SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 3083

METODE ZA PROCJENU LJUDSKE POZE

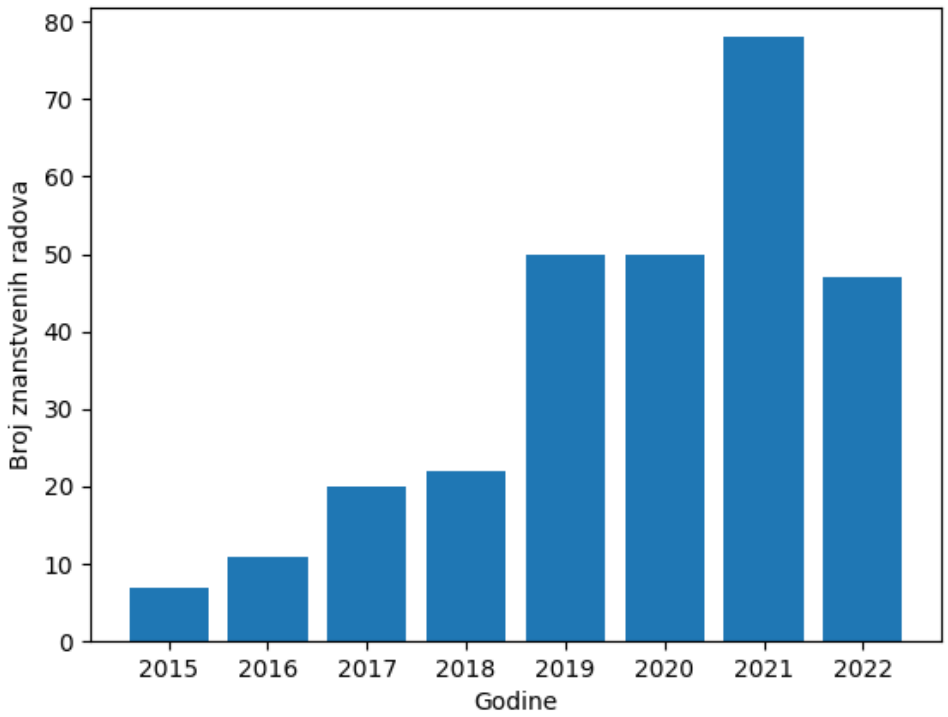
Danijel Popović

Zagreb, veljača 2023.

SADRŽAJ

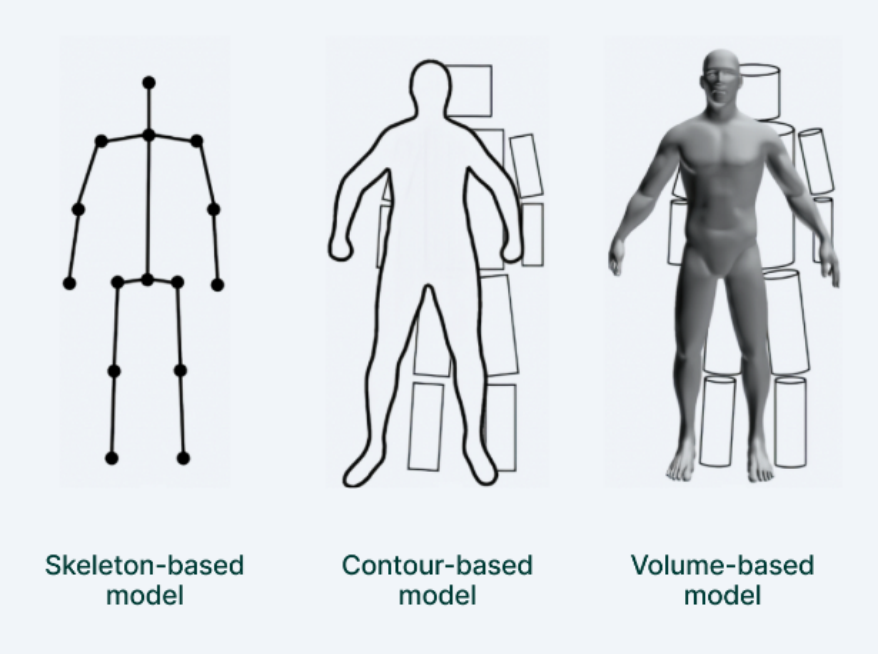
1. Uvod

U današnjem svijetu praktična implementacija sustava koji koriste računalni vid postaje svakodnevica. Samovozeći automobili, *Snapchat*/*Instagram* filteri, prepoznavanje lica (engl. *Face Recognition*) i sl. su aplikacije koje su trenutno aktualne i s vremenom samo postaju bolje. Kao jedan od glavnih problema u području računalnog vida uz klasifikaciju, detekciju i praćenje objekata jest procjena ljudskih poza u 2D/3D. Jako zahtjevan zadatak u području računalnog vida je veoma interesantan zbog svoje primjene u industriji videoigara, augmentiranoj stvarnosti, zdravstvu i sportu. Rast u popularnosti rješavanja ovog problema je vidljiv na slici 1.1.



Slika 1.1: Broj radova tijekom godina, [25]

Zadatak procjene jest određivanja skupa koordinata koje predstavljaju dijelove tijela



Slika 1.2: Vrste reprezentacije ljudske poze, [59]

(engl. *keypoints*) (glava, rame, itd.) koje će služiti kao reprezentacija ljudske poze. Postoje 3 tipa reprezentacije ljudskog tijela (slika 1.2):

1. reprezentacija koristeći kostur tijela (engl. *skeleton-based model*)
2. reprezentacija koristeći konture tijela (engl. *contour-based model*)
3. volumetrijska reprezentacija (engl. *volume-based model*)

Ponuđena rješenja se mogu podijeliti u dvije velike skupine:

1. Standardni (klasični) pristup
2. Pristupi dubokog učenja

U radu su opisani spomenuti pristupi u poglavljima Standardni pristup i Pristup dubokog učenja. U tim poglavljima napravljena je još jedna podjela:

1. Procjena poze u 2D
2. Procjena poze u 3D

Za svaku podjelu opisane su vrste mreža koje su korištene u većini radova (konvolucijske i graf duboke neuronske mreže,te mehanizmi pozornosti). Također je napravljena dodatna podjela po načinu rješavanja problema procjene poze (monokularan vs pristup

s više pogleda za procjenu poze u 3D, te pristup s jednom etapom (engl. *stage*) vs s više etapa za procjenu poze u 2D). Opisani su jedni od trenutno najboljih radova. Potom su se usporedili rezultati standardnog pristupa i pristupa dubokog učenja. Na samom kraju opisan je implementacijski postupak metoda dubokih učenja, te usporedba provedenih eksperimenata sa službenim rezultatima.

2. Standardni pristup

2.1. Procjena poze u 2D

U posljednjih nekoliko desetljeća mnogo je truda bilo uloženo u kreiranju robusnih modela za procjenu poza u kontroliranim i nekontroliranim uvjetima. Tipične metode uključuju *pictorial*, hijerarhijske i nestablaste modele.

Standardni pristup procjene poze je model s *pictorial* strukturom koji prostorne korelacije između dijelova tijela prikazuje kao grafičke modele stablaste strukture s *kinematic priors* koji povezuju povezane udove. Ovi modeli rade dobro kada su svi udovi vidljivi na slici ali skloni su pogreškama poput *double-counting* koje se dese kada odnosi između varijabli nisi obuhvaćeni modelom stablaste strukture. Primjeri radova koji implementiraju ovaj pristup su [41] i [42].

Hijerarhijski modeli, poput [44] i [45], prikazuju odnose između dijelova na različitim skalama (engl. *scales*) i veličinama u hijerarhijskoj strukturi stabla. Ideja ovog pristupa jest da veći dijelovi su oni koje je lakše naći u slici i mogu pomoći u detekciji manjih dijelova.

Nestablasta struktura, korištena u [46] i [47], inkorporira interakcije koje uvode petlje kako bi promijenile stablastu strukturu s dodatnim bridovima koje hvataju simetriju i dalekometne odnose. Ovakvi modeli trebaju napraviti kompromis između preciznog modeliranja prostornih odnosa i modeliranja koje dozvoljava efikasno vrijeme zaključivanja.

Sekvencijske predikcije uče implicitni prostorni model s potencijalno kompleksnom interakcijom između varijabli. Popularan rad ovog pristupa jest Pose Machine [48].

Iako navedene metode daju obećavajuće rezultate za procjenu poze unutar kontroliranih uvjeta, inače degradiraju značajno u nekontroliranim uvjetima zbog kompleksnih varijacija poza, iluminacija, itd.

2.2. Procjena poze u 3D

Percepcija dubine iz 2D je klasičan problem koji je privlačio pozornost mnogih znanstvenika i umjetnika još od Renesanse, kada je Brunelleschi koristio matematički koncept perspektive da bi prenio smisao prostora u svojim slikama Florentinskih zgrada. Stoljećima kasnije, slični znakovi perspektive su korišteni u računalnom vidu za zaključivanje duljine, površine i omjera daljina proizvoljnih scena [49]. Osim informacije o perspektivi, klasični sustavi računalnog vida su pokušali koristiti druge značajke poput sjenčenja [50] ili teksture [51] kako bi dobili dubinu. Moderni sustavi tipično rješavaju problem služeći se nadziranim učenjem kako bi zaključiti diskriminirajuće značajke slike za procjenu dubine.

Jedan od prvih algoritama za procjenu dubine je koristio različit pristup [52], odnosno iskorištavao je poznate 3D strukture objekata u sceni.

Zaključivanje 3D dijelova tijela iz njihovih 2D projekcija se može primijetiti u radu [54]. Tu je pokazano da ako imamo duljinu kosti problem svodi na binarno stablo odluke gdje svaka podjela korespondira u 2 moguća stanja za dio tijela (engl. *joint*) u odnosu na njegovog roditelja. Ovo binarno stablo se može orezati (engl. *prune*) s obzirom na ograničenja dijelova tijela (engl. *joint constraints*) iako rijetko rezultira jedinstvenim rješenjem.

Rad [55] je koristio veliki skup poza za rješavanje višeznačnosti ovisno o upitima najbližih susjeda. Ideju iskorištavanja najbližih susjeda za poboljšanje rezultata je koristio pristup iz rada [56] koji inkorporira vremenska ograničenja tijekom pretraživanja. Drugi načini skupljanja znanja o 3D ljudskoj pozi iz skupa podataka su kreirajući *overcomplete bases* koje mogu reprezentirati ljudske poze kao rijetke kombinacije (engl. *sparse combinations*) (korišteno u radovima [57], [58]), podizanjem poze na Hilbertov prostor reproduciranih jezgri (engl. *reproducible kernel Hilbert space*) [43] ili kreiranjem novih *priors* iz specijaliziranih skupova podataka koji se sastoje ekstremnih ljudskih poza [57].

3. Pristup dubokog učenja

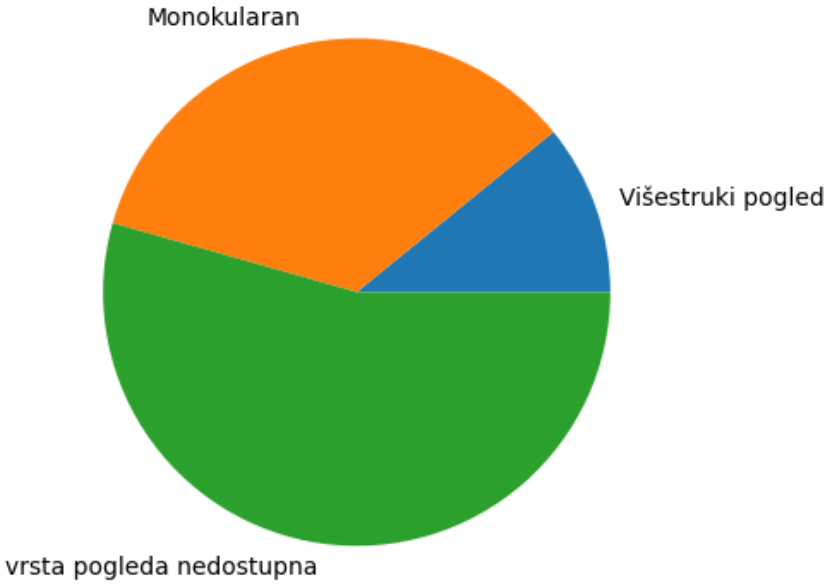
U posljednje vrijeme više radova je fokusirano na rješavanje problema procjene ljudske poze u 2D/3D prostoru koristeći metode dubokog učenja. Metode s najboljim MPJPE (engl. *mean per joint position error*) ,odnosno PCKh (engl. *Percentage of Correct Keypoints*) rezultatima na skupu podataka Human3.6M [40], odnosno MPII [39] koriste duboko učenje kako bi dobili točne procjene. MPJPE je Euklidska udaljenost između točnih i predviđenih koordinata poza. PCKh je modifikacija PCK metrike koja mjeri jesu li predviđene i točne koordinate dijelova tijela unutar određenog praga udaljenosti. Za PCKh@0.5 taj prag je 50% duljine koja spaja glavu i torzo (engl. *head bone link*).

Pionir preokreta u pristupu procjene (iz standardnog u pristup dubokog učenja), je rad *DeepPose* [28].

Postoje međutim dva distinktna pristupa rješavanja problema procjene poze u 3D prostoru. To su pristupi s jednim pogledom (engl. *single view*) i s više pogleda (engl. *multi-view*). Pristup jednog pogleda (monokularan pristup), kako ime sugerira, se služi samo jednim pogledom, odnosno slikom/videozapisom, snimljenog jednom kamerom kako bi metoda generirala procjenu. Dok pristup s više pogleda zahtjeva više slika/videozapisa iz različitih uglova (kamera) da bi dobio točnu ljudsku pozu. Logično možemo zaključiti da pristupi s više pogleda imaju globalno bolje rezultate. Taj zaključak potkrjepljuje i činjenica da najbolje metode za Human3.6M [40] i MPI-INF-3DHP skupove koriste ovaj pristup.

Unatoč tome, iz slike 3.1 možemo primijetiti da je značajno više radova obavljeno rješavajući problem procjene koristeći monokularan pristup, dok pristup s više pogleda je u prošlosti pretežito služilo kao potpora za monokularne, dajući točne vrijednosti poza (engl. *ground-truth*) koje su služile prilikom treniranja modela [27].

U narednim poglavljima biti će opisani radovi za određivanje poze u 2D i 3D, te jedne od trenutno najboljih metoda za procjenu 2D odnosno 3D poze.



Slika 3.1: Podjela radova po vrsti pogleda, [25]

3.1. Procjena ljudske poze u 2D

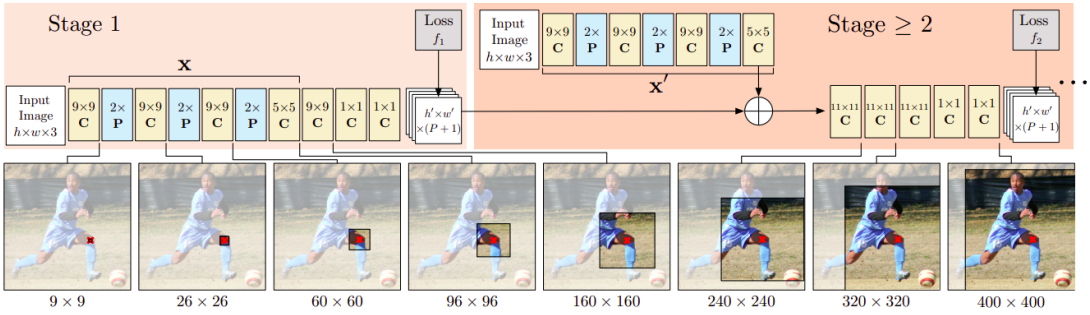
3.1.1. Konvolucijske mreže u 2D procjeni

Rastom popularnosti konvolucijskih mreža i njihovom uspješnošću u izvlačenju značajki slika, radovi kreću ih sve više primjenjivati. Danas su temeljni dio skoro svih metoda za procjenu ljudske poze u 2D.

Rad [29] umjesto direktne procjene poze, koristi kombinaciju konvolucijskih mreža i Markovljevog slučajnog polja. Konvolucijska mreža, nazvana *Part-Detector*, kombinira reprezentacije značajki različitih rezolucija s preklapajućim receptivnim poljima te kao izlaz daje toplinske karte (engl. *heatmaps*), odnosno vjerojatnosti da pikseli predstavljaju određeni dio tijela.

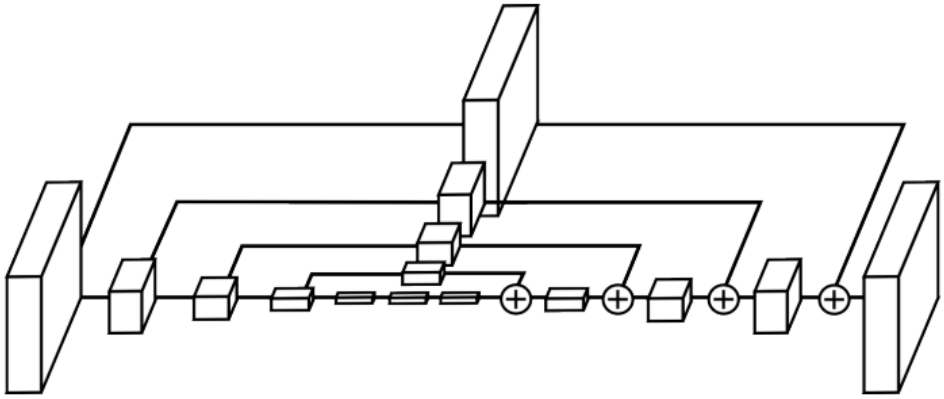
Convolutional Pose Machines [13], odnosno *CPM* je jedan od radova koji koristi sekvencijalni pristup procjene. CPM se sastoji od niza konvolucijskih mreža koje više puta proizvode 2D *belif maps* za lokacije svakog dijela poze (Slika 3.2). Značajke slike i *belif maps* prijašnje etape se koriste kao ulaz za sljedeću etapu. Prva etapa (rozi blok imenovan *Stage 1*) koristi samo lokalnu informaciju unutar slike kako bi kreirala *belief*

maps. Kažemo da je informacija lokalna jer receptivno polje je ograničeno na mali dio oko izlaznog piksela. Struktura mreže prve etape se sastoji od 5 konvolucijskih slojeva popraćenih s dva konvolucijska sloja s jezgrom veličine 1×1.
Korištenjem izlaza prijašnjih etapa i značajki slike (blok *Stage* ≥ 2 na slici 3.2), naknadne etape omogućuju klasifikatoru slobodnu kombinaciju dostupnih kontekstualnih informacija. Zbog prijašnjih informacija klasifikatori odabiru one značajke koje im omogućuju bolju predikciju. U ovim etapama receptivno polje postaje drastično veliko kako bi se prepoznali dalekometni odnosi između 2D polja lokacija (povećanje receptivnog polja je ostvareno nizanjem konvolucijskih slojeva, što je vidljivo na slici 3.2). Arhitektura je *end-to-end* što znači da se sve može zajedno trenirati koristeći algoritam propagacije unatrag (engl. *backpropagation*).



Slika 3.2: Pregled *CMP* metode, [13]

TransPose [30] koristi kombinaciju konvolucijskih mreža i *transformer* mreže za dobivanje boljih rezultata procjena ljudske poze. Konvolucijske mreže su neizostavni dio *backbone* arhitekture (više o ovom modelu će biti opisano u poglavlju Pozornost u 2D procjeni).
Za popularne *Hourglass* [31] mreže konvolucija je nužna. Struktura ove mreže je prikazana slikom 3.3. Konvolucijski slojevi i slojevi sažimanja po maksimalnoj vrijednosti su korišteni za procesiranje značajki. Procesiranje značajki se obavlja sve dok rezolucije značajki ne postanu jako niske. Kvadrati koji se nalaze na slici 3.3 su rezidualni moduli (engl. *residual module*) koji obavljaju procesiranje. Prije prolaska kroz sloj sažimanja, mreža se grana i primjenjuje više operacija konvolucije na *pre-pooled* značajke. Nakon što izlazne značajke poprime najniže moguće rezolucije, mreža koristi operacije naduzorkovanja i kombinacija značajki.
Unatoč njihovoj korisnosti postoje mnogi nedostatci konvolucijskih mreža. Uspješnost konvolucijskih mreža je usko povezana s receptivnim poljem, koje se može povećati produbljivanjem mreže ili dodavanja više parametara konvolucijskim slojevima. Prva opcija dovodi do pojave nestajućih gradijenata dok druga zahtjeva korištenja velike



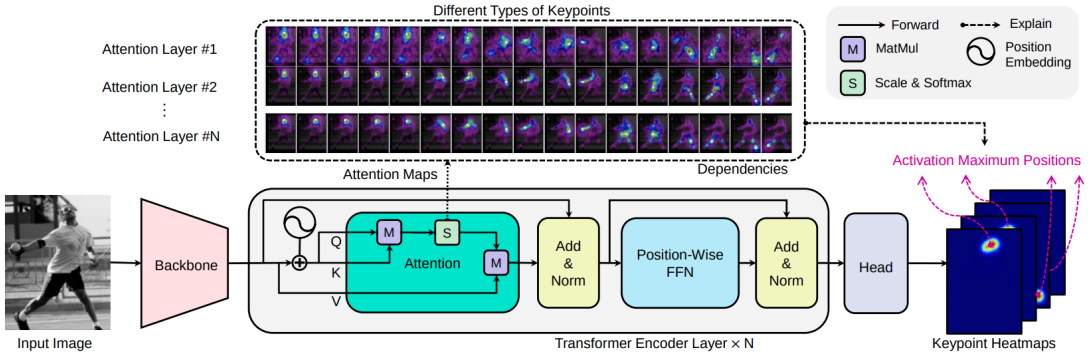
Slika 3.3: Pregled *Hourglass* arhitekture, [31]

računalne moći. Kako bi se zaobišli ovi problemi radovi implementiraju razne arhitekture i kombinacije konvolucijskih mreža s drugim mehanizmima kako bi postigli što točnije rezultate procjene poza.

3.1.2. Pozornost u 2D procjeni

Vizualna pozornost je jedan od osnovnih mehanizama koji ljudi svakodnevno koriste, pa ima smisla pokušati taj mehanizam implementirati za rješavanje problema računalnog vida. Mehanizmi pozornosti su sve popularniji alati korišteni za poboljšanje procjena ljudske poze na slikama. Dosta metoda postiže točne rezultate kombinirajući ih s konvolucijskim mrežama, dok rad *ViTPose* [38] pokušava iskoristiti apsolutnu moć vizualnih *transformators*, odnosno vizualne pozornosti, postižući time najbolje rezultate na MPII skupu podataka [39]. Detaljnije o njemu će biti opisano u poglavlju Trenutno najbolji rad za procjenu 2D poze. Dok konvolucijski modeli su napravili nevjerojatan napredak u procjeni ljudske poze, koje prostorne odnose smatraju relevantnim za procjenu dijelova tijela ostaje misterija.
Rad [30] predlaže model nazvan *TransPose*, koji koristi *transformer* arhitekturu za procjenu ljudske poze. Dodatan cilj *TransPose* modela jest pokušaj povećanja interpretabilnosti rezultata. Pozornosni slojevi ugrađeni u transformer omogućuju modelu da razumije dalekometne odnose između dijelova tijela (engl. *keypoints*) efikasno i također mogu otkriti koji dijelovi tijela naviše utječu na procijenjeni dio tijela.
Slika 3.4 daje pregled strukture mreže gdje možemo primjetiti da se mreža sastoji od 3 dijela:

1. *Backbone*, rozi blok



Slika 3.4: Pregled TransPose arhitekture

2. *Transformer Encoder Layer*, sivi blok

3. *Head*, svijetlo ljubičasti blok

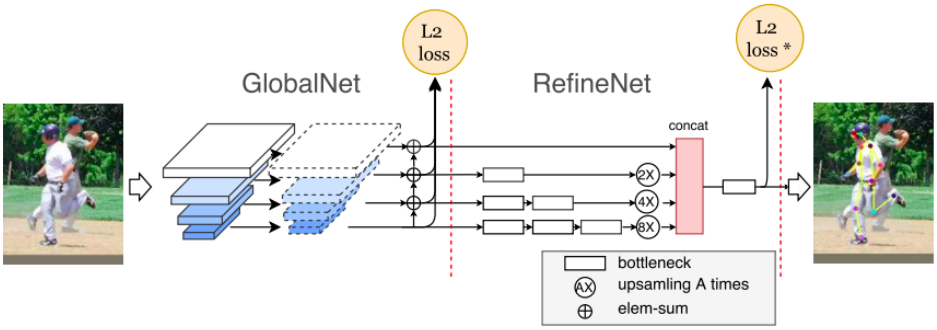
Backbone dio koristi konvolucijske mreže kako bi dobio značajke slike. Model je treniran na dvije različite arhitekture: *ResNet* i *HRNet*.
Transformer Encoder Layer na ulazu dobiva značajke slike koje formalno zapisujemo $X_f \epsilon R^{d \times H \times W}$. Ulazna značajka se potom spljošti (engl. *flatten*) te dobivamo $X \epsilon R^{L \times d}$, gdje $L = H \times W$. Potom X prolazi kroz N pozornosnih i potpuno povezanih slojeva (ovo je zapravo *Multi-Head Attention* koji je objašnjen u poglavlju Pozornost u 3D procjeni).
Head dio je zadužen za pretvorbu izlaza *Transformer Encoder Layer* dijela u toplinske karte dijelova tijela (na slici 3.5 su označene *Keypoint Heatmaps*).

3.1.3. Pristupi s jednom etapom vs pristupi s više etapa

Još jedna podjela trenutnih metoda procjene ljudske poze je na dvije skupine: pristupe s jednom i više etapa (engl. *stage*).

3.1.3.1. Pristup s jednom etapom

Mnogo pristupa s jednom etapom se koncentriraju na dizajniranju strukturno jednostavne mreže. *Hourglass* [31] je jedna od najosnovnijih primjera ovog pristupa. Mnogo istraživača je fokusirano na unapređivanju *hourglass* pristupa. *Cascaded Pyramid Network*, odnosno *CPN*, [17] uzima prednosti metoda *hourglass* i *Feature pyramid networks*, odnosno *FPN*, [34] što dovodi do boljih rezultata procjene poze. Za rješavanje procjene poza korištene su dvije podmreže *GlobalNet* i *RefineNet* (vidljivo na slici



Slika 3.5: Pregled *CPN* arhitekture

3.5).

GlobalNet koristi *ResNet* arhitekturu kao *backbone* i može efikasno locirati dijelove tijela, poput očiju, ali ne uspijeva točno odrediti lokacije kukova. Razlog tome jest što za određivanje lokacija kukova nisu dovoljne samo lokalne informacije.

Kako bi se detektirali dijelovi tijela poput kukova, odnosno *hard keypoints*, izlazi *GlobalNet* mreže se predaju *RefineNet* mreži. *RefineNet* poboljšava procjenu tako što prenosi informaciju kroz različite razine. Na kraju integrira informacije različitih razina koristeći operacije naduzorkovanja (engl. *upsamling*) i spajanja (engl. *concatenate*).

3.1.3.2. Pristup s više etapa

Ovi pristupi se fokusiraju na izgradnju više etapa za poboljšanje performanci. *CPM* [13], prvi predstavlja pristup s više etapa koristeći nekoliko konvolucijskih i slojeva sažimanja (engl. *pooling*). *Stacked Hourglass Network* [31] se sastoji od naslaganih *hourglass* mreža.

Različit pristup od *hourglass* metoda ima OpenPose [8] koji predviđa dijelove koristeći *bottom-up* pristup. Prvo detektira dijelove tijela te ih onda povezuje. Ovaj algoritam može raditi u realnom vremenu (engl. *real-time*) zbog toga je jako popularan u praktičnoj primjeni.

Generalno arhitekture s više etapa rade bolje od arhitektura s jednom etapom. Također dizajn svake etape je jako bitan, što se može primijetiti kod *CPM*, gdje prva etapa ima drugačiju ulogu od ostalih.

3.1.4. Trenutno najbolji rad za procjenu poze u 2D

Trenutno najbolji rad za procjenu 2D poze je *ViTPose* model, uzimajući PCKh@0.5 kao metriku. Sortirana lista radova po PCKh@0.5 metrici se nalazi na [25].

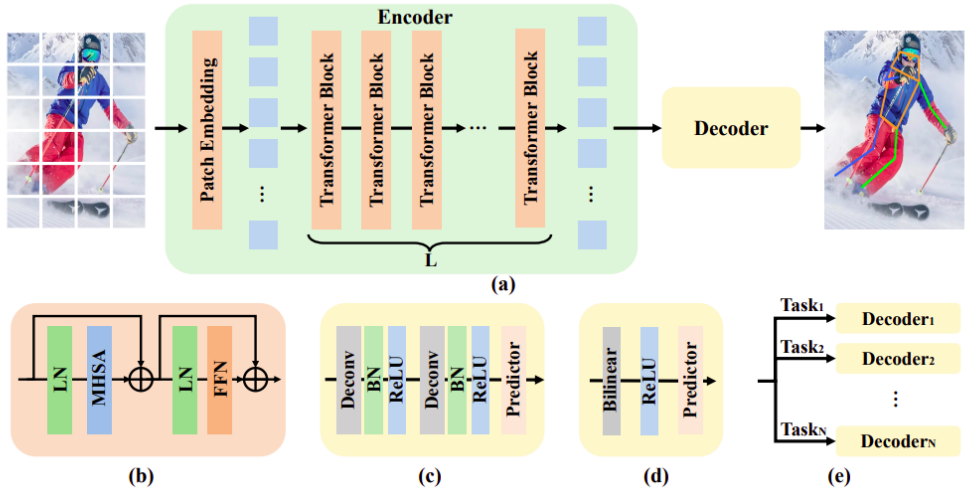
3.1.4.1. ViTPose

ViTPose [38] je rad koji je specifičan zbog korištenja samo vizualnog transformera za procjenu poze. ViTPose koristi jednostavan nehijerarhijski vizualni transformer kao *backbone* dio mreže za izvlačenje značajki za danu instancu osobe i dekoder s malim brojem parametara (engl. *lightweight decoder*) za procjenu poze (slika 3.6 (a)). Kao ulaz, model prima sliku koju je potrebno pretvoriti u *tokens* koristeći *patch embeddings*, odnosno pretvaramo ulaznu sliku u višedimenzionalnu značajku koju nazivamo *embedding tokens*. Potom te značajke obrađuje nekoliko *multi-head self-attention* i potpuno povezanih slojeva.

Testirane su dvije vrste dekodera za procesiranje značajki dobivenih iz enkoder dijela. Prvi dekoder jest standardni dekoder (slika 3.6 (c)) koji se sastoji od dva *deconvolution* bloka. *Deconvolution* blok u sebi sadrži *deconvolution* sloj, *batch* normalizaciju te *ReLU* aktivacijsku funkciju.

Drugi dekoder (slika 3.6 (d)) je jednostavni dekoder. Sastoji se od bilinearne interpolacije, *ReLU* funkcije i konvolucijskog sloja.

Slika 3.6 (e) pokazuje fleksibilnost modela u obavljanju različitih vrsta zadataka. Na izlaz enkodera možemo postaviti više dekoder blokova gdje će svaki biti zadužen za određivanje poze za određenu vrstu skupa poza (gotovo svaki skup podataka koji sadrži ljudske poze ima različite načine označavanja poza).



Slika 3.6: (a) pristup *ViTPose* modela, (b) *Transformer* blok, (c) Standardni dekoder, (d) Jednostavni dekoder, (e) Dekoder za više klasa, [38]

3.2. Procjena ljudske poze u 3D

3.2.1. Monokularan pristup

3.2.1.1. Pregled radova koji koriste monokularni pristup

Kao što je i ranije naglašeno postoje određene razlike u pristupima rješavanja problema procjene 3D ljudske poze. U ovom poglavlju nude se jedne od mogućih podjela na osnovu pristupa dobivanja procjene i tipu arhitekture.

3.2.1.1.1. 2D u 3D vs 3D direktna procjena

Generalno gledajući postoje dva pristupa rješavanja problema 3D procjene:

- direktna procjena u 3D prostor iz 2D slike
- dobivanje 2D koordinata dijelova tijela i podizanje u 3D prostor.

Direktna procjena ne radi procjenu 2D poze kao međukorak, već pokušava procijeniti 3D pozu ljudskog tijela na osnovu samo ulazne slike. Drugi pristup kao međukorak procjenjuje 2D lokacije dijelova tijela te potom projektira ih u 3D prostor. Projekcija 2D koordinata je jako zahtjevan proces zbog neodređenosti dubine što omogućuje da se 2D poze mogu prikazati na beskonačno načina u 3D prostoru. Direktna procjena ipak ima problem što generalno ne uspijeva iskoristiti sve informacije koje su na raspolaganju i to rezultira u lošijim rezultatima u odnosu na projekciju 2D koordinata. U nastavku je dan kratki pregled radova koji koriste navedene metode.

Direkna procjena u 3D prostor iz 2D slike

Rad [11] je uočio kako pristupi koji su bazirani na toplinskim kartama, imaju problem kreiranja diferencijalnih metoda koje bi učinile pristupe *end-to-end*. Uz to izlaz toplinskih karti je manje rezolucije u odnosu na ulaznu sliku, te to uzrokuje neizbježne kvantizacijske pogreške (engl. *quantization errors*). Zbog toga razloga rad predlaže određivanje 3D koordinata dijelova tijela jednadžbom (3.1).

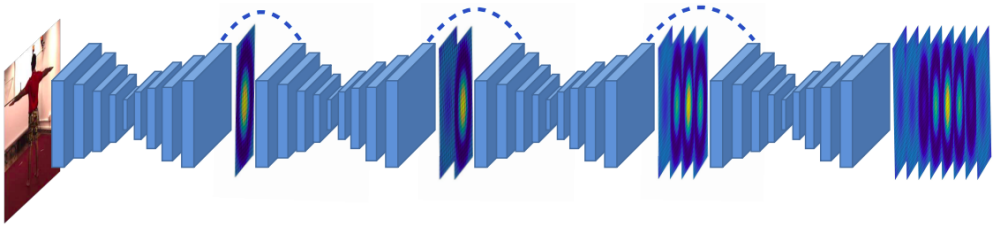
$$\int_{p\epsilon\Omega} p\tilde{H}_k(p) \tag{3.1}$$

gdje Ω je integralna domena a, $\tilde{H}_k(p)$ je normalizirana toplinska karta. Ova metoda se može koristiti kod postojećih rješenja koja koriste toplinske karte i time kombinirati prednosti direktne procjene i toplinskih karti.

U radu [12] smatraju da je reprezentacija 3D ljudskih poza kritični problem pristupa baziranih na konvolucijskim mrežama te predlažu pristup koji ima dva doprinosa. Prvi

je diskretizacija 3D prostora oko subjekta, odnosno osobe, i treniranje konvolucijske mreže da predvidi vjerojatnosti voksela za svaki dio tijela. Odnosno rad koristi volumetrijsku reprezentaciju poze u 3D prostoru. Drugi doprinos je korištenje *coarse-to-fine* postupka za procjenu. Slika 3.7 prikazuje postupak procjene poze. Primjećujemo da je ulaz u mrežu samo jedna RGB slika. Mreža se sastoji od više potpunih konvolucijskih komponenti (engl. *fully convolutional components*). 3D toplinske karte su izlazi komponenti. Isprekidane linije predstavljaju proces spajanja značajki slika i dobivenih toplinskih karti, te je novonastala vrijednost predana na ulaz sljedeće komponente.

Rad *Self-Attention Network for Human Pose Estimation* [1] naglašava manu konvolucijskih pristupa u neiskorištavanju posebnih odnosa poput simetrije tijela. *METRO* [5] i *Mesh Graphormer* [2] direktnom procjenom rješavaju problem procjene ljudske poze i problem rekonstrukcije ljudskog oblika (engl. *human mesh reconstruction*). Kako sva tri rada koriste mehanizme pozornosti više o njima će biti u poglavlju Pozornost u 3D procjeni.



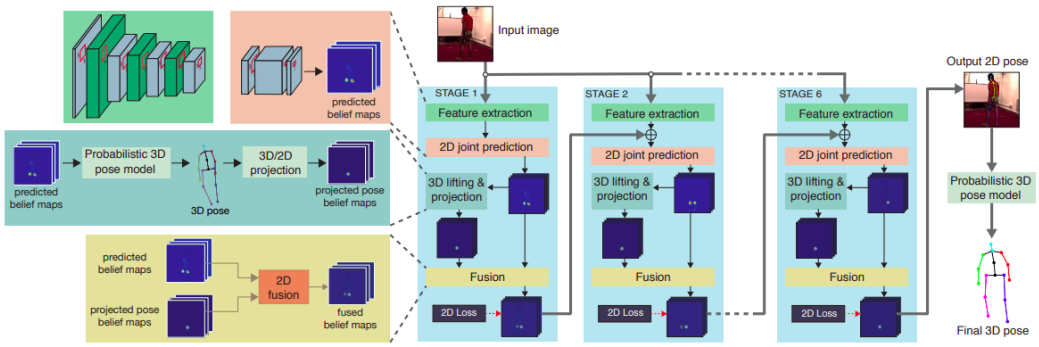
Slika 3.7: Postupak direktne procjene koristeći volumetrijsku reprezentaciju, [12]

Podizanje iz 2D u 3D prostor

Radovi koji koriste ove pristupe dijele proces procjene poze u 2 podzadatka: procjena 2D lokacija dijelova tijela i podizanje iz 2D u 3D prostor. Ovisno o radu, problem pokušavaju riješiti rješavajući 2. ili 1. podzadatak.

Radovi poput *Lifting from the Deep* [6] rješavaju oba podzadatka tako da prvi podzadatak je zapravo varijacija *Convolutional Pose Machine* arhitekture [13] koja se isključivo bavi procjenom 2D poza. Slika 3.8 daje pregled ovog pristupa. Ulaz metode jest jedna RGB slika dok kao izlaz dobivamo 3D pozu. Kao što je ranije navedeno ova metoda koristi pristup s više etapa gdje svaka etapa kreira *belief maps* za lokacije 2D polja. Svaka etapa kao ulaz dobiva ulaznu sliku i *belief maps* prethodne etape. Zadatak svake etape jest naučiti kombinirati *belief maps* i projektirane *belief maps* poza. Projektirane *belief maps* poza dobivaju se koristeći vjerojatnosni 3D model poza. Ove dvije *belief*

map vrijednosti su potom spojene i predane kao izlaz trenutne etape. Točnost 2D i 3D polja lokacija se progresivno povećava u svakoj etapi. Čitava arhitektura je diferencijabilna te se može trenirati *end-to-end* koristeći algoritam propagacije unatrag. Metode



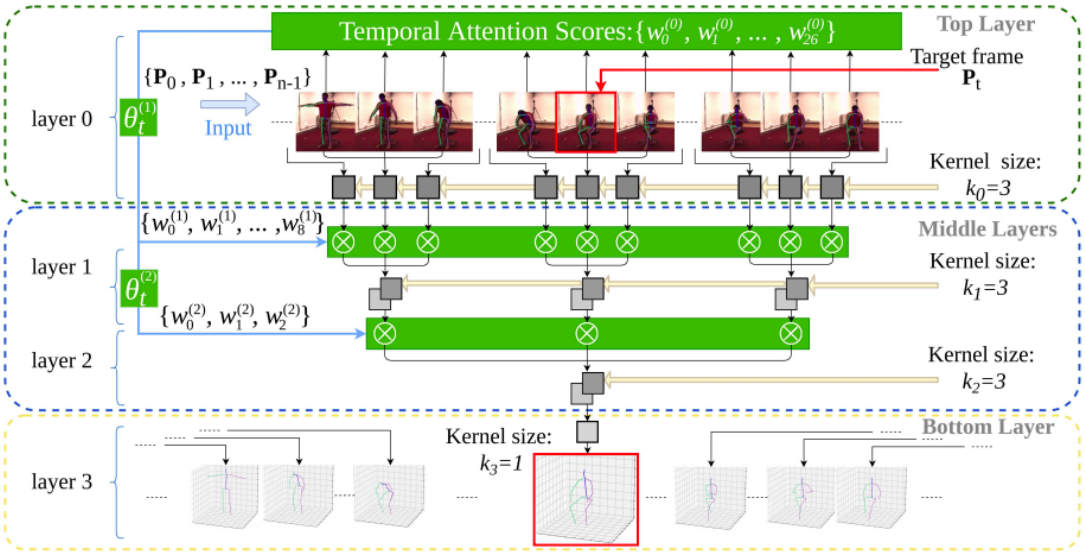
Slika 3.8: Pregled *Lifting from the Deep* metode, [6]

poput *PoseFormer* [14] i *CondDGConv* [9] rješavaju prvi podzadatak služeći se gotovim 2D detektorima poza poput OpenPose [8], HR-Net [16], CPN [17], te se fokusiraju više na rješavanju drugog podzadatka (*PoseFormer* je opisan u poglavlju Pozornost u 3D procjeni, dok *CondDGConv* je opisan u poglavlju Trenutno najnapredniji rad monokularnog pristupa za procjenu poze u 3D). Ovaj pristup je bio popularizirani radom *A simple yet effective baseline for 3d human pose estimation* [18] zbog svoje jednostavnosti i brzine te činjenica da kada je ovaj rad izašao (2017. godine) bio je bolji od tadašnje najbolje metoda za oko 30% na Human3.6M skupu podataka.

3.2.1.1.2. Pozornost u 3D procjeni

Real-time 3D Human Pose Reconstruction [3] naglašavaju dva nedostatka trenutnih metoda. To su vremenska nepovezanost i malo receptivno polje. Zbog toga kreiraju metodu koristeći mehanizme pozornosti kako bi odredili značajne okvire videozapisa (engl. *frames*) i značajne dijelove izlaznih tenzora za svaki sloj duboke mreže. Za dobivanje većeg receptivnog polja korištene su proširene konvolucije (engl. *dilated convolutions*) koje modeliraju dalekometne odnose između okvira videozapisa. Slika 3.9 prikazuje cjelokupnu arhitekturu ovog rada. Na ulazu model prima niz okvira videozapisa. Korištena su dvije procedure koje primjenjuju mehanizme pozornosti: modul vremenske pozornosti (engl. *Temporal Attention module*) i modul jezgrene pozornosti (engl. *Kernel Attention module*). Vremenske pozornosti su prikazne kao zeleni blokovi dok jezgrene pozornosti su svijetlo i tamno sivi blokovi na slici 3.9. Jezgrene pozornosti se sastoje od vremenskih konvolucijskih mreža (engl. *temporal convolu-*

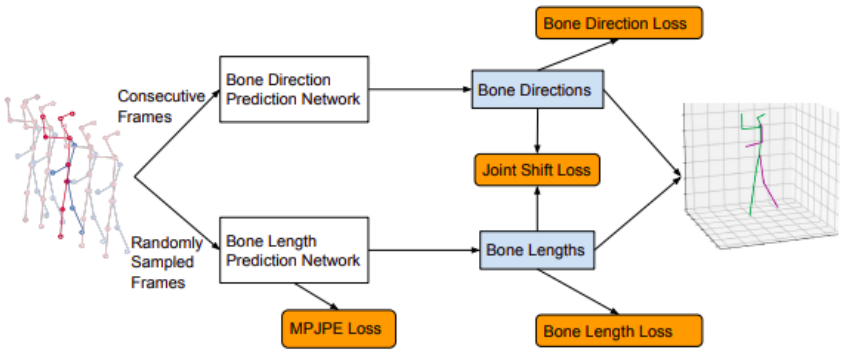
tional network, odnosno *TCN*) i linearnih projekcija (tamno sivi, odnosno svijetlo sivi blokovi). Gledajući funkcionalnost, slojevi se mogu grupirati u 3 skupine: gornji, srednji i donji slojevi. Cilj vremenskih pozornosti jest da kreiraju mjeru doprinosa (engl. *contribution metric*) koja će poslužiti izlaznim tenzorima. Svaki modul vremenske pozornosti kreira skup skalara koji određuju značaj različitih tenzora unutar sloja. Slično stvar radi i jezgrena pozornost koja određuje značaj kanala jezgre za određeni tenzor (engl. *channel weight distribution*).



Slika 3.9: Pregled arhitekture rada *Real-time 3D Human Pose Reconstruction*, [3]

Anatomy-aware 3D Human Pose Estimation [4] koristi drugačiji pristup procjene poze. Zadatak svode na procjenu smjera kostiju i procjenu duljine kostiju, te koristeći te informacije određuju pozu u 3D prostoru. Motivacija za ovakav pristup jest činjenica da duljine kostiju ljudskog kostura ostaju nepromijenjene za vrijeme čitavog videozapisa. Za procjenu smjera kostiju predlažu potpunu konvolucijsku arhitekturu s dugim *skip* vezama. Također koriste implicitni mehanizam pozornosti kako bi dobili podatak o vidljivosti 2D koordinata dijelova tijela koji značajno smanjuje problem neodređenosti dubine kod zahtjevnih poza. Slika 3.10 prikazuje pregled ove metode gdje možemo primijetiti kako za određivanje smjera kostiju koristimo uzastopne okvire a, za duljinu kostiju koristimo nasumične okvire.

Rad [10] predlaže pristup koristeći *PoseFormer* model. Inspiriran je *transformer* arhitekturom, koja je revolucionirala područje obrade prirodnog jezika (engl. *natural language processing*), te ne koristi konvolucijske mreže prilikom procjene ljudske poze. Kreirani model je *spatial-temporal transformer* koji je zadužen za modeliranje odnosa između dijelova tijela za svaki okvir, kao i za modeliranje vremenskih od-

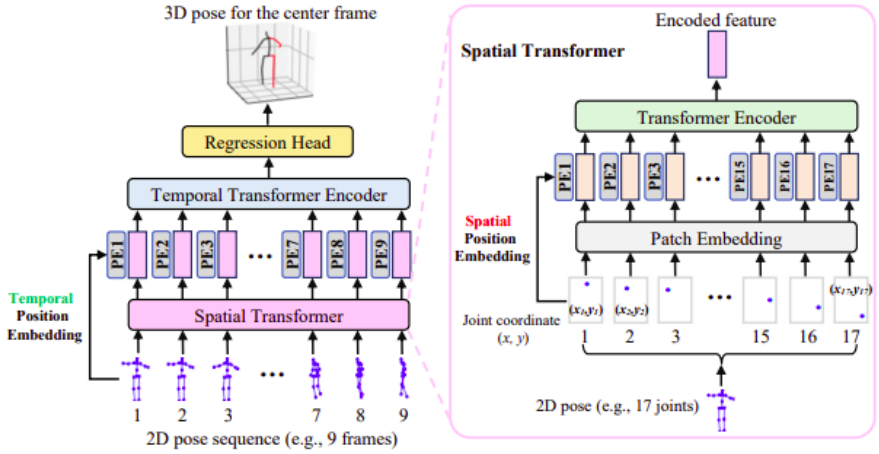


Slika 3.10: Pregled arhitekture rada *Anatomy-aware 3D Human Pose Estimation*, [4]

nosa okvira videozapisa. Slojevi koji koriste prostornu pozornost (engl. *spatial self-attention*) uzimaju prostornu informaciju 2D dijelova tijela i vraćaju latentne reprezentacije značajki za svaki okvir posebno. Modul vremenskog *transformer* analizira globalne ovisnosti između reprezentacija prostornih značajki te generira 3D procjenu ljudske poze. Slika 3.11 daje prikaz arhitekture modela, gdje vidimo da ulaz je niz 2D koordinata dijelova tijela a, izlaz je 3D poza za *center frame*. Ulazni niz formalno je zapisan kao $X \epsilon R^{f \times (J*2)}$, gdje f je broj okvira ,a J je broj dijelova tijela. Što znači da za niz koji se sastoji od 9 okvira i 17 dijelova tijela, dimenzija ulazne varijable X bi bila 9×34. Svaki dio tijela (znači njegova x i y koordinata) se smatraju kao *patch*, te se svaki *patch* predstavi kao višedimenzionalna značajka. Tu proceduru odrađuje *Patch Embedding* blok sa slike 3.11. Višedimenzionalne značajke su sumirane s prostorno pozicijskim značajkama (engl. *spatial positional embedding*), koje služe za očuvanje informacije o poziciji unutar niza. Sumirana vrijednost je predana enkoderu (zeleni blok nazvan *Transformer Encoder* na slici 3.11) koji koristeći *self-attention* mehanizam iskorištava informacije svih dijelova tijela.

Vremenski *transformer* kao ulaz prima sumiranu vrijednost izlaza prostornog *transformer* i vremenskih pozicijskih značajki (engl. *temporal positional embedding*). Potom *Regression Head* blok kao ulaze prima rezultate vremenskog *transformer*. Ovaj blok ulaznu kodiranu značajku Y dimenzije $f \times (J * c)$ (c je *embedding* dimenzija) pretvara u $y \epsilon R^{1 \times (J*c)}$, odnosno smanjuje dimenziju. Taj postupak je nužan jer od niza okvira predanih na ulazu potrebno je dobiti 3D pozu za *center frame*. Potom koristeći višeslojni perceptron (engl. *multilayer perceptron*) i normalizacijski sloj dobivamo 3D pozu za *center frame*.

METRO [5] i *Mesh Graphormer* [2] imaju jako slične pristupe određivanja 3D poza. Što nije začuđujuće uzimajući u obzir da *Mesh Graphormer* se nastavlja na rad

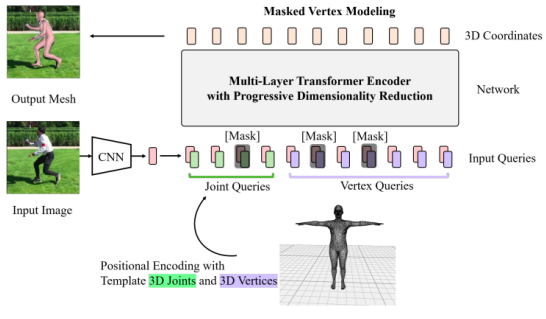


Slika 3.11: Pregled arhitekture *PoseFormer* modela, [10]

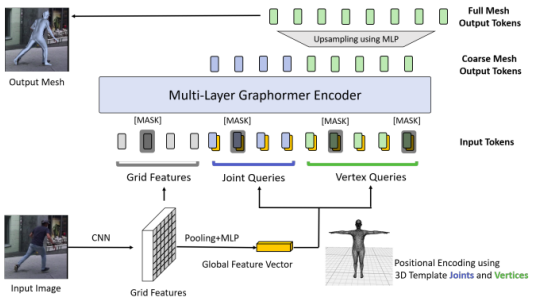
u kojem je predložen *METRO*. Njihove arhitekture možemo vidjeti na slici 3.12. Oba rada koriste konvolucijske mreže kako bi dobili značajke slike (razlika u odnosu na *PoseFormer* model). Značajke slika dobivenih iz konvolucijske mreže su kombinirane s predloškom oblika ljudskog tijela (engl. *template human mesh*) tako što se vektor značajki slike spaja s 3D koordinatama dijelova tijela i 3D koordinatama vrhova predloška. Uz to *Mesh Graphormer* koristi i *grid features* vrijednosti kao ulaz. *Grid features* su dobivene iz posljednjeg konvolucijskog bloka konvolucijske mreže.

Struktura višeslojnih enkodera (*Multi-Layer Transformer Encoder with Progressive Dimensionality Reduction* i *Multi-Layer Graphormer Encoder* blokovi na slikama 3.12 (a) i (b)) je slična i sastoji se od naslaganih *transformer* enkodera i progresivnog smanjivanja dimenzionalnosti (engl. *progressive dimensionality reduction*). *Mesh Graphormer* ima dodan *Graph Residual Block* koji je predstavljen zelenim blokom unutar svojeg enkodera, dok ostatak je identičan s enkoderom *METRO* modela (Slika 3.13).

Kako bi bolje razumjeli kako funkcioniraju mehanizmi pozornosti nužno je objasniti

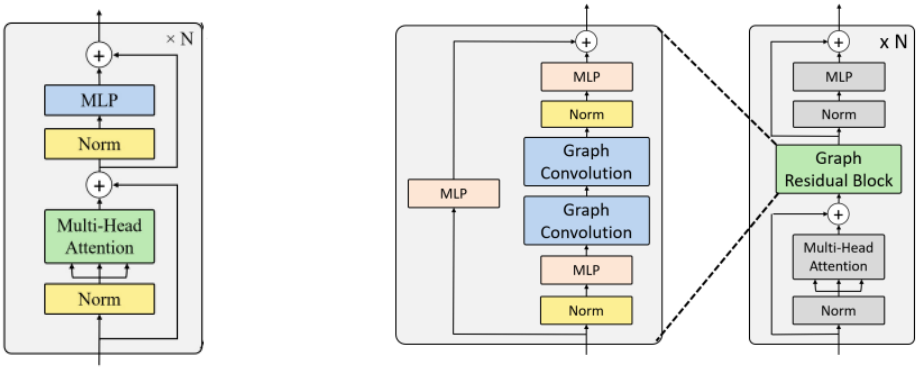


(a) Arhitektura *METRO* modela, [5]



(b) Arhitektura *Mesh Graphormer* modela, [2]

Slika 3.12: Pregled arhitektura



(a) Struktura *METRO* enkodera, [5] **(b)** Struktura *Mesh Graphormer* enkodera, [2]

Slika 3.13: Pregled struktura enkodera

ulogu *Multi-Head Attention* bloka sa slika 3.11 (a) i (b). Započinjemo prvo razumijevanjem funkcija pozornosti (jednadžba (3.2)).

$$Attention(Q, K, V) = softmax(QK^T)V \tag{3.2}$$

Funkcija pozornosti može se opisati kao mapiranje upita Q i skupa parova ključ-vrijednost, (K predstavlja ključ a, V vrijednost). Izlaz je izračunat kao težinski zbroj vrijednosti. Težina koja su dodijeljena određenoj vrijednosti je dobivena funkcijom kompatibilnosti čiji ulazi su upit Q i odgovarajući ključ K. Funkcija pozornosti koja je korištena u *Multi-Head Attention* bloku je *dot-product* funkcija pozornosti (jednadžba (3.2)).

Umjesto pozivanja funkcije pozornosti samo jedanput, ona se može pozivati više puta linearnom projekcijom upita, ključeva i vrijednosti h puta s različitim linearnim projekcijama. Nad svakim od linearno projektiranih upita, ključeva i vrijednosti paralelno se poziva funkcija pozornosti te su izlazi združeni i projektirani posljednji put. Ovu opisanu proceduru radi *Multi-Head Attention* blok a to se formalno zapisuje jednadžbom (3.3).

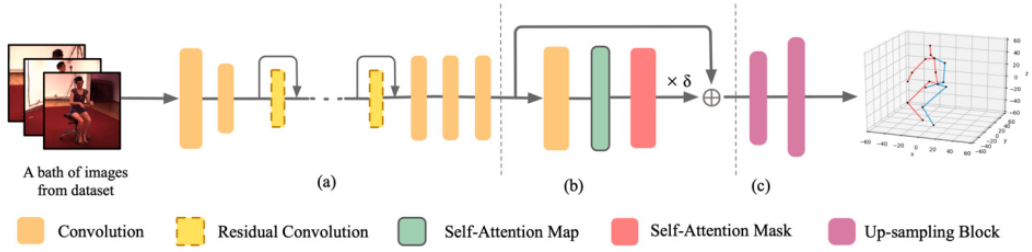
$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{3.3}$$

$$gdje\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

gdje linearne projekcije upita, ključeva, vrijednosti i posljednje projekcije su predstavljene matricama W_i^Q, W_i^K, W_i^V i W_i^O. Možemo primijetiti da su ulazi u *METRO* na slici 3.12 (a) nazvani ulaznim upitima (engl. *input queries*) (ista stvar vrijedi i za *Mesh Graphormer* iako nije na slici eksplicitno navedeno).

Model *Self-Attention Network* [1] ima sličnu ideju kao i dosad, tj. koristi se konvolucijom i mehanizmima pozornosti da bi dobio 3D pozu (slika 3.14). Pristup se sastoji od 3 dijela (na slici 3.14 označeni kao (a), (b) i (c)):

1. Dobivanje značajki slike koristeći *ResNet* arhitekturu.
2. Korištenje *Self-Attention Network* čiji je zadatak razumijevanje dalekometnih odnosa dijelova tijela. Povećanjem δ parametra model će više ovisiti o nelokalnoj informaciji.
3. Naduzorkovanje kako bi se dobio konačni rezultat odnosno 3D poza.



Slika 3.14: Pregled pristupa *Self-Attention Network* rada, [1]

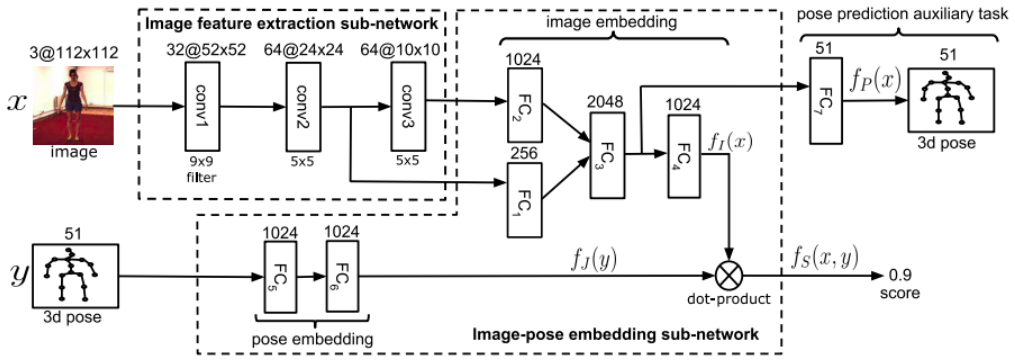
3.2.1.1.3. Konvolucijske mreže u 3D procjeni

Zahvaljujući svojoj efikasnosti konvolucijske mreže postale su popularne u rješavanju velikog broja problema u računalnom vidu, poput detekcije i klasifikacije objekata pa tako i procjena 2D i 3D ljudskih poza. Uglavnom svi radovi koji se baziraju na pristupu dubokog učenja inkorporiraju neki oblik konvolucijskih neuronskih mreža.

Rad *Maximum-Margin Structured Learning* [7] predlaže mrežu koja kao ulaz primi sliku i 3D pozu te na izlazu vraća iznos rezultata koji je visok ako poza odgovara ulaznoj slici, inače je nizak. Mreža se sastoji od konvolucijske mreže za dobivanje značajki slike popraćene s dvije podmreže za transformiranje značajki slike i pose u *embedding* vrijednosti (Slika 3.15). Konačna vrijednost je skalarni umnožak *embedding* vektora. Pomoć u odabiru boljih značajki za predikciji poze odrađuje dio imenovan *prediction pose auxiliary task*

Lifting from the deep kao što je i ranije naglašeno u poglavlju Podizanje iz 2D u 3D prostor, se jako oslanja na konvolucijske mreže da u svakoj novoj etapi generiraju točnije rezultate 2D i 3D polja lokacija.

2D detektori poza (poput *OpenPose* [8]), na koje se mnogi radovi oslanjaju (poput *U-CondDGConv* [9] i *PoseFormer* modela) moraju što bolje odraditi procjene 2D ljudske



Slika 3.15: Pregled pristupa *Maximum-Margin Structured Learning* rada, [7]

poze u čemu im pomažu konvolucijske mreže efektivnim izvlačenjem značajki slika. Osim kod metoda koje koriste toplinske karte, konvolucija se pokazala iznimno korisnim u kombinaciji s mehanizmima pozornosti, što primjećujemo u modelima *METRO* , *Mesh Graphormer* i Self-Attention Network.

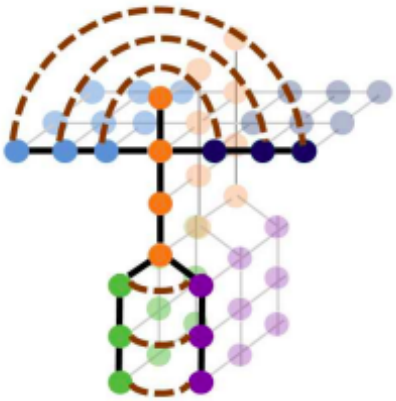
Međutim za dobivanje većeg receptivnog polja konvolucijske mreže se susreću s istim problemima kao i u 2D slučaju (nestajući gradijenti i korištenje velike računalne moći). Ipak zahvaljujući njihovim sposobnostima u otkrivanju lokalnih informacija postale su neizostavan dio skoro svih modela koji se bave procjenom poze ljudskog tijela u 3D prostoru.

3.2.1.1.4. Graf arhitekture

Kako bi riješili problem procjene ljudske poze, reprezentacija ljudske poze je jako bitna. Sama reprezentacija ljudske poze se može prikazati koristeći graf strukturu. Tako je u radu [19] neusmjereni graf korišten za prikaz vremenske i prostorne reprezentacije ljudske poze (Slika 3.16) te su korištene graf konvolucijske mreže za procjenu 3D poze. Međutim neusmjereni grafovi ne uspijevaju uzeti u obzir hijerarhijsku strukturu kostiju što je jedna od značajnijih karakteristika ljudske anatomije.

Rad [20] koristi usmjerene necikličke grafove koji su bazirani na kinematičkoj ovisnosti između dijelova tijela i kostiju.

U-shaped Conditional Directed Graph Convolutional Network [9] za reprezentaciju koristi usmjerene grafove te kako bi iskoristio nelokalne ovisnosti između čvorova uvodi uvjetne konvolucije (detaljniji opis ovog rada je dan u poglavlju Trenutno najnapredniji rad monokularnog pristupa za procjenu poze u 3D).



Slika 3.16: Prostorne i vremenske ovisnosti prikazane koristeći neusmjereni graf, [19]

3.2.1.2. Trenutno najnapredniji rad monokularnog pristupa za procjenu poze u 3D

U ovom potpoglavlju bit će opisana jedna od najboljih metoda, uzimajući MPJPE kao metriku. Trenutno najbolji radovi za monokularni pristup su: MixSTE [24] i U-CondDGConv [9]. Sortirana lista radova po MPJPE metrici se nalazi na [25].

3.2.1.2.1. U-CondDGConv

U-shaped Conditional Directed Graph Convolutional Network je model koji radi procjenu 3D poze iz monokularnih videozapisa koristeći usmjerene grafove za reprezentaciju ljudskih poza.

Pristup *U-CondDGCN* modela

Slika 3.17 prikazuje pristup kojim *U-CondDGCN* metoda dolazi do procjenjene poze. Kao što možemo primjetiti metoda se sastoji od 3 dijela:

1. poduzorkovanje (engl. *Downsampling Stage*)
2. naduzorkovanje (engl. *Upsampling Stage*)
3. spajanje (engl. *Merging Stage*).

Svaki dio se sastoji od posebnih blokova koji imaju svoju ulogu prilikom određivanja procjene. Korišteni blokovi (vidljivi na slici 3.17) su:

- *Spatial-temporal directed graph convolution*, odnosno *ST-DGConv*
- *Spatial-temporal conditional directed graph convolution*, odnosno *ST-CondDGConv*
- vremensko poduzorkovanje

– vremensko naduzorkovanje

– potpuno povezani blok

Uloga svakog dijela i bloka koji se koristi za određivanje poze će biti objašnjena u narednim paragrafima. Važna stvar za primjetiti jest da ulaz ove metode, za razliku od većine metoda nije slika, odnosno videozapis, već vremenski niz usmjerenih grafova. Ulaz kojeg čini niz 2D poza dobivenih iz monokularnih videozapisa se može procijeniti koristeći popularne 2D procjenjivače (engl. *2D pose estimators*). Jednadžba (3.4) prikazuje ulazni niz:

$$P_{2D} = \{X_{t,j} \epsilon R^2 | t = 1, 2, 3, ...T;\ j = 1, 2, ..., J\} \tag{3.4}$$

gdje T i J označavaju broj okvira (engl. *frames*) i broj dijelova ljudskog tijela (engl. *joints*).

Potom je konstruiran vremenski niz usmjerenih grafova iz niza ljudskih poza iz 2D slike (ovaj proces je prikazan *Directed Graph Construction* strijelicom na slici 3.17). Čvorovi usmjerenog grafa označavaju glavne dijelove ljudskog tijela dok bridovi su kosti između dijelova tijela. Struktura grafa je prikazana na slici 3.20. Direkcije bridova modeliramo po konvenciji definiranoj u BVH (*Biovision hierarchical data*) formatu, koji je vidljiv na slici 3.19. Kuk je početni čvor jer je on centar gravitacije ljudskog tijela (označen je crvenom bojom na slikama 3.17 i 3.20). Značajke koje su asocirane s čvorovima i bridovima su inicijalizirane koristeći lokacije djielova tijela (engl. *joints' locations*) i njihove derivacije prvog reda (razlika između dijete i roditelj čvora). Vremenski niz usmjerenog grafa se može formulirati koristeći (3.5)

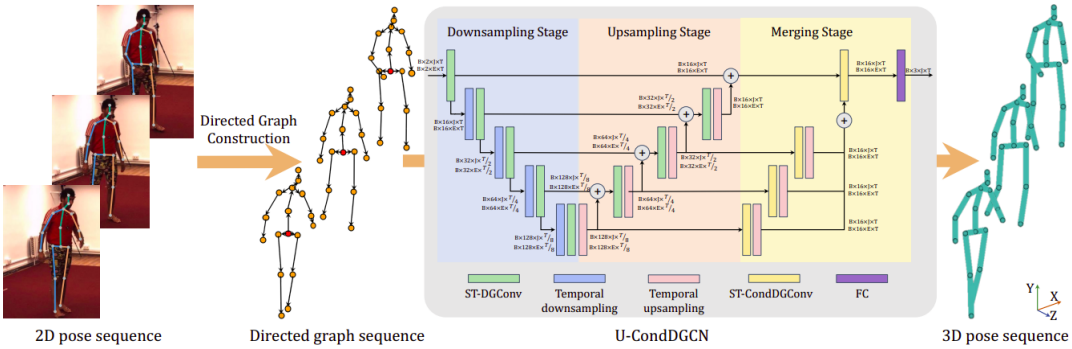
$$G_{2D} = \{G_t = (N, \varepsilon) | t = 1, 2, 3..., T\} \tag{3.5}$$

gdje N je skup čvorova, ε je skup usmjerenih bridova. Zatim možemo primijeniti *U-CondDGCN* za procjenu niza poza u 3D prostoru, te na izlazu dobivamo 3D poze koje formalno prikazujemo jednadžbom (3.6)

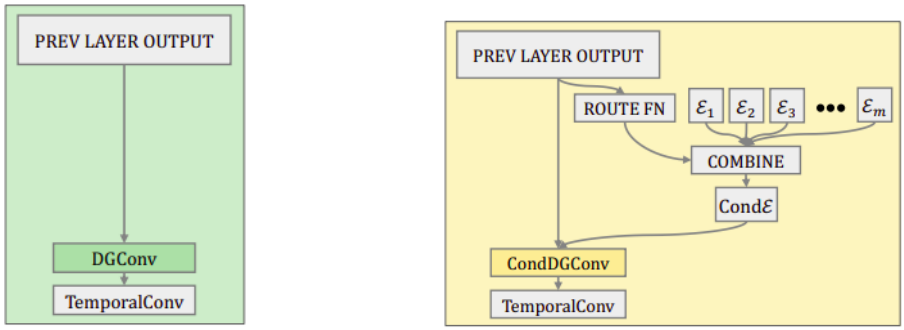
$$P_{3D} = \{X_{t,j} \epsilon R^3 | t = 1, 2, 3..., T;\ j = 1, 2, 3, ...J\} \tag{3.6}$$

Blokovi mreže

Značaj ovog rada leži u tipovima blokova koji su korišteni za rješavanje zadatka procjene te ovdje su opisane njihove funkcije. Da bi spojili prostorne i vremenske značajke u radu koriste 5 tipova blokova: ST-DGConv, ST-CondDGConv, vremenskon poduzorkovanje, vremensko naduzorkovanje i potpuno povezani blok.



Slika 3.17: Pregled UConvDGCN metode



(a) ST-DGConv **(b)** ST-CondDGConv

Slika 3.18: Prikaz ST-DGConv i ST-CondDGConv bloka

ST-DGConv blok se sastoji od konvolucije usmjerenog grafa, odnosno *DGconv* popraćene vremenskom konvolucijom. *DGconv* koristi prostorne odnose spajajući značajke susjednih bridova i čvorova. Za iskorištavanje prednosti vremenskih odnosa, korištene su vremenske konvolucije, što su zapravo 1D konvolucije. Prikaz strukture ovog bloka je vidljiv na slici 3.18 (a)

ST-CondDGConv se razlikuje od prijašnjeg bloka koji je baziran na fiksnoj usmjerenoj graf konekciji ε koja je definirana pomoću prirodne strukture ljudskog tijela. Ali ne uspijeva uhvatiti nelokalne ovisnosti koje mogu biti od velike koristi prilikom procjene. Inspirirani uvjetnim konvolucijama (engl. *conditional convolution*) koje omogućuju različito uzorkovanje podataka koristeći različite konvolucijske jezgre, rad predlaže *ST-CondDGConv* za uvjetovanje konekcija usmjerenog grafa na ulaznim pozama, tako da različite poze mogu poprimiti odgovarajuće konekcije da bi iskoristili raznoliku nelokalnu ovisnost. Struktura bloka je vidljiva na slici 3.18 (b). Na spomenutoj slici postoje nizovi baza matrica povezanosti $\{\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_m\}$ (m je broj baza) i *routing* funkcija (na slici 3.18 (b) imenovana ROUTE FN).

Struktura *routing* funkcije se sastoji od sloja sažimanje, točnije *global average pooling*

sloja, popraćenog potpuno povezanim slojem i sigmoidnom aktivacijom funkcijom. *Routing* funkcija se koristi za predviđanje *bleding* težina koje služe za određivanje uvjetne konekcije.
Nakon provlačenja izlaza prijašnjeg sloja kroz *routing* funkciju, izlaz, odnosno *blending* težine se koriste za linearnu kombinaciju (engl. *linear combination*) baza kako bi se kreirala uvjetna konekcija $Cond\varepsilon$. Uvjetna konekcija je dana kao ulaz CondDGConv bloku kako bi se spojile lokalne i nelokalne prostorne informacije. Konačno korištena je vremenska konvolucija kako bi se spojile vremenske i prostorne informacije.
Vremensko poduzorkovanje je zapravo ST-DGConv blok gdje je *stride* vrijednost unutarnje vremenske konvolucije postavljena na 2. Koristi se za poduzorkovanje vremenske rezolucije za veće receptivno polje.
Vremensko naduzorkovanje je bilinearna interpolacija nad vremenskom osi za dobivanje veće vremenska rezolucija.
Potpuno povezani sloj predviđa 3D pozu iz izvučenih značajki usmjerenih grafova.

Dijelovi UCondDGCN metode

Kao što je ranije navedeno, UCondDGCN ima 3 dijela. Svaki od njih koristi kombinaciju prethodno opisanih blokova kako bi obavio određen zadatak i time dao procjenu poze. Zadatci koje ti dijelovi obavljaju su :

1. *Downsampling Stage* radi poduzorkovanje za spajanje informacija u *long-time* rasponima koristeći vremensko sažimanje (engl. *temporal pooling*)

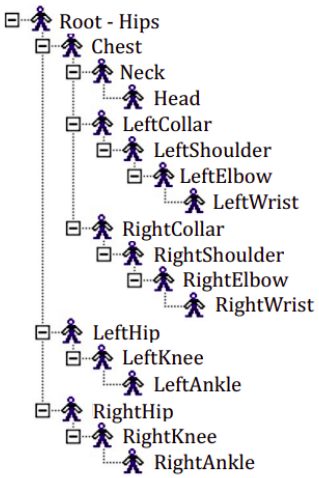
2. *Upsampling Stage* radi naduzorkovanje za dobivanje natrag vremenske rezolucije i koristi *skip* konekcije između poduzorkovanja i naduzorkovanja za integraciju *low-level* detalja

3. *Merging Stage* radi spajanje za kombinaciju mapa značajki različitih nivoa (engl. *multi-scale feature maps*) za predviđanje konačne 3D poze

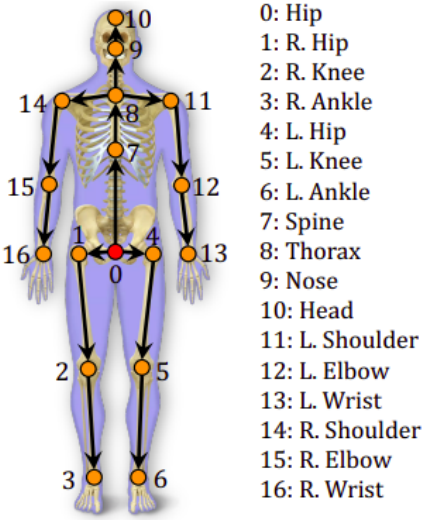
ST-DGConv može bolje koristiti prijašnje znanje o prirodnoj strukturi ljudskog tijela, dok ST-CondDGconv omogućava različitim pozama da poprime odgovarajuće nelokalne ovisnosti.

3.2.2. Pristup s više pogleda

U ovom poglavlju bit će opisani radovi koji koriste pristup s više pogleda a, potom će se opisati dvije trenutno najbolje metode za procjenu 3D poze.



Slika 3.19: BHV format podataka



Slika 3.20: Struktura usmjerenog grafa na ljudskom kosturu

3.2.2.1. Pregled radova koji koriste više pogleda

Postoje mnoge metode koje rješavaju procjenu ljudske poze koristeći više pogleda. Generalno se koriste za anotiranje točnih poza za monokularne procjene ,ali novi radovi ih kreću koristiti kao samostalne metode za određivanje ljudske poze u 3D prostoru. *3D human pose estimation in video with temporal convolutions and semi-supervised training* [35] je predstavio procjenjivanje 3D poza u videozapisima s proširenim vremenskim konvolucijama koristeći 2D lokacije. *Learning monocular 3d human pose estimation from multiview images* [33] predlaže korištenje ograničenja kao slabo nadziranja (engl. *weak supervision*) za poboljšanje monokularnog 3D detektora ljudskih poza kada je broj anotiranih podataka premalen. *Epipolar transformers* [32] se oslanjaju na kalibracije kamera ali pritom koriste minimalno parametara. Ovo čini proces učenja lakšim i zahtijeva manje podataka za treniranje. Osim toga mreža trenirana koristeći *epipolarne transformere* može se koristiti na nikad viđenom *multi-camera* sustavu bez dodatnog treniranja. *Learnable triangulation of human pose* odnosno *LT* [27] uči 3D poze koristeći diferencijalnu triangulaciju (njihove metode će biti detaljno opisane u narednim poglavljima).

Također različite reprezentacije poza su prisutne i u metodama koje koriste više poza, pa tako imamo radove poput *A Massively Multiview System for Social Interaction Capture* [23] i *Harvesting multiple views for marker-less 3d human pose annotations* [22] koriste volumetrijske reprezentacije poza.

3.2.2.2. Trenutno najnapredniji radovi koji koriste više pogleda za procjenu poze u 3D

U ovom potpoglavlju opisat ćemo najbolje metode, uzimajući MPJPE kao metriku. Trenutno najbolji radovi ovog pristupa su: *Learnable Human Mesh Triangulation*, odnosno *LMT* [26] i *Learnable Triangulation of Human Pose*, odnosno *LT* [27]. Sortirana lista radova po MPJPE metrici se nalazi na [25].

3.2.2.2.1. Learnable Human Mesh Triangulation

Ovaj rad predlaže metodu koja procjenjuje 3D oblik osobe (engl. *mesh*) baziran na *SMPL* modelu koristeći slike iz različitih pogleda dobivenih iz C kamera. Čitava arhitektura (slika 3.21) se sastoji od: *visibility module*, CNN *backbone*, *feature aggregation module*, *vertex regression module* i *fitting module*.

Visibility module procjenjuje vidljivost svakog vrha (engl. *per-vertex visibility*). Izlazi ovog bloka su prikazani strijelicama $v_1, ..., v_C$ na slici 3.21. *Visibility module* radi procjenu vidljivosti vrhova $v_c \epsilon R^N$ za svaki poduzorkovani oblik ljudskog tijela (engl. *subsampled mesh*) za svaku sliku I_c gdje N je broj poduzorkovanih vrhova oblika ljudskog tijela. Poduzorkovani oblik ljudskog tijela se koristi zbog smanjenja potrebne računalne moći za određivanja poze.

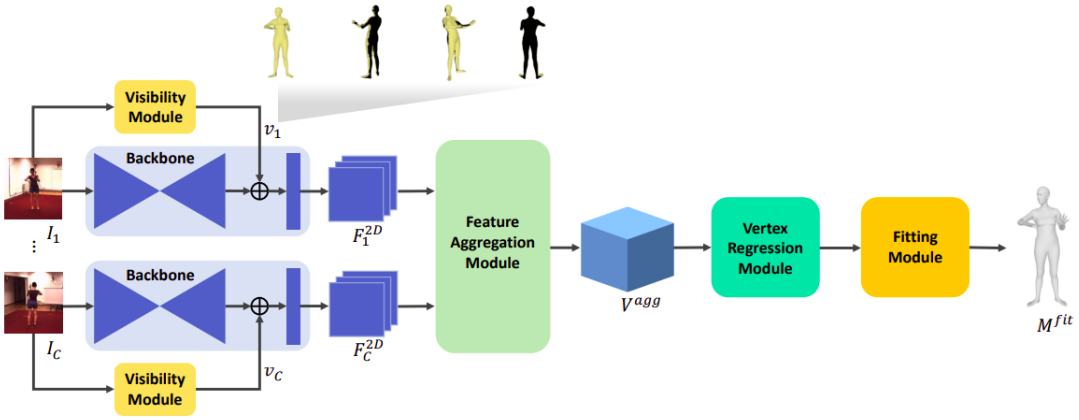
CNN *backbone* računa značajke ulaznih slika F_c^{2D} koristeći ulazne slike I_c i izlaze *visibility module* bloka v_c. Na slici 3.21 to je prikazano *backbone* dijelom i simbolom $\oplus$, gdje se spajaju informacije vidljivosti i izlaza *backbone* dijela konvolucijske mreže.

Feature aggregation module radi *unprojection* ulaznih značajki F_c^{2D} u 3D globalni voksel prostor da bi generirao C volumetričkih značajki V_c^{unproj}, te potom spaja značajke svake kamere da bi kreirao združenu volumetrijsku značajku V^{agg} (zeleni blok na slici 3.21).

Vertex regression module generira 3D koordinate vrhova poduzorkovanog oblika ljudskog tijela iz združenih značajki V^{agg} koristeći 3D konvoluciju i *soft-argmax* operaciju.

Fitting module daje izlaz koji je završna informacija o rotaciji dijelova tijela, koordinatama i obliku koja je dobivna podešavanjem (engl. *fitting*) *SMPL* modela na 3D koordinate vrhova M dobivenih iz *vertex regression* bloka.

Nakon ovog kratkog uvoda da bi bolje razumjeli kako ova metoda funkcionira u više detalja je opisan svaki od gore spomenutih dijelova.



Slika 3.21: Learnable Human Mesh Triangulation (LMT) metoda, $\oplus$ simbol predstavlja spajanje (engl. *concatenation*), žuti dijelovi oblika ljudske poze (engl. *mesh*) prikazuje vidljive vrhove za taj pogled

Visibility module

Visibility module računa mapu vidljivosti vrhova v_c iz slike I_c. Ovaj modul se implementira koristeći *I2L-MeshNet* [36], koji je jedan od najboljih modela za rekonstrukciju oblika ljudskog tijela iz jedne slike. Kao ulaz *I2L-MeshNet* mreži predana je slika I_c i dobiven je oblik ljudskog tijela (engl. *human mesh*) unutar koordinatnog sustava čije središte je određeno dijelom tijela koji predstavlja zdjelicu (engl. *pelvis joint*). Međutim za određivanje vidljivih dijelova nužno je znati koordinate kamere. Algebarskom triangulacijom iz *LT* pristupa [27] (objašnjena kasnije) dobivene su koordinate kamere. Algoritam za računanje vidljivih dijelova se onda koristi da bi se dobila mapa vidljivosti za cijeli oblik ljudskog tijela, odnosno vidljivost svakog vrha $v_c \epsilon R^{6890}$. Kako ne bi se desio *overfitting* predloženog modela koristi se poduzorkovani oblika ljudskog tijela.

Backbone

Ovaj dio daje značajke slike $\{F_c^{2D}\}_{c=1}^{C}$ koristeći ulazne slike iz više pogleda i mape vidljivosti dobivene iz *visibility module*. Za konstrukciju predloženog *backbone*, potrebno je izbaciti zadnji klasifikacijski sloj i sloj sažimanja *ResNet-152* arhitekture, koja je prethodno trenirana koristeći *COCO* [15] i *MPII* [39] skup podataka. Zatim treba dodati 3 *deconvolution* sloja i 1x1 konvolucijski sloj. Zadnji *deconvolution* sloj kreira značajke F_c^{deconv}. Ove značajke su prikazane na slici 3.21 kao strjelica koja izlazi iz *Backbone* dijela i ulazi u *concatenate* simbol $\oplus$. Nakon toga mapa vidljivosti v_c je spojena sa spomenutom značajkom. Dodatni konvolucijski sloj je dodan da bi

generirao finalne značajke slike F_C^{2D}.

Feature Aggregation Module

U ovom bloku 2D značajke F_c^{2D} iz *backbone* dijela su konvertirane u volumetrijske značajke V_c^{unproj} korištenjem *unprocjection* tehnike, odnosno prebacivanje u višedimenzionalni prostor. Slika 3.22 daje vizualni prikaz *unprocjection* tehnike. Združena volumetrijska značajka V^{agg} je izračunata spajanjem novonastalih volumetrijskih značajki V_c^{unproj}.

U predloženoj metodi procjena koordinata vrhova poduzorkovanog oblika ljudskog tijela M ovisi o 3D značajkama dobivenih *unprocjection* tehnikom u 3D prostoru, koji je ograničen na oblik kocke (engl. *cuboid*). Zbog toga lokacija i veličina kocke bi trebale biti postavljene tako da kocka sadrži ciljani oblik ljudskog tijela. Koristi se kocka čije središte je određeno lokacijom dijela tijela koji predstavlja zdjelicu i duljina brida iznosi 2 metra.

Procedura ovog bloka započinje projekcijom 3D koordinata vrhova voksela V^{coords} u 2D prostor za svaki pogled zasebno. Za to su potrebne projekcijske matrice kamera (engl, *camera projection matrix*) te kao izlaz ove procedure dobivamo 2D koordinate V^{proj}.

Potom se bilinearnim uzorkovanjem mapiraju 2D značajke F_c^{2D} na svaku lokaciju V_c^{proj} značajki, što je formalno napisano jednadžbom (3.7).

$$V_c^{unproj} = F_c^{2D}\{V_c^{unproj}\} \tag{3.7}$$

gdje $\{.\}$ predstavlja bilinearno uzorkovanje. Nakon toga je C značajki u 3D prostoru spojeno koristeći 3D *softmax* operaciju (jednadžba (3.9)). Jednadžba kojom se dobivaju združene značajke je prikazana u (3.8):

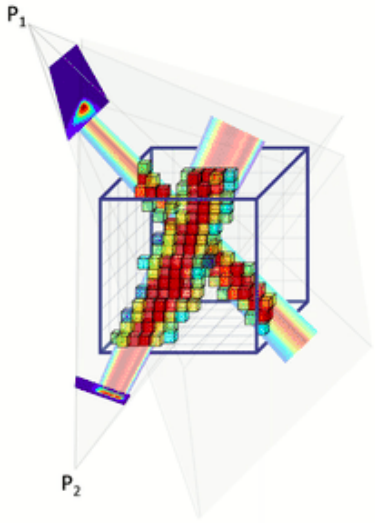
$$V^{agg} = \sum_{c=1}^{C}(d_c \odot V_c^{unproj}) \tag{3.8}$$

$$d_c = \frac{\exp\left(V_c^{unproj}\right)}{\sum_{c=1}^{C}\exp\left(V_c^{unproj}\right)} \tag{3.9}$$

gdje d_c i $\odot$ su *confidence* težine i Hadamardov umnožak (engl. *Hadamard product*).

Vertex Regression Module

Vertex regression blok ima strukturu koja se sastoji od koderskog i dekoderskog bloka unutar kojih se nalaze 3D konvolucije. Ovaj blok generira koordinate vrhova za oblik ljudskog tijela koristeći združene značajke V^{agg}, koje dobiva kao ulaz (vidljivo na slici



Slika 3.22: Prikaz *unprojection* tehnike

3.21). Enkoder prvo računa 3D značajke s rezolucijom $2 \times 2 \times 2$ i dimenzijom kanala koja iznosi 128. Izlaz enkodera je dan kao ulaz dekoder bloku čiji izlaz je volumetrijska značajka V. Nakon toga volumetrijska značajka V se predaje 3D konvoluciji kako bi se dobile 3D toplinske karte (engl. *heatmaps*) koje formalno označavamo s H^{3D}. 3D *soft-argmax* operacija koja je zaslužna za dobivanje koordinata vrhova M oblika ljudskog tijela je prikazana jednadžbma (3.10) i (3.11).

$$\bar{H}_n^{3D} = \frac{\exp(H_n^{3D})}{\sum_{i,j,k} \exp(H_n^{3D}(i,j,k))} \tag{3.10}$$

$$M_n = \sum_{i,j,k} r * \bar{H}_n^{3D}(i,j,k) \tag{3.11}$$

gdje $r = [r_i, r_j, r_k]$ predstavljaju vektor koordinate za voksel s indeksima (i,j,k) u 3D toplinskoj karti. H_n^{3D}, $\bar{H}_n^{3D}$ i M_n označavaju n-ti kanal 3D toplinkse karte, normaliziranu 3D toplinksu kartu i n-ti red vektora matrice koordinata vrhova M. Za treniranje mreže korišten je L1 gubitak za vrhove koje je generirao *vertex regresion* blok, vidljivo jednadžbom (3.12).

$$L_M = \frac{1}{N} \sum_{n=1}^{N} ||M_n - M_n^*||_1 \tag{3.12}$$

gdje M^* predstavlja stvarnu vrijednost oblika ljudskog tijela.

Fitting Module

Ovaj blok je korišten za dobivanje *SMPL* parametara koji odgovaraju koordinatama

vrhova M dobivenih iz prethodnog bloka. Blok koristi optimizacijske parametre poput: *VPoser's latent code* z, globalne rotacije R, *shape* parametra β i globalne translacije t. *VPoser*, označen s $\nu(.)$, koristeći z izračunava *SMPL* parametre θ (jednadžba (3.13)).

$$\theta = \nu(z) \tag{3.13}$$

SMPL dekoder koristi θ zajedno sa R, β i t parametrima kako bi kreirao SMPL oblik tijela (jednadžba (3.14), primjetiti koje su dimenzije izlaza dekodera, odnosno rezultat se sastoji od svih vrhova SMPL oblika).

$$M^{fit} = M(\theta, R, \beta, t)\epsilon R^{6890\times 3} \tag{3.14}$$

Koristeći *coarsening* funkciju $sub(.)$ radi se poduzokovanje vrhova (jednadžba (3.15))

$$M_{sub}^{fit} = sub(M^{fit}) \tag{3.15}$$

Fitting module ažurira sve parametre iterativno da bi smanjio razliku između poduzorkovanog oblika M_{fit} i izlaza *vertex regression* bloka M. *Cost* funkcija koja obavlja *fitting* je prikazana jednadžbama (3.16-3.18).

$$\varepsilon_{fit} = \varepsilon_{data} + \varepsilon_{reg} \tag{3.16}$$

$$\varepsilon_{data} = \frac{1}{N}\sum_{n=1}^{N} ||M_{sub,n}^{fit} - M_n||_2^2 \tag{3.17}$$

$$\varepsilon_{reg} = \lambda_z * \varepsilon_z + \lambda_\beta * \varepsilon_\beta + \lambda_w * \varepsilon_{\theta_w} + \lambda_\alpha * \varepsilon_\alpha \tag{3.18}$$

gdje $M_{sub,n}^{fit}$ i θ_w označavaju n-ti red vektora M_{sub}^{fit} matrice i *axis-angle* reprezentacije dijelova tijela koji predstavljaju zglobove. $\varepsilon_z, \varepsilon_\beta, \varepsilon_{\theta_w}$ su L2 regularizacijski izrazi. ε_α je eksponencijalni regularizaciji izraz koji služi za spriječavanje neprirodnih krivljenja zglobova i koljena. Svaka λ predstavlja regularizacijsku matricu.
Koordinate dijelova tijela se mogu dobiti koristeći M_{fit} i matrice za regresiju u oblika u dijelova tijela G (jednadžba (3.19)). Dobiveni J se koristi za evaluaciju procjene dijelova tijela (engl. *joint coordinate estimation*).

$$J = GM^{fit} \tag{3.19}$$

3.2.2.2.2. Learnable Triangulation of Human Pose

Pristup ovog rada pretpostavlja da imamo sinkronizirane videozapise iz C kamera s poznatom projekcijskom matricom P_c koje snimaju akcije jedne osobe u sceni. Ideja je procijeniti globalne 3D pozicije $y_{j,t}$ iz fiksnog skupa ljudskih dijelova tijela (engl.

joints) čiji su indeksi $j \in (1...J)$ za vremensku oznaku t (engl. *timestamp*). Okviri su procesirani neovisno za svaku vremensku oznaku, odnosno vremenska informacija nije korištena. Zbog toga indeks t može biti zanemaren.
Korišteni su *off-the-shelf* modeli za detekciju ljudi u 2D slikama. Nakon detekcije isječena slika ,koja sadrži osobu u sebi, je predana dubokoj konvolucijskoj neuronskoj mreži koja koristi *ResNet-152* arhitekturu popraćenu nizom transponiranih konvolucija koje proizvode toplinske karte kao međukorak procedure (engl. *intermediate heatmaps*). Posljednji dio mreže je konvolucijska mreža s jezgrom veličine 1×1 koja pretvara *intermediate heatmaps* u vjerojatnosti da piksel predstavlja određeni dio tijela (engl. *joint heatmaps*). Nakon dobivanja 2D lokacija rad predlaže dvije neovisne metode za dobivanje 3D ljudske poze. To su algebarska i volumetrijska triangulacija.

Pristup algebarske triangulacije

Prikaz ovog pristupa je vidljiv na slici 3.23. Ovdje procesiramo svaki dio tijela j neovisno od drugog. 2D pozicije koje se koriste za triangulaciju su dobivene iz toplinskih karti. Iz jednadžbe (3.20) možemo vidjeti da 2D pozicije zahtijevaju informaciju o vrijednosti toplinske karte j-tog dijela tijela za svaki pogled c.

$$H_{c,j} = h_\theta(I_c)_j \tag{3.20}$$

Za procjenu 2D pozicije prvo se računa *softmax* vrijednost (jednadžba (3.21)).

$$H^{'}_{c,j} = \exp(\alpha H_{c,j}) / (\sum_{r_x=1}^{W} \sum_{r_y=1}^{H} \exp(\alpha H_{c,j}(r))) \tag{3.21}$$

gdje parametar α će biti objašnjen kasnije. Potom računamo 2D pozicije dijelova tijela kao centar mase za odgovarajuću toplinsku kartu (jednadžba (3.22)), takozvana *soft-argmax* funkcija:

$$x_{c,j} = \sum_{r_x=1}^{W} \sum_{r_y=1}^{H} r * (H^{'}_{c,j}(r)) \tag{3.22}$$

Važna značajka *soft-argmax* funkcije jest da umjesto dobivanja indeksa s maksimalnom vrijednošću, gradijenti imaju priliku putem algoritma propagacije unatrag propagirati natrag do toplinskih karti. Kako je *backbone* unaprijed treniran koristeći funkciju gubitka koja je različita od *soft-argmax* uveden je *inverse temperature* parametar $\alpha = 100$ tako da na početku treniranja *soft-argmax* daje izlaze bliže pozicijama maksimuma. Da bi se zaključila 3D pozicija dijelova tijela iz 2D procjene $x_{c,j}$ korištena

je linearna algebarska triangulacija. Ova metoda svodi pronalazak 3D koordinata dijelova tijela y_j na rješavanje preodređenog sustava jednadžbi (engl. *overdetermined system of equations*) za vektor 3D koordinata za dio tijela $\tilde{y}$ (jednažba 3.23).

$$A_j * \tilde{y_j} = 0 \tag{3.23}$$

gdje A_j je matrica sastavljena od komponenata matrica projekcija i $x_{c,j}$. Naivni triangulacijski algoritam pretpostavlja da koordinate dijelova tijela svakog pogleda su neovisne te da svi dijelovi daju jednako značajne doprinose za triangulaciju. Međutim u nekim pogledima 2D pozicija se ne može pouzdano procijeniti, dovodeći do nepotrebnih degradacija finalne triangulacije. Ovaj problem se može riješiti primjenjujući metodu konsenzusa slučajnog uzorka (engl. *random sample consensus*, odnosno *RANSAC*) s Huber gubitkom (engl. *Huber loss*). Međutim ovo ima svoje nedostatke. Primjerice *RANSAC* može potpuno odcijepiti protok gradijenta za kamere koje nemaju doprinos. Da bi adresirali ovaj problem dodane su težine w_c koeficijentima matrice koje odgovaraju različitim pogledima (jednadžba (3.24)).

$$(w_j \circ A_j)\tilde{y}_j = 0 \tag{3.24}$$

gdje $w_j = (w_{1,j}, w_{1,j}, w_{2,j}, w_{2,j}, ..., w_{C,j}, w_{C,j})$ a, $\circ$ je Hadamarov umnožak. Težine $w_{c,j}$ su procijenjene konvolucijskom mrežom q^ϕ s parametrima ϕ (sastavljene od 2 konvolucijska sloja, jednog sloja sažimanja i 3 potpuno povezana sloja) (jednadžba (3.25)).

$$w_{c,j} = (q^\phi(g^\theta(I_c)))_j \tag{3.25}$$

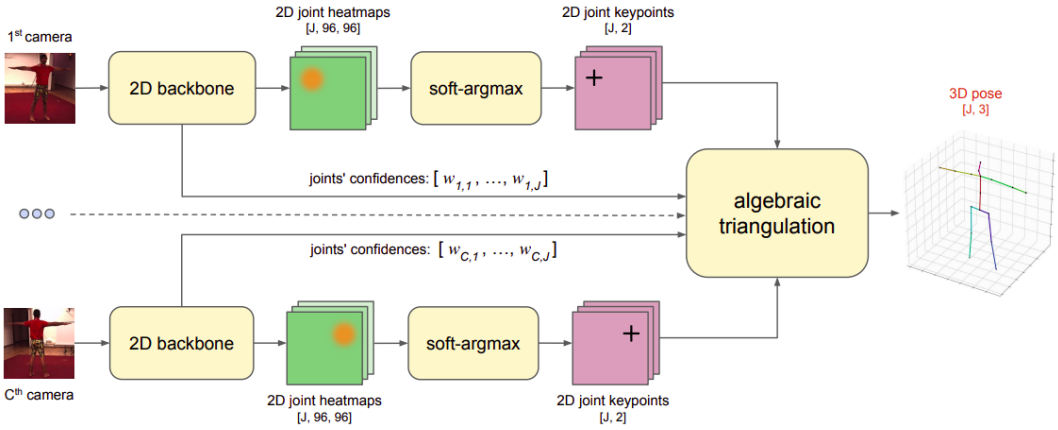
Neuronska mreža ovim putem ima kontrolu doprinosa svake kamere. Jednadžba (3.24) se rješava korištenjem diferencijalne dekompozicije matrice na singularne vrijednosti (engl. *Singular value Decomposition matrix*) (jednadžba (3.26)) gdje $\tilde{y}$ je postavljen kao zadnji stupac matrice V. Konačna nehomogena vrijednost za y je dobivena dijeljenjem vektora 3D homogenih koordinata $\tilde{y}$ s njegovom 4. koordinatom (jednadžba (3.27)).

$$B = UDV^T \tag{3.26}$$

$$y = \tilde{y}/(\tilde{y})_4 \tag{3.27}$$

Pristup volumetrijske triangulacije

Glavni nedostatak algebarske triangulacije jest da slike I_c iz različitih kamera su procesirane neovisno jedna od druge. Znači da ne postoji jednostavan način dodavanja



Slika 3.23: Pristup algebarske triangulacije

pose prior ni filtriranja kamera s pogrešnim matricama projekcije.

Da bi riješili ovaj problem pristup koristi kompleksniju triangulacijsku proceduru (slika 3.24). Koristi se *unprojection* procedura nad mapama značajki kreiranih iz 2D *backbone* bloka. Ovo je napravljeno "punjenjem" (engl. *filling*) 3D kocke putem projekcije izlaza 2D mreže uz projekcijske zrake unutar 3D kocke (vizualno prikazano slikom 3.22, procedura je ista kao i za prethodno opisan rad *LMT*). Kocke su dobivene iz više pogleda te su potom združene i procesirane. Za ovakav pristup volumenske triangulacije 2D izlaz ne mora biti interpretabilan poput *joint heatmaps* (izlazi *backbone* dijela pristupa algebarske triangulacije) te umjesto H_c vrijednosti za *unprojection* tehinku korišten je izlaz sloja konvolucijske neuronske mreže o^γ s 1×1 jezgrom i K izlaznih kanala (težine ovog sloja su označene s γ) (jednadžba (3.28)).

$$M_{c,k} = o^\gamma(f^\theta(I_c))_k \tag{3.28}$$

Za kreiranje volumetrijske rešetke (engl. *volumetric grid*) (potrebna za *unprojection* tehinku), postavljena je 3D kocka veličine $L \times L \times L$ čije središte koordinatnog sustava je postavljeno na dio tijela koji predstavlja zdjelicu (pozicija zdjelice je procijenjena koristeći algebarsku triangulaciju, L je duljina brida kocke u metrima). Koordinatni sustav kocke ima Y os okomitu na tlo i nasumičnu orijentaciju X osi.

Za svaki pogled projektirane su 3D koordinate kocke V^{coords} na ravninu, $V_c^{proj} = P_c V^{coords}$. Prijašnje spomenuto "punjenje" kocke je zapravo bilinearno uzorkovanje mapa $M_{c,k}$ odgovarajućih pogleda koristeći 2D koordinate iz V_c^{proj}. Formalno prikazano jednadžbom (3.29).

$$V_{c,k}^{view} = M_{c,k}\{V_c^{proj}\} \tag{3.29}$$

gdje $\{.\}$ označava bilinearno uzorkovanje. Onda volumetrijske mape svih pogleda se združuju kako bi se formirao ulaz sljedećim procesima koji je neovisan o broju pogleda. Postoje 3 načina putem kojih možemo združiti vrijednosti:

1. Izravna sumacija voksel podataka (jednadžba (3.30)).

$$V_k^{input} = \sum_c V_{c,k}^{view} \tag{3.30}$$

2. Sumacija voksel podataka s normaliziranim multiplikatorima d_c (jednadžba (3.31)):

$$V_k^{input} = \sum_c d_c V_{c,k}^{view} / \sum_c d_c \tag{3.31}$$

3. Računanje opuštene verzije maksimuma (engl. *relaxed version of maximum*). Prvo je izračunat *softmax* svakog voksela V_c^{view} svih kamera, kreirajući distribuciju volumetrijskih koeficijenta $V_{c,k}^w$. Distribucija bi imala sličnu ulogu kao i d_c parametar iz prethodnog načina (jednadžba (3.32)).

$$V_{c,k}^w = \exp(V_{c,k}^{view}) / \sum_c \exp(V_{c,k}^{view}) \tag{3.32}$$

   Onda voksel mape iz svakog pogleda sumiramo s volumetrijskim koeficijentom V_c^w, (jednadžba (3.33)).

$$V_k^{input} = \sum V_{c,k}^w \circ V_c^{view} \tag{3.33}$$

Združene volumetrijske mape su predane volumetrijskoj konvolucijskoj mreži (engl. *volumetric convolutional neural network*) u_ν (težine su ν) kreirajući interpretabilne 3D toplinske karte izlaznih dijelova tijela, (jednadžba (3.34)):

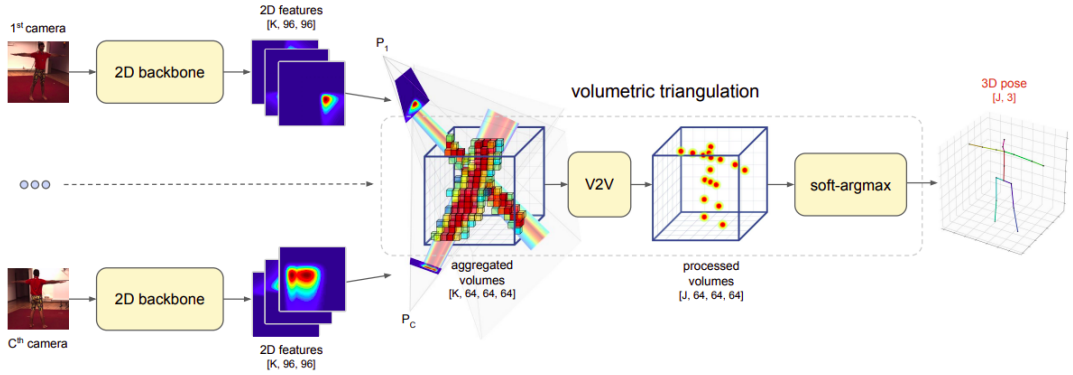
$$V_j^{output} = (u^\nu(V^{input}))_j \tag{3.34}$$

Slijedi računanje *softmax* vrijednosti za V_j^{output}, (jednadžba (3.35), W, H, D označavaju širinu, visinu i dubinu):

$$V_j^{'output} = \exp(V_j^{output}) / (\sum_{r_x=1}^{W} \sum_{r_y=1}^{H} \sum_{r_z=1}^{D} \exp(V_j^{output}(r))) \tag{3.35}$$

i procjenjujemo centar mase za svaki volumetrijski *joint heatmaps* da zaključimo poziciju dijelova u 3D (jednadžba (3.36)). Ovaj dio je prikazan na slici 3.24 blokom *soft-argmax*, pa nije ni čudno da je ovo zapravo ista jednadžba kao i (3.22) samo što se razlikuju dimenzije (ovdje je 3D dok u ranijoj jednadžbi je 2D).

$$y_j = \sum_{r_x=1}^{W} \sum_{r_y=1}^{H} \sum_{r_z=1}^{D} r * V_j^{'output}(r) \tag{3.36}$$



Slika 3.24: Pristup volumetrijske triangulacije

Podizanje na 3D dozvoljava dobivanje više robusnih rezultata, kako loše predikcije su prostorno izolirane od točnih unutar kocke te mogu biti odbačene koristeći konvolucijske operacije. Mreža također inkorporira parametre kamere i dopušta modeliranje *pose prior*.

Funkcije gubitka

Za obje metode (algebarsku i volumetrijsku) gradijenti prolaze od izlaznih predikcija 3D koordinata y_j do ulaznih slika što čini mrežu *end-to-end*.

Za slučaj algebarske triangulacije korištena je *soft* verzija *per-joint mean square error*, odnosno MSE, gubitaka da bi napravili treniranje robusnijim za iznimke. Ova varijanta dovodi do boljih rezultat nakon usporedne s običnim MSE i L1 gubitkom (jednadžba (3.37)).

$$L_j^{alg}(\theta, \phi) = \begin{cases} MSE(y_j, y_j^{gt}) & , ako: \ MSE(y_j, y_j^{gt}) < \varepsilon \\ MSE(y_j, y_j^{gt})^{0.1} * \varepsilon^{0.9} & , inae \end{cases} \tag{3.37}$$

Ovdje ε označava prag za funkciju gubitka koji je postavljen na $(20cm)^2$. Konačni gubitak je prosječna vrijednost svakog dijela tijela i svakog pogleda unutar *batch*.

Za volumetrijsku triangulaciju korišten je L1 gubitak sa *weak heatmap* regularizatorom. On maksimizira predikcije za voksel koji u sebi ima točni dio tijela (jednadžba (3.38)).

$$L^{vol}(\theta, \gamma, \nu) = \sum_j |y_j - y_j^{gt}| - \beta * \log(V_j^{output}(y_j^{gt})) \tag{3.38}$$

Bez drugog izraza u jednadžbi (3.38) neki bi dijelovi tijela kreirali volumetrijske toplinske karte koje nisu interpretabilne. Najvjerojatnije razlog tome jest nedovoljna veličina skupa za treniranje. Stavljanje β na malu vrijednost kao npr $\beta = 0.01$ čini

ih interpretabilnim. Također mali β nema nikakvog utjecaja na konačnu metriku te ga je moguće izostaviti ako interpretabilnost je nepotrebna. Također je testiran gubitak iz algebarskog pristupa umjesto L1 ali davao je lošije rezultate.

4. Usporedba standardnog pristupa i pristupa dubokog učenja

Nakon pregleda dvaju pristupa možemo vidjeti sličnosti i razlike među njima. Većina radova se oslanja na prijašnje radove koji su rješavali isti problem. Njihova sličnost proizlazi iz toga da su novonastali radovi poboljšanja prijašnjih metoda i time očuvaju korisne metode iz prijašnjih. Dok razlika je uočljiva u načinu kako dolaze do novijih te ujedno i boljih rezultata. To vidimo u samoj logici pristupa, gdje dubokim modelima se nastoji poboljšati proces učenja davajući im bolje podatke i kompleksnije arhitekture, dok klasični pristup pokušava kreirati algoritme koji bi trebali uspješno određivati ljudske poze na osnovu postojećih slika iz skupa. Problem kod klasičnih pristupa jest da ti algoritmi postaju previše ovisni o tim podacima odnosno bilo kakvi slučajevi koji nisu pokriveni tim algoritmom mogu rezultirati u katastrofalnoj procjeni. Duboki modeli se ipak lakše snalaze s novim podacima zbog svoje mogućnosti razumijevanja.

Također postoje i razlike unutar samih metoda koje koriste duboke učenje. Razlike su u tipovima arhitektura, funkcijama gubitka, tipu ulaznih podataka i sl. Isti razlozi za razlike između tih metoda su i njihove sličnosti. Kako dosta metoda se nastavlja na prijašnje radove onda je sasvim logično koristiti iste arhitekture te poboljšati rezultate na drugi način. Primjer ovoga su metode koje ne pokušavaju dobivati 2D lokacije dijelova tijela već koriste gotove 2D detektore i fokusiraju se na implementaciju pretvaranja iz 2D koordinata u 3D koordinate.

5. Implementacija i evaluacija metoda dubokog učenja

5.1. Priprema skupa podataka

Human3.6M [40] je skup podataka koji se sastoji od snimaka 5 ženskih i 6 muških osoba koje rade pojedine radnje kojih ukupno ima 15. Popis radnji i njihovih skraćenica koje će biti korištene kao nazivi stupaca za naredne tablice (do tablice 5.2 do 5.9) su prikazani u tablici 5.1. Skraćenice su upotrebljene zbog čitkosti stupaca (poželjan je kraći naziv stupaca zbog velikog broja radnji, što nije moguće koristeći hrvatske nazive) i lakše razumljivosti (engleske skraćenice jasno govore o kojoj je radnji riječ).

Snimke su snimljene iz 4 različita pogleda odnosno koristeći 4 različite kamere. Ukupno skup se sastoji od 3.6 milijuna ljudskih poza u 3D i njihovih korespondirajućih slika. Subjekti S1, S5, S6, S7, S8 su korišteni u trening skupu, dok S9 i S11 su korišteni kao validacijski skup.

Prije samog korištenja podataka potrebno je odraditi obradu podataka (engl. *preprocessing*). Kako je sami skup podataka jako velik (oko 260 GB) potrebno je osigurati dovoljno memorije prilikom dobivanja podataka da se potrebni podaci kreiraju kako bi se skup mogao koristiti. Postoji mnogo načina pripreme podataka ali u ovom radu je korišten *GitHub* repozitorij [37]. Unutar repozitorija je objašnjena procedura koja se sastoji do pozivanja 3 metode: *download_all.py*, *extract_all.py* i *process_all.py*. Nakon pozivanja metoda dobivamo skup podataka koji kasnije koristimo u eksperimentima.

5.2. Learnable Triangulation of Human Pose

Ovaj rad je ranije opisan u prijašnjim poglavljima i sastoji se od dvije metode koje su jako uspješne u određivanju 3D ljudskih poza. To su pristup algebarske triangulacije (engl. *Algebraic triangulation*) i pristup volumetrijske triangulacije (engl. *Volumetric*

Naziv radnje	Engleski naziv	Skraćenica
Davati upute	*Directions*	*Dir.*
Raspravljati	*Discussion*	*Disc.*
Jesti	*Eating*	*Eat*
Radnje tijekom sjedenja na tlu	*Activities while seated*	*SitD.*
Pozdravljati	*Greeting*	*Greet*
Fotografirati	*Taking photo*	*Photo*
Pozirati	*Posing*	*Pose*
Kupovati	*Making purchases*	*Purch.*
Pušiti cigare	*Smoking*	*Smoke*
Čekati	*Waiting*	*Wait*
Hodati	*Walking*	*Walk*
Sjediti na stolici	*Sitting on chair*	*Sit*
Razgovarati na mobitel	*Talking on the phone*	*Phone*
Šetati psa	*Walking dog*	*WalkD.*
Hodati zajedno	*Walking together*	*WalkT.*

Tablica 5.1: Prikaz radnji koje su se koristile za kreiranje Human3.6M skupa podataka

triangulation). Eksperimenti se sastoje od evaluacije metoda promjenom ulaznih podataka i arhitekture.

Promjenu ulaznih parametara postižemo koristeći izobličene slike (engl. *distorted images*). *Undistort* slike, odnosno slike koje su korištene tijekom treniranja, se dobivaju koristeći *grid-sampling*. Tehnički govoreći, ova procedura je ostvarena koristeći metodu *cv2.remap* iz *OpenCV* biblioteke. Metoda radi proces uzimanja piksela iz jedne lokacije unutar slike i postavljanja tih piksela u drugu lokaciju u novoj slici.

Promjena arhitekture je ostvarena ne korištenjem *confidence* težina koje su naučene tijekom treniranja modela. Težine su smještene u 2D *backbone* dijelu mreže a izlaz tih težina se koristi prilikom algebarske triangulacije, što je vidljivo na Slika 3.23 (izlazi težina su imenovani *joints' confidences*, važno je primijetiti da izlazi dolaze iz 2D *backone* dijela i idu do *algebraic triangulation* bloka).

U nastavku slijedi opis eksperimenata koji su odrađeni na spomenutim metodama.

5.2.1. Algebarska triangulacija

Pokrenuto je 5 eksperimenata koristeći pristup algebarske triangulacije. To su:

1. Koristeći *confidence* težine i korištenje neizobličenih slika (red Algebarska t. u Tablicama 5.2 i 5.3)

2. Ne koristeći *confidence* težine i korištenje neizobličenih slika (red Algebarska t. (*w/o conf*) u Tablicama 5.2 i 5.3)

3. Koristeći *confidence* težine i korištenje izobličenih slika (red Algebarska t. (*w/o undistort images*) u Tablicama 5.2 i 5.3)

4. Ne koristeći *confidence* težine i korištenje izobličenih slika (red Algebarska t. (*w/o undistort and conf*) u Tablicama 5.2 i 5.3)

5. Koristeći filtrirane podatke (red Algebarska t. (filtrirani podaci) u Tablici 5.3)

Protokol 1 (relativno sa zdjelicom)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Algebarska t. (službeni rezultati)	20.4	**22.6**	20.5	19.7	22.1	**20.6**	19.5	**23.0**	**25.8**	33.0	23.0	21.6	**20.7**	23.7	21.3	22.6
Algebarska t.	**18.97**	22.76	**19.98**	**19.49**	**21.72**	20.69	**19.11**	22.39	26.1	**31.81**	**22.85**	**20.94**	23.5	**20.12**	**21.12**	**22.23**
Algebarska t. (w/o conf)	22.97	26.64	24.88	23.53	30.04	24.76	21.19	26.06	46.18	44.25	30.15	23.73	26.31	22.55	24.19	28.36
Algebarska t. (w/o undistort images)	19.57	23.16	20.82	20.06	22.77	21.72	19.67	22.81	26.33	31.87	23.52	21.73	24.24	21.84	22.83	22.96
Algebarska t. (w/o undistort images and conf)	23.41	27.14	25.8	24.14	30.83	25.99	21.72	26.64	53.92	44.35	30.82	24.47	27.23	24.1	26.15	29.66

Tablica 5.2: Evaluacija algebarske triangulacije na Human3.6M validacijskom skupu (pogreška relativna u odnosu na zjedlicu).

Protokol 1 (apsolutna pogreška)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Algebarska t. (službeni rezultati)	18.1	**20.0**	**17.6**	**17.0**	18.9	**19.3**	**17.4**	19.2	**21.9**	23.2	**19.5**	18.0	**18.3**	20.5	**17.9**	**19.2**
Algebarska t.	**17.52**	20.52	17.7	48.05	**18.75**	19.31	17.5	**19.12**	22.07	57.45	19.59	46.37	20.68	**18.12**	17.98	25.02
Algebarska t. (w/o conf)	22.04	24.9	23.46	51.1	27.82	23.88	20.34	22.76	43.36	68.49	28.06	49.43	24.39	21.95	22.8	31.74
Algebarska t. (w/o undistort images)	18.04	21.8	19.35	49.52	21.63	21.52	17.85	19.81	22.67	57.49	21.49	47.05	22.25	22.49	22.88	26.69
Algebarska t. (w/o undistort images and conf)	22.02	25.98	25.21	52.25	30.22	25.95	20.34	23.41	51.35	68.49	29.59	49.61	26.44	26.18	27.65	33.74
Algebarska t. (filtrirani podaci)	**17.52**	20.52	17.7	17.07	**18.75**	19.31	17.51	**19.12**	22.08	**22.7**	19.6	**17.71**	20.68	**18.12**	17.98	19.22

Tablica 5.3: Evaluacija algebarske triangulacije na Human3.6M validacijskom skupu (apsolutna pogreška).

U Tablici 5.2 zanimljivo je primijetiti da službeni rezultati i implementirano rješenje se razlikuju u pojedinim kategorijama čak za više od 1 mm. Možemo uočiti da *Average* rezultat implementiranog je bolji od službenog rezultata za 0.37mm. Premda je mala razlika u tablici 5.2 ali uzimajući u obzir da razlika između najboljeg i drugog najboljeg modela na [25] 0.11 mm, možemo reći da za procjenu ljudske poze ovo značajna razlika. Razlog ovog rezultata može biti razlika u težinama korišteni za dobivanje službenih rezultata i korištenih težina.

Rezultati eksperimenta koji ne koristi *confidence* težine (red Algebarska t. (*w/o conf*) u tablici 5.2) su lošiji od službenih rezultata i vlastite implementacije koje ih koriste, što je vidljivo u tablici 5.2 (važno je primijetiti razliku reda Algebarska t. (*w/o conf*) i redova Algebarska t. (službeni rezultati) i Algebarska t.)
Eksperiment koji koristi izobličene slike ima lošije rezultate od prva dva eksperimenta (prvi eksperiment koristi neizobličene slike i *confidence* težine, drugi koristi neizobličene slike ali ne koristi *confidence* težine) zbog razloga što model na ulazu dobiva izobličene slike te time se otežava proces procjene poze.
Završni eksperiment tablice 5.1 (red Algebarska t. (*w/o undistort and conf*)) jest evaluacija modela koristeći izobličene slike i ne koristeći *confidence* težine. Naravno ovaj rezultat kao što je i očekivano daje najgore rezultate. Ali pogreška je samo za 1 mm veća od eksperimenta koji nije koristio *confidence* težine. Time zaključujemo da model je više ovisan o tim težinama ,nego o tome jesu li ulazni podaci izobličeni, i da je ih nužno koristiti prilikom procjene.
Kod vlastitih eksperimenata model najbolje procjenjuje poze iz klase *Dir.* (najmanja MPJPE pogreška 18.97) dok najteže procjenjuje poze iz klase *SitD.* (najmanja MPJPE pogreška 31.81). Iz službenih rezultata možemo vidjeti da model najbolje procjenjuje poze iz klase *Pose* (najmanja MPJPE pogreška 19.5) dok najteže procjenjuje poze iz klase *SitD.* (najmanja MPJPE pogreška 33.0)

U tablici 5.3 su prikazani rezultati apsolutne MPJPE pogreške za 5 eksperimenata. Kao i kod prošle tablice zanimljivo je vidjeti da službeni rezultati nisu nužno bolji od vlastito implementiranog rješenja. Već spomenuti razlog ove razlike je također ovdje primjenjiv. Bitno je naglasiti da službena rješenja za ovu vrstu evaluacije koriste filtrirane podatke. Tako da stvarnu usporedbu službenih rezultata i rezultata implementiranog rješenja možemo vidjeti u razlici između drugog i zadnjeg retka tablice 5.3. (usporedba redova Algebarska t. (službeni rezultati) i Algebarska t. (filtrirani podaci))
Za razliku od prijašnje tablice ovdje prosječno službeno rješenje, odnosno *Avg.* stupac, je bolje za 0.02 mm.
Što se tiče utjecaja izobličenih slika i ne primjenjivanja *confidence* težina isti zaključak vrijedi kao i za prijašnju tablicu.
Filtrirani podaci koji su korišteni za evaluaciju modela izbacuju scene koje imaju pogreške u *ground-truth* podacima. Za validacijski skup subjekta S9 pojedine kategorije imaju podatke koje u pogrešno pomaknute u 3D prostor uspoređujući ih sa stvarnom pozicijom. Tu razliku možemo primijetiti kod rezultata 3. i 7. retka tablice 5.3 za

stupce *Greet*, *SitD.* i *Wait* (redovi Algebarska t. i Algebarska t. (filtrirani podaci)). Kod vlastitih eksperimenata model najbolje procjenjuje poze iz klase *Greet* (najmanja MPJPE pogreška 17.07) dok najteže procjenjuje poze iz klase *Sit* (najmanja MPJPE pogreška 22.07). Iz službenih rezultata možemo vidjeti da model najbolje procjenjuje poze iz klase *Greet* (najmanja MPJPE pogreška 17.0) dok najteže procjenjuje poze iz klase *SitD.* (najmanja MPJPE pogreška 23.2)

U sljedećim tablicama prikazane su relativne i apsolutne MPJPE pogreške subjekata S9 i S11, odnosno subjekata koji su korišteni u evaluacijskom skupu.

Protokol 1 (relativno sa zdjelicom)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Subjekt S9	21.36	26.17	23.5	21.32	23.13	22.17	20.9	23.3	29.18	40.84	24.41	24.96	25.0	20.64	21.63	24.85
Subjekt S11	**15.45**	**15.02**	**15.81**	**17.32**	**20.26**	**19.14**	**16.71**	**21.21**	**21.6**	**21.29**	**20.24**	**16.59**	**21.0**	**19.48**	**20.58**	**18.7**

Tablica 5.4: Relativna MPJPE pogreška S9 i S11 subjekata

Protokol 2 (apsolutna pogreška)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Subjekt S9	20.46	23.39	20.32	75.51	20.66	21.25	19.78	20.63	25.44	90.57	21.26	75.85	22.88	20.57	20.58	31.82
Subjekt S11	**13.17**	**13.99**	**14.6**	**15.68**	**16.77**	**17.26**	**14.44**	**17.17**	**17.16**	**18.9**	**16.8**	**14.43**	**17.0**	**15.05**	**15.23**	**15.85**

Tablica 5.5: Apsolutna MPJPE pogreška S9 i S11 subjekata

Kao što možemo primjetiti da subjekt S9 je teži za procijeniti od S11 subjekta za svaku radnju u obje tablice.

5.2.2. Volumetrijska triangulacija

Pokrenuta su 2 eksperimenta koristeći pristup volumetrijske triangulacije. To su:

1. Korištenje neizobličenih slika (red Volumetrijska t. u Tablicama 5.3 i 5.4)
2. Korištenje izobličenih slika (red Volumetrijska t. *(w/o undistort images)* u Tablicama 5.3 i 5.4)

Protokol 1 (relativno sa zdjelicom)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Volumetrijska t. (službeni rezultati)	18.8	20.0	**19.3**	18.7	20.2	**19.3**	18.7	22.3	23.3	**29.1**	21.2	20.3	**19.3**	21.6	19.8	20.8
Volumetrijska t.	**17.36**	**19.63**	19.44	**18.36**	**19.96**	19.36	**17.79**	**20.68**	**23.28**	29.39	**20.58**	**19.42**	21.16	**18.68**	**19.15**	**20.32**
Volumetrijska t. (w/o undistort images)	18.52	20.5	20.31	19.61	21.41	20.71	18.58	21.54	23.41	29.5	21.73	20.62	22.05	21.21	21.97	21.44

Tablica 5.6: Evaluacija volumetrijske triangulacije na Human3.6M validacijskom skupu (pogreška relativna u odnosu na zjedlicu)

U tablici 5.6 prikazane su relativne MPJPE pogreške za 2 eksperimenta i službeni rezultati rada. Kao i do sada primjećujemo razlike između vlastite implementacije i

Protokol 1 (apsolutna pogreška)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Volumetrijska t. (službeni rezultati)	16.9	**18.1**	**16.6**	**16.0**	**17.1**	**17.9**	**16.5**	18.5	**19.6**	**20.1**	**18.2**	**16.8**	**17.2**	19.0	**16.6**	**17.7**
Volumetrijska t.	**16.64**	18.35	16.9	47.55	17.49	18.27	**16.5**	**17.78**	20.07	55.84	18.32	45.81	19.0	**17.21**	16.75	23.75
Volumetrijska t. (w/o undistort images)	17.23	19.76	18.58	49.16	20.34	20.33	17.21	18.78	20.48	55.87	20.23	46.56	20.75	21.6	21.87	25.47

Tablica 5.7: Evaluacija volumetrijske triangulacije na Human3.6M validacijskom skupu (apsolutna pogreška)

službenih rezultata (redovi Volumetrijska t. (službeni rezultati) i Volumetrijska t. u tablici 5.6). U ovoj metodi testiranje obavljamo koristeći izobličene ulaze. Testiranjem je ustanovljen isti učinak kao i kod algebarskog pristupa, dajući lošije rezultate uporabom izobličenih slika (razlika između reda Volumetrijska t. (w/o undistorted images) i redova Volumetrijska t. (službeni rezultati) i Volumetrijska t.).

Kod vlastitih eksperimenata model najbolje procjenjuje poze iz klase *Dir.* (najmanja MPJPE pogreška 17.36) dok najteže procjenjuje poze iz klase *Sit* (najmanja MPJPE pogreška 23.28). Iz službenih rezultata možemo vidjeti da model najbolje procjenjuje poze iz klasa *Greet* i *Pose* (najmanja MPJPE pogreška 18.7) dok najteže procjenjuje poze iz klase *SitD.* (najmanja MPJPE pogreška 23.3)

U tablici 5.7 prikazane su apsolutne MPJPE pogreške za 2 eksperimenta i službeni rezultati rada. Najbolju prosječnu vrijednost ima službeni rad koji je koristio filtrirane podatke zbog nepravilnih *ground-truth* podataka koji su objašnjeni prije. Ovaj put nije bilo moguće evaluirati podatke nad filtriranim podacima, jer potrebni podaci za predviđene 3D dijelove tijela nisu javno dostupni.

Kod vlastitih eksperimenata model najbolje procjenjuje poze iz klase *Pose* (najmanja MPJPE pogreška 16.5) dok najteže procjenjuje poze iz klase *SitD.* (najmanja MPJPE pogreška 55.84). Iz službenih rezultata možemo vidjeti da model najbolje procjenjuje poze iz klase *Greet* (najmanja MPJPE pogreška 16.0) dok najteže procjenjuje poze iz klase *SitD.* (najmanja MPJPE pogreška 20.1).

Protocol 1 (relativna pogreška)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Volumetrijska t. S9	19.66	22.43	22.37	20.82	21.6	21.34	19.87	22.2	26.19	38.79	22.6	23.57	23.02	19.95	20.44	23.02
Volumetrijska t. S11	**13.95**	**13.28**	**15.98**	**15.47**	**18.26**	**17.28**	**14.99**	**18.73**	**19.02**	**18.44**	**17.21**	**14.91**	**18.06**	**17.09**	**17.78**	**16.66**

Tablica 5.8: Relativna MPJPE pogreška S9 i S11 subjekata

Protocol 1 (apsolutna pogreška)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Volumetrijska t. S9	19.5	20.79	19.33	75.55	19.74	20.75	18.71	19.5	23.56	89.83	20.3	75.85	21.54	19.95	19.44	30.69
Volumetrijska t. S11	**12.4**	**12.82**	**14.03**	**14.54**	**15.16**	**15.65**	**13.52**	**15.55**	**14.97**	**16.27**	**15.01**	**13.27**	**14.75**	**13.78**	**13.9**	**14.38**

Tablica 5.9: Apsolutna MPJPE pogreška S9 i S11 subjekata

Kao i za slučaj algebarske triangulacije primjećujemo da S9 je teži bio za procijeniti nego S11.

5.2.3. Potencijalne ideje za buduće radove

Tijekom rada na ovom zadatku pokušaj implementacije novog 2D backbone dijela je bio cilj. Ideja je bila zamjenjivanje treniranog 2D *backbone* dijela u algebarskim i volumetrijskim triangulacijama s postojećim modelom koji ima dosta impresivne rezultate u procjenama 2D ljudskih poza, *fine-tuning* novonastalne mreže i usporedba dobivenih rezultata sa službenim rezultatima. Za vlastiti pokušaj upotrijebljen je *TransPose*. Osim svojih impresivnih rezultata, razlog odabira jest njegova dostupnost u *torch.hub* repozitoriju. No zbog nedovoljno računalnih resursa koji su bili potrebni da se ostvari ovaj zadatak, spomenutu ideju nudim kao prijedlog za buduća istraživanja.

Prije svega potrebno je osigurati da imamo dovoljno memorije da spremimo podatke potrebnih za treniranje i evaluiranje modela. U poglavlju prije sam naveo da Human3.6M skup podataka ukupno sadrži 260 GB. Nakon pre-processing novonastali podaci će zauzimati 162 GB. Znači prije nego što započne se rješavanja zadatka potrebno je osigurati minimalno 422 GB memorije. Ovaj problem se može riješiti postavljanjem Human3.6M skupa podataka na vanjski tvrdi disk te nakon obrade podataka memorijsko zauzeće će iznositi samo 162 GB.

Zatim je potrebno osigurati da oprema (tj. laptop ili računalo) s kojom se trenira ima dovoljno RAM-a. Za evaluaciju modela je potrebno je oko 4 GB, dok za treniranje je potrebno više od toga. Za evaluaciju *batch size* je 4, dok bilo koji drugi *batch size* zahtjeva količinu RAM-a veću od 4 GB.

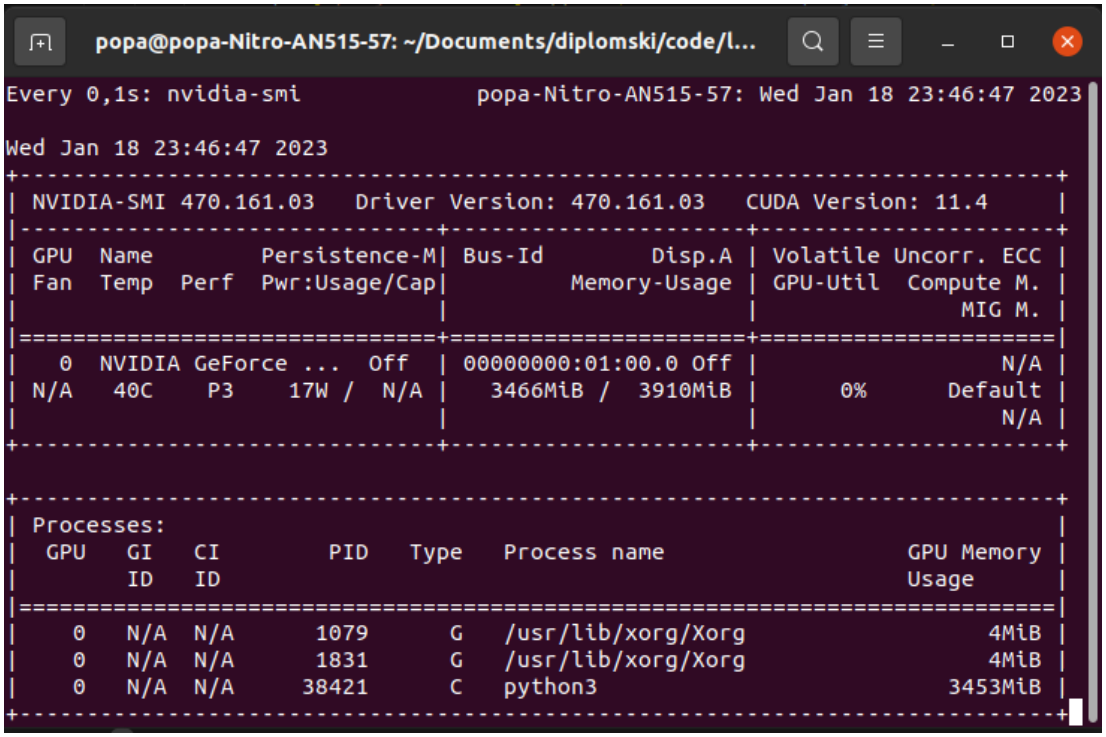
Čak i nakon smanjivanja skupa podataka koji se koristi za treniranje i dalje nije bilo dovoljno memorije, što znači da glavni problem u memoriji kod treniranja modela su optimizatori i memorija potreba za spremanje gradijenta.

Google Colab je uzet kao opcija koristeći manji skup *Human3.6M* podataka, odnosno samo S1 subjekt. No ostali dio *preprocessing* dijela nije uspio se uspješno izvršiti te *Colab* se pokazao kao još jedan neuspjeli pokušaj.

Konačno je zaključeno da je jedini način implementiranja ove ideje korištenje računala koje ima dovoljno *hard-disk* memorije i RAM-a da bi se izvršio *fine-tuning*.

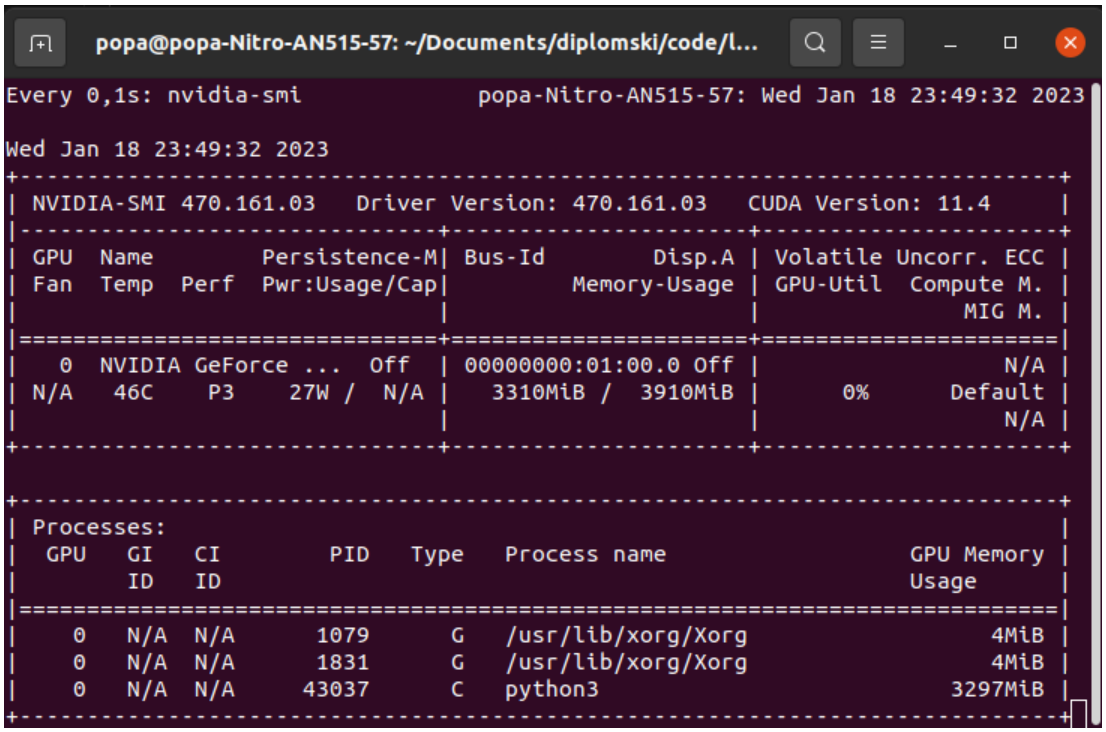
Za računanje potrebne radne memorije moramo uzeti u obzir:

- Memoriju za parametre mreže
- Memoriju za izlaz iz međuslojeva
- Memoriju za gradijent svakog parametra
- Memoriju za optimizatore
- Memoriju za implementaciju (engl. *miscellaneous memory*)



Slika 5.1: Memorija korištena za evaluiranje pristupa algebarske triangulacije, primjetiti polje u kojem piše 3466 MiB/ 3910 MiB

Nakon računanja za potrebe treniranja ove mreže ustanovljeno je da je potrebno 2.37 GB memorije samo za model. Još je potrebno nadodati memoriju za optimizatore, što je u ovom slučaju *AdamW*, i memoriju za implementaciju. Memorija potrebna za evaluaciju modela korištenih u opisanim eksperimentima je prikazana na slikama 5.1 i 5.2.



Slika 5.2: Memorija korištena za evaluiranje pristupa volumetrijske triangulacije, primjetiti polje u kojem piše 3310 MiB/ 3910 MiB

6. Zaključak

U ovom diplomskom radu istražene su metode za procjenu ljudske poze za 2D i 3D. Istražene su i opisane metode koristeći klasične pristupe te pristupe dubokog učenje, potom njihova usporedba. Detaljnije su opisane metode dubokog učenja zbog njihove šire primjenjivosti za rješavanje problema procjene ljudske poze. Podjela tih metoda je napravljena po vrsti pogleda, odnosno na monokularne i pristupe s više pogleda. Unutar tih podjela su napravljene i dodatne podjele kako bi se naglasila njihova razlika. Razlike su bile prikazane opisom različitih tipova mreža, odnosno konvolucijskih mreža, graf mreža te mehanizama pozornosti. Potom su istaknute neke od najboljih metoda iz svake skupine i opisane detaljno. Nakon toga se prikazuju implementacijski rezultati tih metoda, te njihova međusobna usporedba ali i usporedba sa službenim rezultatima. Krajnje je dan naglasak na mogući nastavak rada i koraci koji se moraju uzeti u obzir prilikom rješavanja ovih problema.

LITERATURA

[1] Xia, Hailun, and Tianyang Zhang. "Self-Attention Network for Human Pose Estimation." Applied Sciences 11.4 (2021): 1826.
[2] Lin, Kevin, Lijuan Wang, and Zicheng Liu. "Mesh graphormer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[3] Liu, Ruixu, et al. "Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[4] Chen, Tianlang, et al. "Anatomy-aware 3d human pose estimation with bone-based pose decomposition." IEEE Transactions on Circuits and Systems for Video Technology 32.1 (2021): 198-209.
[5] Lin, Kevin, Lijuan Wang, and Zicheng Liu. "End-to-end human pose and mesh reconstruction with transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[6] Tome, Denis, Chris Russell, and Lourdes Agapito. "Lifting from the deep: Convolutional 3d pose estimation from a single image." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[7] Li, Sijin, Weichen Zhang, and Antoni B. Chan. "Maximum-margin structured learning with deep networks for 3d human pose estimation." Proceedings of the IEEE international conference on computer vision. 2015.
[8] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[9] Hu, Wenbo, et al. "Conditional directed graph convolution for 3d human pose estimation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.
[10] Zheng, Ce, et al. "3d human pose estimation with spatial and temporal transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[11] Sun, Xiao, et al. "Integral human pose regression." Proceedings of the European conference on computer vision (ECCV). 2018.
[12] Pavlakos, Georgios, et al. "Coarse-to-fine volumetric prediction for single-image 3D human pose." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[13] Wei, Shih-En, et al. "Convolutional pose machines." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016.
[14] Zheng, Ce, et al. "3d human pose estimation with spatial and temporal transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[15] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.
[16] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[17] Chen, Yilun, et al. "Cascaded pyramid network for multi-person pose estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[18] Martinez, Julieta, et al. "A simple yet effective baseline for 3d human pose estimation." Proceedings of the IEEE international conference on computer vision. 2017.
[19] Cai, Yujun, et al. "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
[20] Shi, Lei, et al. "Skeleton-based action recognition with directed graph neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[21] Biovision BVH, https://research.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/BVH.html, datum pristupanja: 20.10.2022.
[22] Pavlakos, Georgios, et al. "Harvesting multiple views for marker-less 3d human pose annotations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[23] Joo, Hanbyul, et al. "Panoptic studio: A massively multiview system for social motion capture." Proceedings of the IEEE International Conference on Computer Vision. 2015. [24] Zhang, Jinlu, et al. "MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[25] 3D Human Pose Estimation on Human3.6M, https://paperswithcode.com/sota/3d-human-pose-estimation-on-human36m, datum pristupanja: 20.10.2022.
[26] Chun, Sungho, Sungbum Park, and Ju Yong Chang. "Learnable human mesh triangulation for 3D human pose and shape estimation." arXiv preprint arXiv:2208.11251 (2022).
[27] Iskakov, Karim, et al. "Learnable triangulation of human pose." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[28] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[29] Tompson, Jonathan J., et al. "Joint training of a convolutional network and a graphical model for human pose estimation." Advances in neural information processing systems 27 (2014).
[30] Yang, Sen, et al. "Transpose: Keypoint localization via transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[31] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European conference on computer vision. Springer, Cham, 2016.
[32] He, Yihui, et al. "Epipolar transformers." Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2020.
[33] Rhodin, Helge, et al. "Learning monocular 3d human pose estimation from multi-view images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[34] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[35] Pavllo, Dario, et al. "3d human pose estimation in video with temporal convolutions and semi-supervised training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[36] Moon, Gyeongsik, and Kyoung Mu Lee. "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image." European Conference on Computer Vision. Springer, Cham, 2020.
[37] Human3.6M dataset fetcher, https://github.com/anibali/h36m-fetch, datum pristupanja: 2.1.2023.
[38] Xu, Yufei, et al. "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation." arXiv preprint arXiv:2204.12484 (2022).
[39] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and

state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014.
[40] Ionescu, Catalin, et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence 36.7 (2013): 1325-1339.
[41] Li, Yanghao, et al. "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[42] Andriluka, Mykhaylo, Stefan Roth, and Bernt Schiele. "Pictorial structures revisited: People detection and articulated pose estimation." 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009.
[43] Ionescu, Catalin, Joao Carreira, and Cristian Sminchisescu. "Iterated second-order label sensitive pooling for 3d human pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
[44] Sun, Min, and Silvio Savarese. "Articulated part-based model for joint object detection and pose estimation." 2011 International Conference on Computer Vision. IEEE, 2011.
[45] Tian, Yuandong, C. Lawrence Zitnick, and Srinivasa G. Narasimhan. "Exploring the spatial hierarchy of mixture models for human pose estimation." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.
[46] Dantone, Matthias, et al. "Human pose estimation using body parts dependent joint regressors." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
[47] Karlinsky, Leonid, and Shimon Ullman. "Using linking features in learning non-parametric part models." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.
[48] Ramakrishna, Varun, et al. "Pose machines: Articulated pose estimation via inference machines." European Conference on Computer Vision. Springer, Cham, 2014.
[49] Zisserman, A., I. D. Reid, and A. Criminisi. "Single view metrology." IEEE ICCV. 1999.
[50] Zhang, Ruo, et al. "Shape-from-shading: a survey." IEEE transactions on pattern analysis and machine intelligence 21.8 (1999): 690-706.
[51] Lindeberg, Tony, and Jonas Garding. "Shape from texture from a multi-scale perspective." 1993 (4th) International Conference on Computer Vision. IEEE, 1993.
[52] Roberts, Lawrence G. Machine perception of three-dimensional solids. Diss. Massachusetts Institute of Technology, 1963.

[54] Lee, Hsi-Jian, and Zen Chen. "Determination of 3D human body postures from a single view." Computer Vision, Graphics, and Image Processing 30.2 (1985): 148-168.
[55] Jiang, Hao. "3d human pose reconstruction using millions of exemplars." 2010 20th International Conference on Pattern Recognition. IEEE, 2010.
[56] Gupta, Ankur, et al. "3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[57] Akhter, Ijaz, and Michael J. Black. "Pose-conditioned joint angle limits for 3D human pose reconstruction." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[58] Bogo, Federica, et al. "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image." European conference on computer vision. Springer, Cham, 2016.
[59] A Comprehensive Guide to Human Pose Estimation, https://www.v7labs.com/blog/human-pose-estimation-guide, datum pristupanja: 2.3.2023.

METODE ZA PROCJENU LJUDSKE POZE

Sažetak

Svakodnevni susret sa sustavima koji koriste metode računalnog vida je neizbježan. Zahvaljujući dubokom učenju problemi računalnog vida ponovno postaju jako popularni. Među jednim od najzanimljivijih jest procjena ljudske poze u 2D/3D. U ovom radu istražene su te potom opisane metode standardnog (klasičnog) pristupa te pristupa koji koriste duboko učenje. Unutar te podjele naglašene su razlike njihovih pristupa i detaljno opisani postupci njihovih najboljih metoda. Implementirane su dvije metode za procjenu poze te je napravljena evaluacije provedenih eksperimenata i službenih rješenja.

Ključne riječi: procjena ljudske poze u 2D, procjena ljudske poze u 3D

Title

Abstract

Everyday encounter with systems that use computer vision methods is inevitable. Thanks to deep learning, computer vision problems are becoming very popular again. Among one of the most interesting is the estimation of human pose in 2D/3D. In this paper, the methods of the standard (classical) approach and approaches that use deep learning were investigated and then described. Within that division, the differences in their approaches are emphasized and the procedures of their best methods are described in detail. Two methods for pose estimation were implemented and an evaluation of the conducted experiments and official solutions was made.

Keywords: human pose estimation in 2D, human pose estimation in 3D