom posljednjih nekoliko godina ova razlika u preciznosti se počela smanjivati zbog tehnološkog napretka. Računalna snaga koja je danas dostupna je nekoliko redova veća zbog upotrebe grafičkih procesnih jedinica opće namjene (GPGPU) i CPU / GPU klastera, što nam omogućuje treniranje snažnijih, ali složenih modela. Također, postoji mnogo više dostupnih podataka nego prije. Izgradnjom modela na velikim podacima prikupljenim iz stvarnih scenarija, možemo učiniti ASR sustave robusnijima a time i pogodnima za svakodnevnu uporabu. Standardna arhitektura sustava za automatsko prepoznavanje govora prikazana na slici 5.2 sastoji se od četiri glavne komponente: obrada signala i izvlačenje

značajki, akustični model (AM), jezični model (LM), model za generiranje vjerojatnosti prepoznate govorne sekvence.



Slika 5.2. Arhitektura ASR sustava Izvor:[69]

Prva faza odgovorna je za pred-obradu audio signala što uključuje metode za uklanjanje šuma i izobličenja. Također, u ovom koraku govorni signal se obično pretvara iz vremenske domene u frekvencijsku domenu gdje audio signal opisujemo značajkama prikladnim za ulaz u akustični model. Akustički model koristi znanje o akustici i fonetici te generira ocjenu za svaku ulaznu sekvencu značajki varijabilne duljine. Uloga jezičnog modela je na izlazu dati ocjenu kao vjerojatnost pretpostavljenog niza riječi, učeći korelaciju između riječi iz tekstualnih korpusa. I na kraju, komponenta za pretraživanje hipoteza kombinira AM i LM ocijene i kao rezultat prepoznavanja na izlazu daje sekvencu riječi s najvišom ocjenom.

Ovisno o tipu prepoznavanja govornog signala ASR sustavi se mogu podijeliti u različite kategorije: izolirane riječi, kontinuirani govor i spontani govor. Kod prepoznavanje izoliranih riječi sustav očekuje jednu izgovorenu riječ i obično zahtijeva od govornika da kod svakog izgovaranja postoji tišina na obje strane uzorka. Za prepoznavanje višestrukih riječi sustav zasebno pokreće prepoznavanje izolirane riječi te zahtjeva malo vremena između izgovorenih riječi. Sustav za prepoznavanja kontinuiranog govora mora imati mogućnost obrade prirodnog govora korisnika, što iziskuje posebne metode implementacije, dok kod prepoznavanja spontanog govora sustav mora imati mogućnost rukovati s različitim prirodnim značajkama govora kao što su riječi koje se izgovaraju zajedno, pogrešni izgovori i ne-riječi.
Svaki govornik ima jedinstvena svojstva koja utječu na glas. Na temelju tih svojstava sustav je podijeljen u dvije glavne klase: sustavi ovisni i neovisni o govorniku. ASR modeli koji ovise o govorniku, generalno su jeftiniji i lakši za implementirati. Daju dobre rezultate za samo jednog određenog govornika dok je rezultat prepoznavanja govora za ostale govornike jako loš. Kod sustava neovisnih o govorniku često je moguće ASR model do trenirati glasom korisnika radi dobivanja što kvalitetnijih rezultata iako ovi sustavi generalno rade ne ovisno o govorniku. Složenost sustava kao i njegova točnost ovise o veličini vokabulara s kojim sustav raspolaže. Neke aplikacije zahtijevaju prepoznavanje nekoliko riječi, dok druge zahtijevaju veliki vokabular. Ovisno o tipu vokabulara sustavi za automatsko prepoznavanje govora mogu se klasificirati kao sustavi s malim, srednjim i velikim vokabularom.
U prošlosti, konvencionalni ASR sustavi su koristili standardne tehnike za izvlačenje govornih značajki poput MFCC i LPC koeficijenata koji su se koristili kao ulaz u akustični model.

Akustični model je dizajniran da može na svome ulazu primiti vektore značajki varijabilne duljine i da može ne smetano raditi sa varijabilnosti audio signala, uzrokovanom kompliciranom interakcijom govornikovih karakteristika (npr. spol, bolest ili stres), stilom i brzinom govora, okolnom bukom i popratnim razgovorom, izobličenjem kanala (npr. razlika u mikrofonu) te sa razlikama u dijalektu.

Dugo vremena automatsko prepoznavanje govora temeljilo se na „plitkim“ arhitekturama baziranim na skrivenim Markovljevim modelima (HMMs) gdje je svako stanje karakterizirano sa Gaussian mixture modelom (GMM). Daljnji tehnološki napredak u prepoznavanju govora je iziskivao izradu bolji i kompleksnijih akustičnih značajki pogodnih za varijacije GMM-HMM modela. Iako su se ovi pristupi pokazali uspješnima, te se još uvijek koriste postoji nekoliko nedostataka kao što su zahtijevanje značajne količine specifičnog znanja o zadatku, treniranje različitih modula ASR sustava se obavlja odvojeno što često iziskuje dodatnu prilagodbu parametara.
Rješenje ovih izazova leži u dubljim arhitekturama koje mogu barem funkcionalno oponašati ljudski sustav govora za kojeg je poznato da ima dinamičku i hijerarhijsku strukturu kao u proizvodnji tako i u percepciji govora. U ovome radu ćemo dati kratak pregled dubokih arhitekture koja se koristi u realizaciji akustičnog modela u trenutno najsuvremenijim ASR sustavima, čime se uklanja potreba za GMM-HMM pristupom.

5.2. Arhitekture dubokog učenja za automatsko prepoznavanje govora

Pojavom dubokog učenja drastično se poboljšala točnost prepoznavanja ASR sustava. Takvi sustavi postepeno zamjenjuju tradicionalne sustave, a osim povećanja točnosti i bolje robusnosti prednost je i to što mogu biti trenirani kao „end-to-end“ sustavi. Ovi end-to-end sustavi sve više dobivaju na popularnosti zbog pojednostavljenog procesa izgradnje modela i sposobnosti izravnog mapiranja govora u tekst bez unaprijed definiranih poravnavanja. U literaturi postoje dvije najvažnije vrste end-to-end arhitektura za ASR: RNN u kombinaciji sa connectionist temporal classification i CNN.
Connectionist Temporal Classification (CTC)[70] je objektivna funkcija koja omogućava treniranje RNN za transkripciju sekvenci bez potrebe za prethodnim usklađivanjem ulaznih i ciljnih sekvenci. Ideja je prikazati izlaz iz mreže kao distribuciju vjerojatnost svih mogućih sekvenci uvjetovanih ulaznom sekvencom. Izlazni sloj mreže sadrži po jednu jedinicu za svaku transkripcijsku oznaku (slovo ,fonem, nota) i jednu dodatnu jedinicu koja označava prazninu.Za zadanu sekvencu transkripcije, postoji onoliko mogućih poravnanja koliko postoji različiti načini razdvajanja s prazninama.

Graves et al. [38] je u svome radu koristio BRNN-LSTM mrežu za klasifikaciju fonema na TIMIT bazi. Pokazano je da dvosmjerna povratna neuralna mreža nadmašuje standardnu RNN i višeslojni perceptron u smislu brzine i preciznosti. Autor je također u svome radu [39] pokazao

da BRNN u realizaciji akustičnog modela kod tradicionalnog ASR sustava sa HMM može uvesti poboljšanje u odnosu na DNN i GMM. Graves et al. [40] je također predložio korištenje RNN za realizaciju end-to-end sustava za automatsko prepoznavanje govora te je prezentirao jako dobre rezultate bez upotrebe jezičnog modela. Sak et al. [44] je predložio ASR sistem temeljen na RNN te je prezentirao najnovija dostignuća kod akustičnog modeliranja za velike riječnike. U svome radu Weninger et al. [43] je prezentirao metodu poboljšanje govornih značajki temeljenu na dubokim dvosmjernim RNN mrežama.
Osim RNN koji se danas pojavljuje u većini relevantne literature, postoje i istraživanja u području automatskog prepoznavanja govora temeljena na konvoluciskim neuralni mrežama. Kod većine istraživanja koristi se svojstvo CNN mreža za pronalaskom odgovarajućih akustičnih značajki pa je tako Palaz et al. [27] u svome radu pokazao da je moguće istrenirati end-to-end model za klasifikaciju sekvenci fonema koristeći sirovi signal. Abdel-Hamid et al. [28] je u svome radu pokazao prednosti CNN kod izvlačenja kvalitetnih akustičkih značajki. U ovom radu kao ulazni podaci korišten je spektrogram koji se sastoja o od MFSC (mel-frequency spectral coefficients) koeficijenata i njihovih delta i delta-delta promjena. Predstavljeni rezultati pokazuju da uporaba CNN smanjuje pogrešku kod prepoznavanja fonema te glasovne pretrage po velikom rječniku od 6%-10% u usporedbi sa DNN na TIMIT bazi.

Automatsko prepoznavanje govora dostiglo je visoku razinu performansi, ali obično ne uspijeva u stvarnom, bučnom okruženju a kao jedan od bitnih razloga je neusklađenost uvjeta u kojima sustav je treniran i u kojima se koristi. Iako se uvođenjem duboko učenja ukazalo na veću robusnost ASR sustava kroz treniranje modela na velikim podacima koji pokrivaju mnoge akustične uvijete i dalje su potrebne metode pred-procesiranja signala da bi se ostvarilo pouzdano udaljeno prepoznavanje govora.

5.3. Udaljeno prepoznavanje govora

Udaljeno prepoznavanja govora (eng. Far-field speech) je neophodna tehnologija za govornu interakciju u stvarnim uvjetima koja za cilj ima omogućiti pametnim uređajima da prepoznaju udaljeni ljudski govor (1-10m). Sustav za prepoznavanje dalekog govora obično se sastoji od „front-end „ modula za obradu signala i „back-end“ modula za prepoznavanje govora. Front-end modul za cilj ima filtrirati govor od buke i jeke, koristeći različite tehnika poboljšanja govora kao što su dereverberation i beamformin. Back-end modul je sličan običnom sustavu za prepoznavanje govora i ima za cilj prepoznati i pretvoriti očišćeni govor u tekst.

U mnogim aplikacijama postoji potreba za razdvajanjem višestrukih izvora zvuka ili izvlačenje interesantnog izvora, pritom minimizirajući ostale neželjene ometajuće signale i buku. Izdvojeni signali mogu se zatim izravno slušati ili dalje obrađivati [71].
Znanstveno je dokazano da ljudi mogu jasno razaznati samo jedan razgovor u jako bučnom okruženju, kao što je koktel zabava. Unatoč tome što ovaj dobro poznati problem predmet

istraživanja već desetljećima, problem koktel zabave još uvijek zahtjeva daljnje istraživanje [72]. Kao što je naglašeno u nekim radovima [73], korištenjem jednog kanala nije moguće poboljšati razumljivost i kvaliteta signala u isto vrijeme. Način savladavanja ovog ograničenja jest dodavanje prostorne informacije dostupnim vremenskim/frekvencijskim informacijama. Ovo se postiže korištenjem dva ili više govornih kanala odnosno pomoću dva ili više mikrofona. Postoje dvije kategorije više-kanalnih algoritama: Blind Source Separation (BSS) i Beamforming.

Najpoznatija klasa metoda za obradu signala iz niza mikrofona je formiranje snopa. Ove metode nastoje kombinirati signale primljene na pojedinim mikrofonima na takav način da se pojačava signal koji dolazi iz određenog smjera, dok se signali koji dolaze iz drugih smjerova prigušuju. Dvije često korištene metode beamforminga su Delay-and-sum (DS) i Minimum variance distortionless response (MVDR). DS je jednostavna i izravna metoda za izvođenje beamforminga. Koristi činjenicu da mikrofoni smješteni na različitim prostornim pozicijama primaju isti signal izvora ali s različitim kašnjenjima. Štoviše, kašnjenja ovise o smjeru iz kojeg dolazi izvorni signal. Ako znamo kašnjenja koja odgovaraju željenom smjeru, možemo ih koristiti za pomak signala. Operacija pomicanja signala poravnava željeni signal u svim kanalima, dok signali koji dolaze iz različitih smjerova ostaju neusklađeni. Tada možemo jednostavno usredniti sve takve pomaknute signale koji će uzrokovati prigušenje signala iz nepoželjnih smjerova. Da bi koristili ovu metodu, moramo znati pojedinačna kašnjenja izvora signala za svaki mikrofon. To može biti ili unaprijed poznato (iz arhitekture niza mikrofona i položaja izvora) ili, u češćem slučaju, moramo procijeniti kašnjenja primljenih signala. Najpopularnije tehnike za procjenu vremenskog kašnjenja temelje se na unakrsnoj korelaciji. Međusobna korelacija odražava razinu sličnosti dvaju signala kao funkciju njihovog relativnog kašnjenja, te je za očekivati da kašnjenje koje stvara maksimum ove funkcije pripada stvarnom kašnjenju pri kojem je govorni signal stigao do mikrofona.

Unatoč svojoj jednostavnosti, DS je vrlo čest izbor za realizaciju tzv. *beamforming-a*. Glavni nedostatak DS je u tome što ocjenjuje svoje parametre uzimajući u obzir samo poziciju željenog izvora, a ne pozicije interferirajućih zvukova. Druga metoda *beamforming-a* MVDR je uspješnija u rješavanju ovog problem nastojeći eksplicitno minimizirati učinak buke

Da bi se postigla maksimalna redukcija buke, koristi se procjena matrice kovarijance šuma koja predstavlja koliko su signali buke međusobno povezani između mikrofona. Poznavanje ovih korelacija daje informacije o smjerovima izvora buke i omogućuje *beamforming-u* da potisne signale koji dolaze iz tih smjerova.

Govorni signal snimljen udaljenim mikrofonom ne sadrži samo aditivnu buku, nego i refleksije istog signala iz zidova i drugih objekata - efekt poznat kao odjek. Ova vrsta izobličenja ima vrlo različita svojstva, što uzrokuje neuspjeh mnogih konvencionalnih metoda za smanjenje buke. Kao posljedica toga, potrebno je koristiti posebne metode za redukciju odjeka u ulaznom signalu. Jedna od takvih metoda je Weighted prediction error (WPE). WPE je pokazala

uspješnost u nedavnoj literaturi [74, 75] gdje je pokazano da može učinkovito iskoristiti na više-kanalne signale i da može se lako povezati s *beamforming* tehnikama.

Pokazalo se također da se postizanja robusnosti ASR sustava može ostvariti uz modeliranje ljudskog slušnog sustava. Tako je Russo et al. [76, 77] u svojim radovima pokazao da modeliranjem auditornog mehanizma u „front-end"-u ASR sustava dovodi do značajnog poboljšanja performansi uz prisutnost buke. Autori su u [76] predložili novu vrstu auditornih značajki dobivenu slijedom filtriranja audio signala pomoću Gammatone filterbank i procesiranja pomoću Inner Hair cell (IHC) modela. Pokazano je poboljšanje performansi u bučnim uvjetima u odnosu na standardne MFCC koeficijente. U [77] autori su predložili metodu rekonstrukcije signala bazilarne membrane. Ova metoda se može koristiti u fazi pred-procesiranja signala što rezultira poboljšanjem kvalitete signala. Također autori su predložili korištenje novih koeficijenata za prepoznavanje govora temeljenih na odazivu bazilarne membrane.

6. Zaključak

Razvoj senzora i područja senzorike omogućili su razvoj interaktivnih inteligentnih uređaja. Integracija navedenih uređaja omogućuje razvoj novih aplikacija u širokom rasponu domena kao što su, među ostalim, pametni domovi, e-zdravlje i inteligentni transportni sustavi itd.
Ove okolnosti pridonijele su situaciji pojave i razvoja pametne okoline koje iziskuju razradu novih koncepata i pristupa u interakciji između čovjeka i pametnih okolina te iziskuju izlazak iz današnjih standardnih oblika sučelja koja moraju nuditi naprednije, intuitivnije i prirodnije načine interakcije.

Analiza područja međuljudske interakcije naglasila je nužnost za istraživačkim područjima usmjerenim na interakciju čovjeka i računala (HCI) kroz prirodne i intuitivne modalitete sa naglaskom na prepoznavanje gesta ruku/tijela, analizu bio-signala i korištenje elemenata afektivnog računarstva te prepoznavanje govora i srodne problematike.

U ovom radu dan je pregled područja naprednih sučelja za interakciju čovjeka i računala temeljenih na metodama i konceptima dubokog učenja. Objašnjeni su temeljni principi te su obrađene dominante arhitekture neuralnih mreža koje se u literaturi najčešće koriste za realizaciju prirodnog korisničkog sučelja. Objašnjeni su koncepti potrebni za realizaciju sučelja za upravljanje gestama, sučelja mozak-računalo te sučelja za upravljanje govorom. Ukazalo se na probleme, prepreke i izazove rada ovih sučelja u stvarnim uvjetima.
Pregledom relevantne literature, usmjerene na arhitekture dubokog učenja, pokazalo se da se uvođenjem ovih metoda ostvario značajan napredak u promatranim znanstvenim područjima. Dostupnost velike količine potrebnih podataka osiguralo je bolju robusnost naprednih sučelja u realnim uvjetima no trenutni pokazatelji kvalitete rada ali i mogućnosti, ukazuju da je nužan nastavak i daljnji istraživački napor kako bi performanse ovakvih umjetnih sustava postigle ljudsku razinu komunikacije i interakcije realnim uvjetima. Bez obzira na detaljne specifičnosti pojedinih rješenja i prijedloga, za zaključiti je da će istraživanja u ovim poljima i dalje biti inspirirana i motivirana ljudskim sustavom.

LITERATURA

[1] https://www.datacamp.com

[2] https://www.medium.com

[3] Peng, M., Wang, C., Chen, T. and Liu, G., 2016. NIRFaceNet: A convolutional neural network for near-infrared face identification. *Information*, *7*(4), p.61.

[4] https://www.superdatascience.com

[5] https://towardsdatascience.com

[6] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105)

[7] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[8] Ajmal, H., Rehman, S., Farooq, U., Ain, Q.U., Riaz, F. and Hassan, A., 2018, April. Convolutional neural network based image segmentation: a review. In *Pattern Recognition and Tracking XXIX* (Vol. 10649, p. 106490N). International Society for Optics and Photonics.

[9] Wang, P., Li, W., Liu, S., Gao, Z., Tang, C. and Ogunbona, P., 2016, December. Large-scale isolated gesture recognition using convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 7-12). IEEE.

[10] Jain, M., van Gemert, J. and Snoek, C.G., 2014. University of amsterdam at thumos challenge 2014. *ECCV THUMOS Challenge*, *2014*.

[11] Li, S., Liu, Z.Q. and Chan, A.B., 2014. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 482-489).

[12] Kang, B., Tripathi, S. and Nguyen, T.Q., 2015, November. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 136-140). IEEE.

[13] John, V., Boyali, A., Mita, S., Imanishi, M. and Sanma, N., 2016, November. Deep learning-based fast hand gesture recognition using representative frames. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1-8). IEEE.

[14] Simonyan, K. and Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (pp. 568-576).

[15] Wu, J., Ishwar, P. and Konrad, J., 2016. Two-stream cnns for gesture-based verification and identification: Learning user style. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 42-50).

[16] Jain, A., Tompson, J., LeCun, Y. and Bregler, C., 2014, November. Modeep: A deep learning framework using motion features for human pose estimation. In *Asian conference on computer vision* (pp. 302-315). Springer, Cham.

[17] Sun, L., Jia, K., Yeung, D.Y. and Shi, B.E., 2015. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4597-4605).

[18] Weinzaepfel, P., Harchaoui, Z. and Schmid, C., 2015. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 3164-3172).

[19] Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S. and Kautz, J., 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4207-4215).

[20] Huang, J., Zhou, W., Li, H. and Li, W., 2015, June. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)* (pp. 1-6). IEEE.

[21] Molchanov, P., Gupta, S., Kim, K. and Kautz, J., 2015. Hand gesture recognition with 3D convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1-7).

[22] Li, X., Song, D., Zhang, P., Yu, G., Hou, Y. and Hu, B., 2016, December. Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 352-359). IEEE.

[23] Lin, W., Li, C. and Sun, S., 2017, September. Deep convolutional neural network for emotion recognition using EEG and peripheral physiological signal. In *International Conference on Image and Graphics* (pp. 385-394). Springer, Cham.
[24] Miranda-Correa, J.A. and Patras, I., 2018, May. A Multi-Task Cascaded Network for Prediction of Affect, Personality, Mood and Social Context Using EEG Signals. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 373-380). IEEE
[25] Yang, Y., Wu, Q., Qiu, M., Wang, Y. and Chen, X., 2018, July. Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.
[26] Yanagimoto, M. and Sugimoto, C., 2016, November. Recognition of persisting emotional valence from EEG using convolutional neural networks. In *2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA)* (pp. 27-32). IEEE.
[27] Palaz, D., Collobert, R. and Doss, M.M., 2013. End-to-end phoneme sequence recognition using convolutional neural networks. *arXiv preprint arXiv:1312.2137*.
[28] Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G. and Yu, D., 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, *22*(10), pp.1533-1545.
[29] Parcollet, T., Zhang, Y., Morchid, M., Trabelsi, C., Linarès, G., De Mori, R. and Bengio, Y., 2018. Quaternion convolutional neural networks for end-to-end automatic speech recognition. *arXiv preprint arXiv:1806.07789*.
[30] Abdel-Hamid, O., Mohamed, A.R., Jiang, H. and Penn, G., 2012, March. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)* (pp. 4277-4280). IEEE.
[31] Palaz, D., Doss, M.M. and Collobert, R., 2015, April. Convolutional neural networks-based continuous speech recognition using raw speech signal. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4295-4299). IEEE.
[32] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C.L.Y. and Courville, A., 2017. Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*.
[33] Qian, Y., Bi, M., Tan, T. and Yu, K., 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(12), pp.2263-2276.
[34] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W.J., Espi, M., Higuchi, T. and Araki, S., 2015, December. The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*(pp. 436-443). IEEE.
[35] Lekshmi.K.R, Sherly, E. Automatic Speech Recognition using different Neural Network Architectures – A Survey, (*IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 7 (6) , 2016, 2422-2427
[36] Kawakami, K., 2008. *Supervised sequence labelling with recurrent neural networks* (Doctoral dissertation, Ph. D. thesis, Technical University of Munich).
[37] Schuster, M. and Paliwal, K.K., 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), pp.2673-2681
[38] Graves, A. and Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5-6), pp.602-610.
[39] Graves, A., Jaitly, N. and Mohamed, A.R., 2013, December. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding* (pp. 273-278). IEEE.
[40] Graves, A. and Jaitly, N., 2014, January. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning* (pp. 1764-1772).
[41] Graves, A., 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.
[42] Graves, A., Mohamed, A.R. and Hinton, G., 2013, May. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). IEEE.
[43] Weninger, F., Geiger, J., Wöllmer, M., Schuller, B. and Rigoll, G., 2014. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. *Computer Speech & Language*, *28*(4), pp.888-902.

[44] Sak, H., Senior, A. and Beaufays, F., 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*.
[45] Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F. and Chandran, V., 2018. Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework. *IEEE Access*, *6*, pp.49325-49338.
[46] Alhagry, S., Fahmy, A.A. and El-Khoribi, R.A., 2017. Emotion recognition based on EEG using LSTM recurrent neural network. *Emotion*, *8*(10).
[47] Li, Z., Tian, X., Shu, L., Xu, X. and Hu, B., 2017, August. Emotion recognition from eeg using rasm and lstm. In *International Conference on Internet Multimedia Computing and Service* (pp. 310-318). Springer, Singapore.
[48] Pigou, L., Van Den Oord, A., Dieleman, S., Van Herreweghe, M. and Dambre, J., 2018. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, *126*(2-4), pp.430-439.
[49] Du, Y., Wang, W. and Wang, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1110-1118).
[50] Veeriah, V., Zhuang, N. and Qi, G.J., 2015. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 4041-4049).
[51] Neverova, N., Wolf, C., Taylor, G.W. and Nebout, F., 2014, September. Multi-scale deep learning for gesture detection and localization. In *Workshop at the European conference on computer vision* (pp. 474-490). Springer, Cham.
[52] Chai, X., Liu, Z., Yin, F., Liu, Z. and Chen, X., 2016, December. Two streams recurrent neural networks for large-scale continuous gesture recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 31-36). IEEE.
[53] Wang, P., Song, Q., Han, H. and Cheng, J., 2016. Sequentially supervised long short-term memory for gesture recognition. *Cognitive Computation*, *8*(5), pp.982-991.
[54] Hogan, K., 2003. *Can't get through: eight barriers to communication*. Pelican Publishing.
[55] Kelly, S.D., Manning, S.M. and Rodak, S., 2008. Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, *2*(4), pp.569-588.
[56] Kendon, A., 1986. The biological foundations of gestures: Motor and semiotic aspects. Current Issues in the Study of Gesture.
[57] https://docs.opencv.org/
[58] Asadi-Aghbolaghi, M., Clapés, A., Bellantonio, M., Escalante, H.J., Ponce-López, V., Baró, X., Guyon, I., Kasaei, S. and Escalera, S., 2017. Deep learning for action and gesture recognition in image sequences: A survey. In *Gesture Recognition* (pp. 539-578). Springer, Cham.
[59] Khalid, M.B., Rao, N.I., Rizwan-i-Haque, I., Munir, S. and Tahir, F., 2009, February. Towards a brain computer interface using wavelet transform with averaged and time segmented adapted wavelets. In *2009 2nd International Conference on Computer, Control and Communication* (pp. 1-4). IEEE.
[60] Kołodziej, M., Majkowski, A. and Rak, R., 2010. Matlab FE_Toolbox-an universal utility for feature extraction of EEG signals for BCI realization. *Przegląd Elektrotechniczny*, *1*.
[61] Scherer, K.R., Bänziger, T. and Roesch, E. eds., 2010. *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press.
[62] Mao, Q., Dong, M., Huang, Z. and Zhan, Y., 2014. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, *16*(8), pp.2203-2213.
[63] Zhang, Y.D., Yang, Z.J., Lu, H.M., Zhou, X.X., Phillips, P., Liu, Q.M. and Wang, S.H., 2016. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access*, *4*, pp.8375-8385.
[64] Russell, J.A., 1980. A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), p.1161.
[65] Kraljević, L., Russo, M. and Sikora, M., 2017, August. Emotion classification using linear predictive features on wavelet-decomposed EEG data. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 653-657). IEEE.
[66] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I., 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, *3*(1), pp.18-31.

[67] Wichakam, I. and Vateekul, P., 2014, May. An evaluation of feature extraction in EEG-based emotion prediction with support vector machines. In *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 106-110). IEEE.
[68] Kraljevic, L., Russo, M. and Stella, M., 2018. Voice command module for Smart Home Automation. *International Journal of Signal Processing*, *3*.
[69] Yu, D. and Deng, L., 2016. *AUTOMATIC SPEECH RECOGNITION*. Springer london limited.
[70] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J., 2006, June. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376). ACM.
[71] Brandstein, M. and Ward, D. eds., 2013. *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media.
[72] Haykin, S. and Chen, Z., 2005. The cocktail party problem. *Neural computation*, *17*(9), pp.1875-1902.
[73] Loizou, P.C., 2007. *Speech enhancement: theory and practice*. CRC press.
[74] Kinoshita, K., Delcroix, M., Gannot, S., Habets, E.A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B. and Sehr, A., 2016. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), p.7.
[75] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W.J., Espi, M., Higuchi, T. and Araki, S., 2015, December. The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*(pp. 436-443). IEEE.
[76] Russo, M., Stella, M., Sikora, M. and Pekić, V., 2019. Robust Cochlear-Model-Based Speech Recognition. *Computers*, *8*(1), p.5.
[77] Russo, M., Stella, M., Sikora, M. and Šarić, M., 2019. Cochlea-inspired speech recognition interface. *Medical & Biological Engineering & Computing, pp.1-11*