SVEUČILIŠTE U SPLITU
FAKULTET ELEKTROTEHNIKE, STROJARSTVA I BRODOGRADNJE
Poslijediplomski doktorski studij Elektrotehnike i informacijske tehnologije

Kvalifikacijski doktorski ispit

PREPOZNAVANJE LJUDSKIH AKTIVNOSTI

Ivan Đapić Split, 09.10. 2015

SADRŽAJ

1 UVOD 3
1.1 Općenito o prepoznavanju ljudskih aktivnosti 3
1.2 Područja primjene 4
1.3 Osnovne vrste aktivnosti 4
1.3.1 Gesta ili osnovna akcija 4
1.3.2 Akcija 4
1.3.3 Aktivnost 5
1.3.4 Interakcija 5
1.3.5 Grupna aktivnost ili aktivnost grupe 5
2 PREGLED PODRUČJA 6
2.1 Općenito o funkcionalnoj taksonomiji razvrstavanja 6
2.2 Inicijalizacija – prvi korak 7
2.2.1 Inicijalizacija kinematičkog modela kao glavna radnje ove faze obrade 7
2.2.1.1 Početna poza 7
2.2.1.2 Kinematički modeli 8
2.2.2 Trenutna dostignuća u inicijalizaciji 8
2.2.2.1 Kinematički modeli temeljeni na kinematičkoj strukturi 8
2.2.2.2 Kinematički modeli temeljnih na 3D oblicima 10
2.2.2.3 Kinematički modeli temeljeni na boji 10
2.2.2.4 Kinematički modeli temeljeni na dijelovima tijela 11
2.2.3 Trenutno otvoreni problemi u inicijalizaciji 13
2.3 Praćenje – drugi korak 13
2.3.1 Oduzimanje pozadine 16
2.3.1.1 Prezentacija pozadine 16
2.3.1.2 Klasifikacija pozadine 19
2.3.1.3 Ažuriranje pozadine 23
2.3.1.4 Inicijalizacija pozadine 23
2.3.2 Segmentacija temeljena na pokretu subjekta 25
2.3.3 Segmentacija temeljena na boji i obliku subjekta 33
2.3.3.1 Segmentacija temeljena na izgledu bez korištenja konteksta 34
2.3.3.2 Segmentacija temeljena na izgledu i obliku sa korištenjem konteksta 38
2.3.4 Segmentacija temeljena na dubini slike 42
2.3.5 Temporal Correspondences 42
2.3.5.1 Temporal correspondences prije i poslije preklapanja 43
2.3.5.2 Temporal correspondences tijekom preklapanja 45
2.3.6 Trenutno otvoreni problemi u praćenju 46
2.4 Estimacija poze – treći korak 47
2.4.1 Estimacija poze bez korištenja modela subjekta 48
2.4.2 Estimacija poze sa indirektnim korištenja modela subjekta 50
2.4.3 Estimacija poze sa direktnim korištenja modela subjekta 50
2.4.4 Trenutno otvoreni problemi u estimaciji poze 53
2.5 Prepoznavanje radnje – četvrti korak 54
2.5.1 Interpretacije scene 54
2.5.2 Holistički pristupi 55
2.5.3 Akcijske primitivi i gramatika 58
2.5.4 Neuralne mreže 58
2.5.5 Trenutno otvoreni problemi u prepoznavanju aktivnosti 59
2.6 Zaključak 60
3 LITERATURA 61

1 UVOD

1.1 OPĆENITO O PREPOZNAVANJU LJUDSKIH AKTIVNOSTI

Sintagma eng. *Human Activity Recognition* (**HAR**) prema [2] definirana je kao proces zaključivanja o aktivnosti 3D objekata sa video sekvenci koje karakterizira visoka stopa preklapanja objekta sa samim sobom ili drugim objektima i veliki broj spojnih točaka (tj. zglobova kod ljudi). HAR je grana znanosti koja spada u eng. *Computer Vision*-a (**CV** - a). U posljednjih nekoliko godina ova grana CV - a poprima sve veći interes, kako komercijalni tako i sveučilišni, zbog napretka u softveru i hardveru koji su omogućili širu i dublju primjenu HAR – a.

Prvi HAR algoritmi bili su relativno jednostavni jer su operirali direktno na video sekvenci čime nisu bili u stanju raditi analizu složenih radnji. Ovo je imalo za posljedicu da su jedine radnje (ili pokrete) koje su bili u stanju uspješno prepoznati spadale u skup jednostavnih radnji sa ponavljanjem poput šetnje ili mahanja.

Međutim, osmišljavanjem drugačijih, bitno kompliciranijih načina odlučivanja o vrsti aktivnosti sa video sekvence porastao je skup potencijalnih područja primjena HAR aplikacija. Taj „drugačiji“ način podrazumijevao je kompleksniji sustav ili okvir (eng. *framework*) zaključivanja tako što se prepoznavanje aktivnosti ne vrši više primjenom algoritma prepoznavanja direktno na video sekvenci nego se rade slojevi (eng. *layers*) obrade ulaznog video signala od kojih svaki ima zasebnu ulogu u sklopu sustava. Slojevi mogu biti serijski posloženi u vremenu obrade (jedan za drugim, sljedeći ne može početi s radom dok prethodni ne završi) ili pak više njih može raditi zajedno kako bi dali rezultat obrade potreban za sljedeći sloj ili korak obrade podataka, a dozvoljena je i isprepletena logika ova dva pristupa.

Također, HAR je jedina neinvazivna metoda aplikacije zaključivanja ili odlučivanja što znači da ne koristi dodatna „pomagala“ poput senzora, markera itd., a to joj daje niz prednosti u odnosu na invazivne sustave. Međutim, HAR ne isključuje rad sa dodatnim izvorima informacija poput već nabrojenih pomagala.

Ako se HAR definira rječnikom umjetne inteligencije i računalnog vida onda je HAR prema [5]: kategorizacija ljudskih aktivnosti u više različitih klasa pri čemu svaka snimljena aktivnosti ima pripadajuću klasu u koju se klasificira.

1.2 PODRUČJA PRIMJENE

Postoje tri glavna područja primjene u koje se može svrstati većinu HAR aplikacija prema [3], a to su:

i. Aplikacije nadzora (eng. ***surveillance***) – videom se prati jedna ili više osoba tijekom određenog perioda kako se bi ustanovilo da li promatrana osoba ili više njih izvode sumnjive aktivnosti. Primjeri ovakvih aplikacija su: nadzor aerodroma, nadgledanje javnih površina (trgova, parkirališta, plaža itd.), nadzor starih i nemoćnih ljudi itd.

ii. Aplikacije kontroliranja (eng. ***control***) – videom se snimaju pokreti jedne osobe koji se interpretiraju kao naredbe ili komande odvojenom, vanjskom sustavu. Primjer ovakvih aplikacija su: eng. *avatar-i* za video igre, sustavi virtualne stvarnosti itd.

iii. Aplikacije analize (eng. ***analysis***) – jedna ili više osoba se prate kroz duži vremenski period te se prepoznavanje pokreta vrši ili u realnom vremenu ili naknadno, nakon pohranjivanja snimljenog materijala. Detaljno se analiziraju snimljeni pokreti te je najznačajniji doprinos ovakvih aplikacija rezultat analize pokreta. Primjeri aplikacija su: aplikacije koje analiziraju sport (npr. nogomet ili tenis) te daju statistiku npr. tehnički dobro i loše izvedenih udaraca, ekspertne aplikacije u zdravstvu (nadgledanje oporavka pacijenta ili sprečavanja nastupanja bolesti kroz preventivni nadzor i analizu).

1.3 OSNOVNE VRSTE AKTIVNOSTI

U ovom poglavlju dana je definicija osnovnih pojmova i sintagmi na polju prepoznavanja aktivnosti korištenih u ovom radu. Definicije su preuzete iz radova [1], [2] i [3].

1.3.1 Gesta ili osnovna akcija

Gesta, naziva još i osnovna ili nedjeljiva akcija, definira se kao pokret koji se ne da dalje raslojavati te opisuje pokret na razini udova (ekstremiteta)[1]. Primjer geste je 'podignuti nogu' ili 'ispružiti ruku'.

1.3.2 Akcija

Akcija je sekvenca ili vremenski niz gesta jedne osobe koje zajedno (gledane kao cjelina) definiraju pokret cijelog tijela ili samo jednog njegovog djela[2]. Primjer akcije je 'mahanje rukom', 'šetnja' ili 'udarac šakom'.

1.3.3 Aktivnost

Aktivnost se definira kao niza ili skup akcija koje se vremenski mogu odvijati jedna iza druge ili paralelno - istovremeno, a koje analizirane zajedno (kao skup) otkrivaju o kojoj vrsti aktivnosti se radi na videu[3]. Npr. aktivnost 'fizičkog napada' se sastoji od akcije 'hodanja' (do mjesta napada) plus akcije 'udarac šakom'.

1.3.4 Interakcija

Definicijama geste, akcije i aktivnosti definirani su pokreti na razini jedne osobe, od osnovnih do najkompleksnijih. Međutim, postoje i definiciji radnji ili pokreta koji definiraju radnje na nivou više osoba ili osobe i okoliša, a zovu se interakcije.

Dakle, interakcija je aktivnost koja podrazumijeva aktivnosti najmanje dvije osobe ili aktivnost jedne osobe i objekta iz okoliša[1].

Primjer interakcije dvije osobe je 'hrvanje', a primjer interakcije osobe i objekta je 'glačanje'.

1.3.5 Grupna aktivnost ili aktivnost grupe

Grupna aktivnost predstavlja skup interakcija, dakle više osoba ili osoba i predmeta koji istovremeno obavljaju interakciju.

Primjer grupne aktivnosti više osoba je 'poslovni sastanak', a osoba i predmeta 'utrka automobilima'.

2 PREGLED PODRUČJA

U ovom poglavlju biti će dan pregled trenutnih dostignuća kao i potencijalnih doprinosa polju kroz neriješene probleme iz pojedinog segmenta polja.

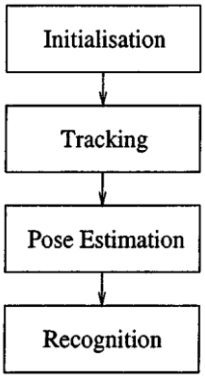
Pregled je uzeo u obzir sve radove objavljene na 6 trenutno vodećih svjetskih baza znanja do trenutnog datuma (Rujan, 2015). Baze znanja uzete u razmatranje su sljedeće: IEEE Xplore Digital Library (ieeexplore.ieee.org), Elsevier (www.elsevier.com), Springer Link (link.springer.com), ACM Digital Library (dl.acm.org), MIT Press (mitpress.mit.edu), Wiley Online Library (onlinelibrary.wiley.com).

Radovi su pronađeni korištenjem kombinacije ključnih riječi „*human activity recognition*“ te „*human behaviour detection*“ pri čemu su izuzeti svi radovi koji nisu bili temeljeni na monokularnom videu bez senzora kao dodatnih pomagala. Prednost je dana radovima čiji autori imaju veći h indeks te onima koji su objavljeni u časopisima ili zbornicima sa višim h indeksom. Također, u slučaju da je rad dao doprinos na više područja detaljno se opisuje samo na jednom području.

2.1 OPĆENITO O FUNKCIONALNOJ TAKSONOMIJI RAZVRSTAVANJA

Korištenjem funkcionalne taksonomije, prvi puta definirane u [4], biti će dan pregled najnovijih dostignuća na polju HAR-a. Međutim, prije toga potrebno je objasniti što je to funkcionalna taksonomija i od čega se sastoji.

Ideja sintagme funkcionalna taksonomija je svrstavanje svakog rada koji se bavi HAR tematikom u predefinirane skupine (ili funkcije sa slika 2.1) koje predstavljaju globalne module HAR sustava. Zajedničkim radom svih modula omogućen je konačni rezultat, a to je donošenje odluke o vrsti aktivnosti (a time i njenoj kategoriji).



Slika 1 [4] – globalni moduli HAR sustava

Kako bi sustav započeo sa procesiranjem videa potrebno je da bude **inicijaliziran**, tj. da se sagradi odgovarajući model sustava koji se stavlja u početnu točku.

Eng. Tracking ili **praćenje** je proces segmentiranja subjekta kojega se prati i pozadine te praćenja segmentiranih površina kroz video sekvencu.

Eng. Pose Estimation ili **procjena poze** je proces kojim se utvrđuje koju pozu je subjekt izveo unutar promatranog videa.

Eng. Recognition ili **prepoznavanje** je proces zaključivanje o vrsti akcije/ponašanja promatranog subjekta.

2.2 INICIJALIZACIJA – PRVI KORAK

Prvi korak u radu svakog HAR sustava je inicijalizacija sustava na početnu točku ili ishodište od koje kreće proces prepoznavanja. Prema [4] inicijalizacija se definira kao skup potrebnih radnji kako bi se se osiguralo da HAR sustav ispravno interpretira trenutnu scenu (u trenutku početka snimanja), a sve u svrhu ispravnog puštanja u rad HAR sustava. Nažalost, postoji dio posla kojega je potrebno obaviti prije koraka inicijalizacije, a kojega je nemoguće izbjeći (poput izabiranja, a potom izrađivanja kinematičkog modela koji će modelirati pokrete). Međutim, ta količina posla koju je potrebno obaviti prije sa sve većim napretkom polja teži smanjivanju, tj. „prebacivanju“ u samu fazu inicijalizacije čime dobivamo puno bolji sustav, od monolitnije funkcionalnosti do jednostavnijeg održavanja. Većina radnji inicijalizacije spada u skupine sljedećih radnji: kalibracija kamere, prilagodbu sustava na uvjeti na sceni, inicijalizaciju kinematičkog modela.

Neke radnje iz koraka inicijalizacije se mogu ponavljati u daljnjem radu sustava, dakle one nisu rezervirane isključivo za fazu inicijalizacije. U takve radnje spadaju postizanje oštrine slike, uklanjanje umjetnih artefakata itd.

2.2.1 Inicijalizacija kinematičkog modela kao glavna radnje ove faze obrade

Inicijalizacija kinematičkog modela kao najvažnija radnja ove faze prema [4] definira se kao radnja za čije izvođenje su potrebna dva preduvjeta: otkrivanje inicijalne poze subjekta na videu (čovjeka ili ljudi) te posjedovanje modela koji modelira kinematiku promatranog subjekta. Neki radovi ove dvije radnje obavljaju u istom koraku, kao dio inicijalizacije kinematičkog modela.

2.2.1.1 Početna poza

Samo neki sustavi inicijalnu ili početnu pozu prepoznaju automatski iz inicijalne scene. Svi ostali zahtjevu da se poza ili označi ručno od strane ljudskog agenta koji upravlja sustavom ili bude unaprijed navedena koordinata na kojoj će se poza nalaziti.

2.2.1.2 Kinematički modeli

Pod kinematičkim modelom, prema [6], se smatra humanoidni model koji aproksimira izgled, oblik i kinematiku čovjeka. S obzirom na način konstruiranja modela, tj. na podatke koje model koristi razlikuju se sljedeći tipovi modela:

- Modeli temeljeni na kinematičkim strukturama
- Modeli temeljeni na 3D oblicima
- Modeli temeljeni na boji
- Modeli temeljeni na lociranju djela tijela

Što se tiče modela koji su temeljeni na kinematskim strukturama, većina njih definira ograničenja uz koje se model gradi, a to su unaprijed definiran broj zglobova te za svaki zglob broj prostornih dimenzije unutar kojih se pokreti zgloba izvode. Time se postupak inicijalizacije kinetičkog modela temeljenog na strukturi svodi na računanje jedinog parametra kojeg nije moguće unaprijed definirati u kinematičkom modelu, a on je dužina udova od zgloba do vrh.

2.2.2 Trenutna dostignuća u inicijalizaciji

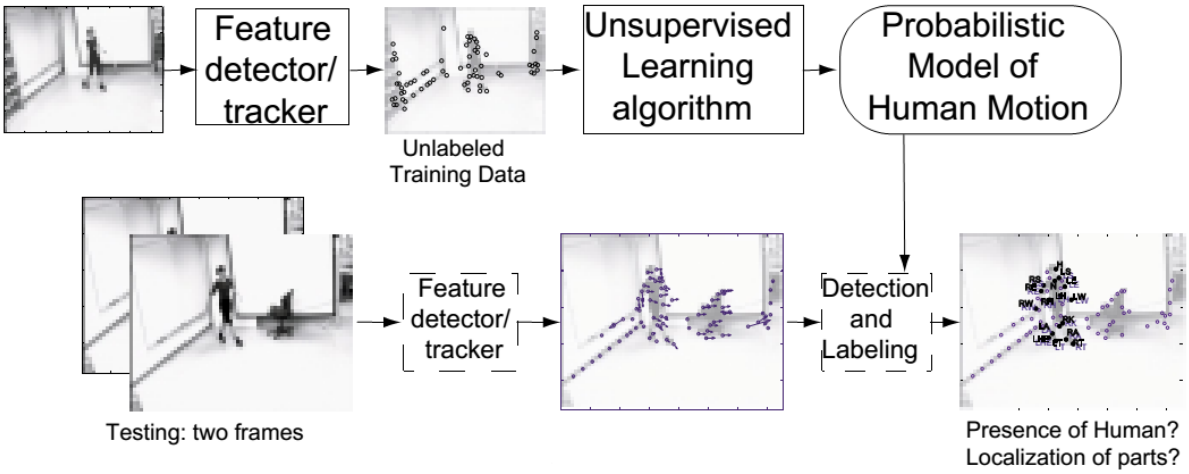
2.2.2.1 Kinematički modeli temeljeni na kinematičkoj strukturi

Inicijalizacija poze sa procjenom duljine udova korištenjem ručno identificiranih zglobova iz videa snimljenog jednom kamerom obrađena je radovima [7], [8], [9] i [10].

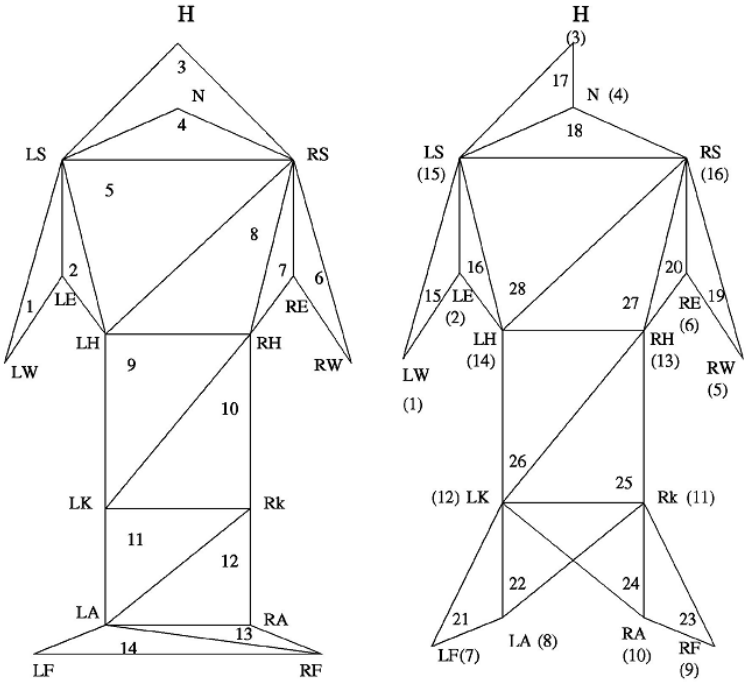
Metoda za automatsku inicijalizaciju kinematskog modela gornjih dijelova tijela obrađena je radovima [11] i [12] tako što se koristila segmentacija pokreta (odvajanje pokretnih segmenta slike od statičkog zaleđa) snimljenih jednom kamerom.

Kao jedan od najznačajnijih radova s obzirom na trenutni broj citata u području automatske konstrukcije modela na osnovu strukture izdvaja se [13]. Autori [13] su konstruirali nekoliko novih algoritama od kojih svaki rješava po jedan dio problema. Kao što je prikazano slikom 2 najprije se pusti video kroz Tomasi- Kanade algoritam kako bi se otkrile sve značajke na slici. Otkrivene značajke su prikazane kao zaokružene grupe eng. *piksel*-a na drugoj sličici slike 2. Potom se dalje radi učenje na osnovu značajki, ali nemaju sve značajke jednaku važnost. Zato se paralelno odvija proces sa 3. i 4. sličice koji na osnovu dvije slike snimljene jedna iza druge (sličica 3. i 4.) daje veću važnost značajkama sa većom

brzinom pomaka što bi mogla sugerirati da se radi o značajkama subjekta. Na osnovu ove informacije algoritam je u stanju napravit nešto što se zove eng. *Decomposable triangulated graph* prikazan na slici 2 sličica 5 i slika 3 (uvećano).



Slika 2 [13] – Kratak prikaz sustava iz [13]



Slika 3 [13] – primjer eng. *Decomposable triangulated graph*; (a) brojevi označavaju jedan od mogućih redoslijeda uklanjanja vrhova, slova su oznake dijelova tijela; (b) primjer drugačijeg označavanja vrhova za oduzimanje i općenito – oni u zagradama označavaju redoslijed eliminacije, a obični brojevi su naziv vrha

Bitno je napomenuti da se može raditi i sa drugim vrstama grafova od ovdje korištenog trokuta, međutim oni ne daju najbolje rezultate jer nisu otporni na preklapanje. Naime, dok je barem jedan čvor neprekloplјen prepoznavanje segmenta subjekta koji je djelomično preklopljen (ostala dva čvora) je i dalje uspješno. Nadalje, značajke koje pripadaju subjektu

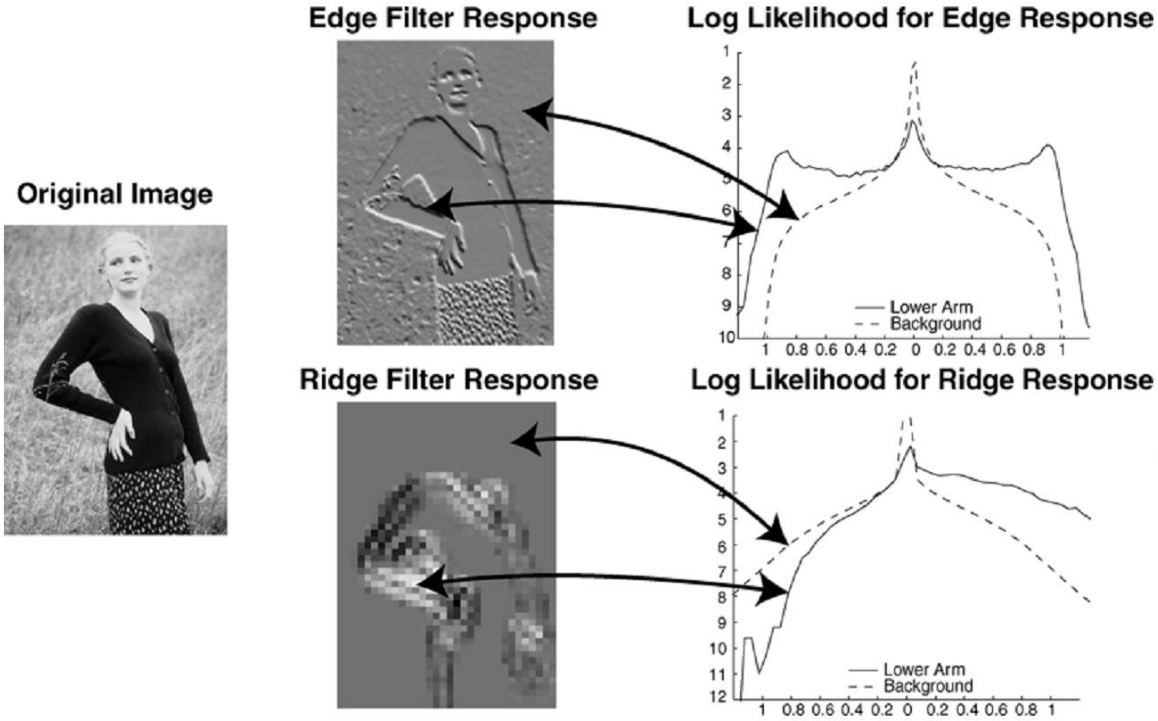
ne moraju biti vidljive u svakoj slici, algoritam uspješno funkcionira i bez podataka o značajkama iz svake slike.

2.2.2.2 Kinematički modeli temeljnih na 3D oblicima

Kako je uvjet za dobivanje ovakvih modela posjedovanje 3D podataka, a njih je nemoguće dobiti kroz monokularni video, onda ovo poglavlje preskačemo (vidi uvod u ovu cjelinu).

2.2.2.3 Kinematički modeli temeljeni na boji

Vrijedan rad temeljen na boji i monokularnom videu je [14] i [15]. Koristi činjenicu da je luminiscencija ekstremiteta na rubu i luminiscencija pozadine bitno drugačija, tj. među te dvije vrijednosti postoji razlika ili pomak. Zbog toga ima smisla provući rubne filtre preko slike jer se njima izdvoje jasne konture ekstremiteta – slika 4 (eng. *edge filter*). Radovi koriste prvu derivaciju normaliziranog Gaussovog filtra čime se dobije rub zvan eng. *steered*, a to znači da rub generira krivulju koja prati orijentaciju ekstremiteta. Ukoliko se takav rub skalira (pomoću neke od funkcija skaliranja) dobije se rasjed ili eng. *ridge* – slika 4 (eng. *ridge filter*). Karakteristika takvih rubova je ta što dio krivulje koja je okomita na os ekstremiteta ima visoko vrijednost osvijetljeno, a dio krivulje paralelan sa osi ekstremiteta ima nisku vrijednost svjetline.



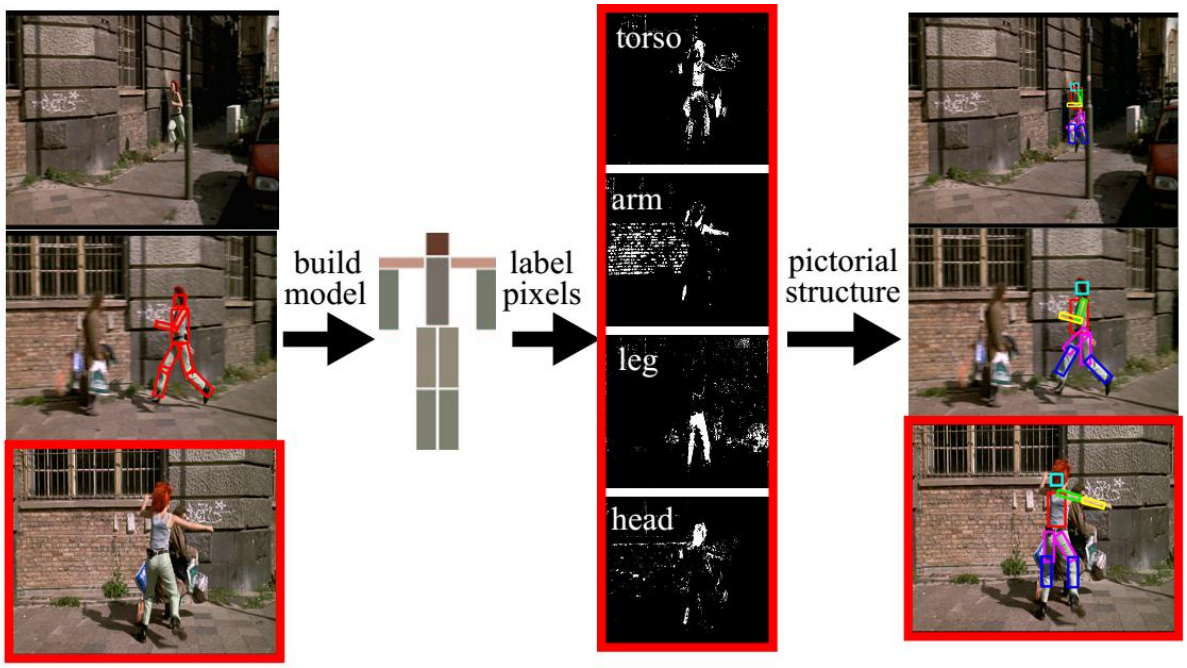
Slika 4 [15] – Računanje funkcije vjerovatnosti razdiobe za ektremitete na osnovu filtra ruba i filtra rasjeda (eng. *ridge*)

Posljednja značajka koju autori [14] i [15] koriste je informacija o pokretu ili eng. *motion* cue. Ta informacija je moguća zbog pretpostavke da se intezitet površine koja predstavlja ekstremitete, dakle dio površine slike čija tekstura predstavlja subjekt, vrlo malo promjeni između dvije uzastopne slike jedne sekvence. Uz pomoć ove tri distribucije (rub, rascjep i informacija o pokretu) napravi se praćenje pokreta subjekta temeljeno na Bayesovoj statistici – algoritam kojega su autori predložili kao doprinos području. Treba još napomenuti da autori algoritam uče ručno, na označenim podacima. Također, algoritam se uči i na sekvencama gdje nema ljudi te time nauči kakva je raspodjela vjerojatnosti scena koje ne sadrže ekstremitete.

2.2.2.4 Kinematički modeli temeljeni na dijelovima tijela

Najnoviji način otkrivanja inicijalnog položaja, a time i praćenja subjekta je lociranje samo nekih dijelova tijela te konstruiranje prediktivnog modela koji uz pomoć takvih, parcijalnih podataka generira odluku o kakvoj se pozi radi sa određenom vjerojatnošću.

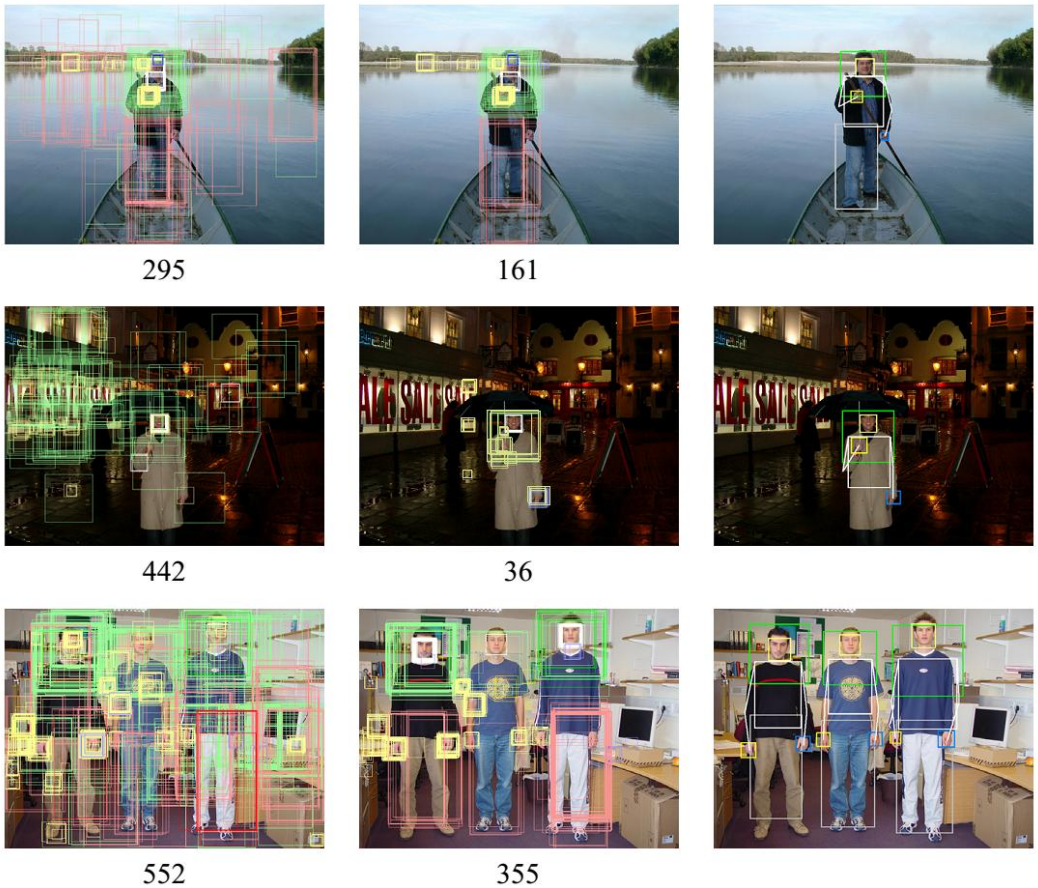
Radovi [16] i [18] (slika 5) najprije propuste filter koji prepoznaje hodanje na svaku sliku sekvence. Filter je skaliran na samo jednu vrijednost te nije u stanju prepoznati manje skalirane subjekte (slika 5 lijevo, prvi red). Također filter ne hvata niti atipične pokrete (slika 5 lijevo, posljednji red). Međutim autori potom rade sljedeće: zahvaljujući poznatoj dužini udova sa slike 5 lijevo - drugi red, uče nešto što zovu kvadratni, logički, regresijski klasifikator za svaki ud, a sve se to događa u RGB modelu boja korištenjem piksela koji prekrivaju površine udova kao pozitivnih vrijednosti te onih koji nisu na površini udova kao negativnih. Naučeni klasifikatori se onda ponovno puštaju na svaku sliku sekvence kako bi



Slika 5 [16] – način funkcioniranja sustava [16] i [18]

se prepoznalo i one pokrete koji inicijalno nisu bili uspješno prepoznati. Slika 5 – sredina prikazuje kako izgleda prepoznavanje dijelova tijela sa takvim klasifikatorom za atipičnu radnju sa sličice lijevo koja je u trećem redu. Posljednji korak je spajanje dijelova tijela u kompletan model koji predstavlja cijelo tijelo (slika 5 desno treći red). Ista logika vrijedi i za sliku 5 - desno - prvi red.

Rad [17] velikom točnošću određuje lokaciju i 2D pozu čovjeka u pozadini s visokom entropijom. Zbog različitih boja odjeće i kože prepoznavanje aktivnosti koristi Adaboost metodu detekcije subjekta te posebno konstruirane detektore lica, torza, nogu i ruku. Detektirani dijelovi spajaju se u 2D površinu koja označava cijelo tijelo korištenjem RANSAC metode u realnom vremenu.



Slika 6 [17] – prvi stupac su sve detekcije na slici, drugi stupac prikazuje reducirane detekcije i treći konačno pronađeni oblik. Brojevi ispod slika kazuju broj detekcija.

Rad [21] funkcionira u tri faze. Najprije kreira nešto što naziva „Integralna“ slika, sintagma označava drugačiji aranžman podataka koji predstavljaju ulaz u sustav obrade, a iz kojih se veoma brzo računaju posebno konstruirane značajke. Druga faza je algoritam temeljen na AdaBoost metodi koji izabire mali skup najznačajnijih značajki sa slike te generira iznimno brze klasifikatore. Treća i posljednja faza obrade je metoda koja kombinira kaskadno klasifikatore, temeljni doprinos joj je brzo i efikasno eliminiranje pozadine čime

ostavlja glavninu vremena obrade objektima slike koji bi mogli predstavljati subjekte. [21] se pokazao kao jedan od trenutno boljih radova na polju automatske detekcije lica ljudi.

Ostali radovi vrijedni spomena su [19] i [20].

2.2.3 Trenutno otvoreni problemi u inicijalizaciji

Inicijalizacija modela monokularnog videa koji uz veliku točnost nastavljaju pratiti subjekt uslijed promjene na odjeći, kosi ili obliku tijela, a bez postavljenih hipoteza o okolišu ili subjektu ostaje neriješeno pitanje do danas.

2.3 PRAĆENJE – DRUGI KORAK

Praćenje ili eng. *tracking* se prema [2] i [4] definira kao proces koji se sastoji od dva koraka:

I. Eng. *figure-ground* segmentacija – proces izdvajanja subjekta kojeg pratimo (čovjeka) iz pozadine. Metode izdvajanja subjekta su većinom prve metode obrade videa te je njihovo pravilno funkcioniranje iznimno bitno. Ova grana dijeli se na: oduzimanje pozadine, segmentaciju korištenjem pokreta subjekta, segmentaciju korištenjem izgleda i oblika subjekta, segmentaciju korištenjem dubine slike.

II. Eng. *temporal correspondences* – proces povezivanja prepoznatog subjekta na trenutnoj slici sekvence sa prethodnom slikom sekvence uz korištenje nečega što se zove eng. *trajectories* ili trajektorije.

Figure-ground segmentacija temelji se na dvije vrste podataka ili informacija koje je moguće dobiti iz slike ili sekvence slika: eng. *temporal* (trenutni, privremeni) podacima i eng. *spatial* (prostorni) podacima.

Globalno gledano, svi algoritmi konstruirani za praćenje mogu se podijeliti u dva glavna skupa: deterministički algoritmi i stohastički algoritmi.

Jedna skupina determinističkih algoritama rade praćenje tako da iterativno traže lokalni maksimum neke funkcija sličnosti između mete praćenja (subjekt) i predloška o subjektu. Najčešće korištena funkcija sličnosti je SSD ili suma kvadrat udaljenosti. Poboljšanje u odnosu na ovaj pristup je napraviti najprije segmentaciju podataka (npr. *mean shift* algoritmom), a onda primijeniti SSD. Sljedeća skupina determinističkih algoritama temeljenih na modelu subjekta koriste unaprijed poznatu informaciju o subjektu kako bi proizveli prezentaciju subjekta. Neki od primjera prezentacije subjekta su: izgled kože,

blobovi, kinematički skeleton, siluete itd. Grupa determinističkih algoritama koji se temelje na izgledu subjekta primjenjuju metoda prepoznavanja subjekta u nekom prostoru poput prostora razapetog svojstvenim vektori ili u prostoru *kernel-a.*

Kod stohastičkih algoritama, praćenje se modelira diskretnim sustavima stanja. Informacija koja opisuje metu ili subjekt praćenja može se predstaviti sekvencom stanja oblika $\{x_k\}_{k=0,1,2,\dots}$. Ukoliko želimo znati točno stanje sustava u trenutku k ono je dana jednadžbom:

$$x_k = \boldsymbol{f_k}(x_{k-1}, v_k)$$

gdje je v_k šum u stanju k, a $\boldsymbol{f_k}$ je vektorska funkcija koja je nelinearna i ovisi o vremenu k. U sustavima koji se mogu opisati linearnim Gaussovim modelima postoji uvijek samo jedan mode razdiobe funkcija raspodjele vjerojatnosti (eng. *PDF* ili Probability Density Function) te Kalmanov filter propagira i ažurira *mean* i kovarinacu takve funkcije. Međutim, kod ne linearnih sustava nemoguće je analitički raščlaniti distribuciju te se onda koriste razne metode aproksimacije distribucije. Najčešće korištena i najpoznatija metode aproksimacije je filter čestica ili sekvencijalni *Monte Carlo* koji rekurzivno konstruira PDF prostora stanja korištenjem Monte Carlo integracije. U računalnom vidu ovaj pristup je poznat kao Kondenzacijski algoritam ili eng. *Condensation algorithm.* Ostali algoritmi temeljeni na filterima čestica koriste konture, neke značajke boje ili modele izgleda. Histogram boja je otporan na šum i djelomično preklapanje, ali ne podnosi promjenu osvjetljenja ili prisustvo boja u pozadini sličnih subjektu. Također, računanje dugo traje ako je regija koja se prati velika površinom te ako je općenito broj regija velik. Metode koje koriste konturu umjesto histograma boja rješavaju problem promjene iluminaciju, ali su iznimno složene za računanje. Filter čestica generalno predstavlja tehniku uzorkovanja Bayesovih sekvenci koja rekurzivno aproksimira *posterior* distribuciju korištenjem konačnog skupa otežanih uzoraka. U praksi, takva tehnika se sastoji od dva koraka: predviđanja i ažuriranja. Neka su do trenutka $t-1$ dana sva prošla zapažanja izrazom:

$$z_{1:t-1} = \{z_1, \dots, z_{t-1}\}$$

Korak predviđanja uz pomoć tranzicije vjerojatnosti, oblika $p(x_t|x_{t-1})$ računa vjerojatnost sadašnjeg trenutka koji ima oblik izražen s

$$p(x_t|z_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})\, dx_{t-1}$$

Kako je završetkom trenutka t poznat događaj z_t onda se prethodan jednadžba može zapisati korištenjem Bayesovog pravila:

$$p(x_t|z_{1:t-1}) = \frac{p(z_t|x_t)p(x_t|z_{1:t-1})}{p(z_t|z_{1:t-1})}$$

Posterior vjerojatnost $p(x_t|z_{1:t-1})$ je aproksimirana konačnim skupom od N uzoraka $\{x_t^i\}_{i=1,\dots,N}$ sa težinskim faktorima w_t^i. Uzorci kandidati označavaju se s $\tilde{x}_t^i$, a uzimaju se iz distribucije oznake $q(x_t|x_{1:t-1}, z_{1:t})$. Slijedi definicija težinskih faktora s obzirom na iznesene oznake:

$$w_t^i = w_{t-1}^i \frac{p(z_t|\tilde{x}_t^i)p(\tilde{x}_t^i|x_{t-1}^i)}{q(\tilde{x}_t|x_{1:t-1}, z_{1:t})}$$

U slučaju poznatog filtera čestica eng. *Bootstrap* izraz $q(x_t|x_{1:t-1}, z_{1:t})$ jednak je $p(x_t|x_{t-1})$, a težinski faktori postaju $p(z_t|x_t)$.

Gledano iz perspektivne implementacije, praćenje se radi većinom korištenjem nečega što se zove trenutni podaci. Kod praćenja objekta sa statičkom pozadinom jedina promjena vrijednosti piksela događa se usred pomicanja subjekta. Trenutni podaci generiraju se korištenjem sljedećih dviju vrsta operacija: oduzimanjem sadržaja slika međusobno ili kreiranjem toka podatak pomoću značajki slike, blobova itd. (slika 7). Trenutni podaci izvlače se brže od prostornih i fokusirani su na subjekt promatranja na videu.

Temporal data		Spatial data	
Subtraction	Flow	Threshold	Statistics
Two images	Points	Chroma-keying	Pixels
Three images	Features	Special clothes	Blobs
Background	Blobs	IR	Contours

Tablica 1 – preuzeto iz [4]

Ako se radi o operaciji oduzimanju vrijednosti piksela dvije ili tri uzastopne slike ona se događa tako da napravimo matematičko oduzimanja neke komponente piksela poput inteziteta ili krome. Razlika dobivena oduzimanjem će predstavljati sve piksela na kojima je bio pokret, osim u slučaju kada subjekt i okolina imaju isti intezitet, tj. kromu. Nadogradnja na ovu operaciju je ažuriranje pozadine novim sadržajem koji se dogodio između dva kadra, a sve to da bi sam koraka oduzimanja bio precizniji.

Pokret ili eng. *flow* označava koherentno gibanje značajki ili točaka od interesa za praćenje između dva susjedna okvira sekvence.

Prostorni podaci generiraju se dvjema osnovnim vrstama operacija: eng. *tresholding* i statističkim pristupima. Prvi način generiranja podataka oslanja se na nekakvu hipotezu uvjetovanu okolišem poput one da ljudska odjeća ima veću vrijednost zasićenosti boje od okoliša koji ju okružuje(rad [33]). Statistički pristupi su napredno iskorištavanje trenutnih podataka dobivenih oduzimanjem tako da se iterativno računaju karakteristike pojedinog piksela ili grupe piksela kako bi segmentirali subjekt od ostatka. Te karakteristike su npr. rub ili neka komponenta piksela poput inteziteta. Primjer statističkog pristupa je kada se na sekvenci slika koje predstavljaju statičku pozadinu za svaki piksel na svakoj slici uzme vrijednost inteziteta kako bi se dobila varijanca i srednja vrijednost inteziteta pozadine. Takva vrijednost se onda uspoređuje sa intezitetima piksela nepoznate sekvence te svi pikseli čija je vrijednost inteziteta bliska izračunatoj, po pretpostavci, čine pozadinu nepoznate sekvence.

Uspješnost korištenja metoda eng. *tresholding-a* bitno je određena kvalitetom postavljenih pretpostavki. Ovo je u suprotnosti za zahtjevom da se aktivnosti prepoznaju na bilo kojoj pozadini i u bilo kojim uvjetima. Međutim, postoji dio aplikacija gdje se i okoliš i subjekt mogu kontrolirati te time *tresholding* dobiva na značaju. Za aplikacije koje imaju neograničene uvjete snimanja boljim su se pokazale statističke metoda zbog svoje adaptivnosti.

2.3.1 Oduzimanje pozadine

Sve do 1998. god. prepoznavanje ljudskih pokreta je bilo moguće samo u zatvorenim prostorima jer nisu postojale metode koje bi taj proces uspješno radile na otvorenom prostoru upravo zbog ovog koraka oduzimanja pozadine koji je zadovoljavajuće radio samo na statičkim sekvencama. Međutim, radom [22] koji je predstavljao veliki doprinos računalnom vidu, riješen je i taj problem. Autori [22] su predložili da se svaki piksel slike predstavi kroz eng. *mixture of Gaussians (MoG)* te da se svaki takav pikse redovito ažurira sa novim Gaussom. Ažuriranje se uglavnom radi rekurzivno čime gubimo mogućnost oduzimanja pozadine kod iznimno brzih promjena sadržaja scene.

Oduzimanje pozadine je složen postupak koji se dijeli na kraće procese:

a) Prezentacija pozadine
b) Klasifikacija pozadine
c) Ažuriranje pozadine
d) Inicijalizaciju pozadine

2.3.1.1 Prezentacija pozadine

Segmentirani entiteti (subjekti i dijelovi pozadine) se trebaju opisati nekakvom prezentacijom. Generalno, sve različite prezentacije podatak slike spadaju u sljedeće dvije grupe: prezentacije subjekta eng. *object based representation* koje se pak temelje na *figure-ground* segmentaciji te eng. *image based representation* koje se konstruiraju direktno iz piksela slike. Tablica 2 pokazuje podskupine ovih dviju glavnih grupa.

Object-based	Image-based
Point	Spatial
Box	Spatio-temporal
Silhouette	Edge
Blob	Features

Tablica 2 – vrste prezentacija podataka slike i njihove podgrupe

Prezentacija eng. *point-om* ili točkom se koristi većinom kod sustava koji rabe pasivne ili aktivne markere uz samu kameru zato što markeri imaju visoki kontrast u odnosu na ostatak sadržaja slike. Ako se tako označeni subjekti snimaju istovremeno sa više kamera moguće je napraviti 3D prezentaciju podatak.

Prezentacija eng. *box-om* ili pravokutnicima je najučestalija metoda prezentacija segmentiranih podataka. Ideja je uhvatiti pravokutnicima regije slike u kojima se nalazi subjekt prepoznat *figure-ground* segmentacijom. Neki sustavi potom rade praćenje subjekta korištenjem ovako konstruiranih pravokutnika dok drugima oni služe samo kao međukorak za ulazak u fazu procjene poze (korak 3 u globanoj obradi podatak).

Prezentacija eng. *silhouette* ili siluetom je najjednostavnija prezentacija. Postiže se tako da se od *figure-ground* segmentacije operacijama oduzimanja sadržaja ili operacijama eng. *thresholding* dobije ostatak koji predstavlja nekakvu siluetu subjekta. Koristi se i kod 2D i kod 3D prostorne prezentacije podatka. 2D prezentacija je obično lakša za implementaciju, ali neki autori koriste kompleksne 2D prezentacije poput zatvorenih, uniformnih B-splinova sa fiksnim brojem kontrolnih točaka međusobno jednako udaljenih. 3D prezentacija siluetom se postiže ili kombiniranjem 2D silueta i nekih dodatnih podataka ili pak stero tehnikama (više kamera odjednom itd.). Isto kao i za prezentacije kutijom ili pravokutnikom (eng. *box*) vrijedi da neki sustavi nastavljaju pratiti siluetu dok je drugima ona ulazak u sljedeću fazu (procjena poze).

Prezentacija eng. *blob* ili blob-om koristi slične tehnike kao *figure-ground* segmentacija tokom ili statistikom podataka.

Kod eng. *image based* prezentacije podaci se predstavljaju neovisno o postojanju subjekta u slici ili se ograničava prezentacija samo na piksele koji čine subjekt, a koji su predstavljeni jednom od metoda eng. *object-based* segmentacije.

Prezentacija eng. *spatial* ili prostorna se radi tako da se slike sekvence ili samo dijelovi slika koji segmentiraju objekta transformiraju u drugačiji prostor, npr. onaj razapet koordinatnim sustavom drugačijim od Kartezijevog, poput sfernog. Ovime se postiže kompaktniji zapis podataka. Transformacije koje se koriste kod prebacivanja podataka u drugi koordinatni sustav su Fourierova analiza, DCT (eng. *Discrete Cosinus Transform),* eng *wavelets* ili „valići“ te PCA (eng. *Principle Component Analysis*).

Ako se u prostornu prezentaciju doda vremenska dimenzija dobije se podskupina sa tablice 2 naziva eng. *spatio-temporal*. Ona je unaprjeđenje prostorne prezentacije zato što vremenska dimenzija otkrije sve značajke na slici vezane za pokret te se one onda mogu uključiti kao dio prezentacije.

Prezentacija korištenjem rubova ili eng. *edges* podrazumijeva liniju sačinjenu od točaka na rubu subjekta koji se prati ili niz kraćih linija koje prate rub subjekta od kojih su susjedne spojene zajedničkom točkom. Ova prezentacija se koristi kada se želi postići veća otpornost na šum.

Posljednja podgrupa je prezentacija značajkama ili eng. *feature based representation.* Značajke se računaju na osnovu jedne od prethodno opisanih prezentacija uz korištenje nekakvih pomoćnih podataka.

Izbor modela prezentacije bitno je određen brzinom kojom se procesiraju podaci da bi se postigla ciljana prezentacija. Uz ovo bitna je i preciznost prezentacije.

Kod prezentacije pozadine najučestaliji je MoG u RGB modelu boja. Međutim neki autori kombiniraju MoG sa normaliziranim RGB modelom kao u radu [23] te YUV modelom boja kao u [24]. Korištenjem MoG-a sa 3D modelima boja dobiju se elipsoidi i sfere kao u [25]. Autori [26] koriste krnje valjke, a [27] krnje stošce. Autori [28] koriste posebno konstruirani eng. *kernel* koji je u stanju vrlo dobro nositi se sa „pretrpanim“ pozadinama i brzim scenama zbog toga što se iznimno brzo prilagodi na promjenu sadržaja. Rad [28] predstavlja svaki piksel pozadine sa sekvencom bita te se pokazao vrlo uspješnim u usporedbi sa ostalim modelima predstavljanja.

Rad [29] najprije prezentira piksel tako da napravi jednu vrstu aritmetičke sredine piksela i njegovih prvih susjeda (slika 7 lijevo). Potom istu sliku predstavi u svojstvenom

prostoru (prostor razapet svojstvenim vektorima matrice) – slika 7 sredina. Konačno, oduzimanjem dvije slike dobije informaciju potrebnu za praćenje koju prosljeđuje HMM i Bayesovim mrežama (slika 7 desno, zaokružena područja). Iznimno je uspješan u prepoznavanju grupnih aktivnosti ili aktivnosti grupe.

Slika 7 [29] – način prepoznavanja i praćenja subjekata radom [29]

Rad [30] koristi auto regresivni model u realnom vremenu koji hvata i predviđa lokaciju subjekta na slici uz pretrpanu pozadinu koja je vrlo dinamična (padanje kiša ili snijega, subjekti koji se kreću iza njišućeg stabla, kretanje subjekta po valovitoj vodenoj površini). Rad je konstruirao novu metriku temeljenu na usporedbi stanja između predviđenog i stvarnog položaja. Također, rad je imao vrlo loše rezultate detekcija i praćenja na vodi (puno lažnih detekcija položaja subjekta).

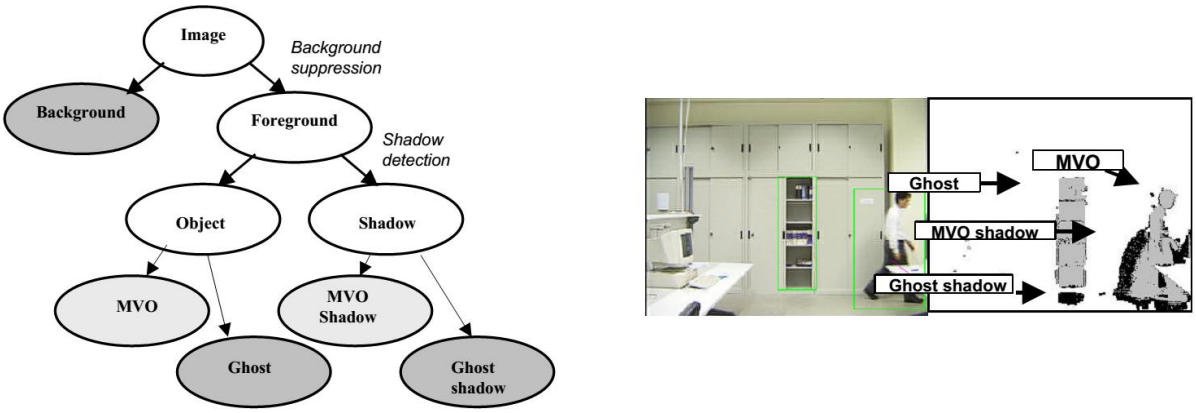
2.3.1.2 Klasifikacija pozadine

Oduzimanjem pozadine nastat će određeni broj segmenata slike, tj. skupina međusobno bliskih piksla koje mogu biti stvarna lokacija subjekta na slici, ali mogu biti i nekakav proizvoljni artefakt zaostao nakon oduzimanja. Zadatak ovog koraka druge faze obrade je klasifikacija takvih segmenata kao pozitivnih ili negativnih ovisna o tome da li predstavljaju lokaciju subjekta na slici ili ne. Jedan od najčešćih uzoraka krive klasifikacije su sjene.

Rad [31] predstavlja recenziju radova koji se bave problematikom sjena vezanom za pravilnu klasifikaciju subjekta.

Rad [32] je konstruirao sustav (prikazan slikom 8) koji ima dvije svrhe. Prva svrha je postići precizno praćenje subjekta, a to znači odjeliti subjekt od njegove sjene. Pri tome se ne koristi nikakav kinematski model nego samo razlika luminiscencije te boje između sjene i subjekta koji ju stvara. Drugi cilj koji autori sustava [32] žele postići je pravilna klasifikacija segmenata subjekta od segmenata koji predstavljaju lažni pomak uzrokovan promjenom u okolišu poput pomicanja komada pokućstva , pomaka uslijed šuma na slici ili nastajanja lažnog pokreta uslijed akcije subjekta (npr. bacanje predmeta). Autori zato definiraju sljedeće pojmove sa slike 8, lijevo: MOV, eng. *moving visual object,* je subjekt koji se dobije idealnom segmentacijom, eng. *background* je skup točaka koje nisu dio pokreta, eng. *shadow* tj. sijena

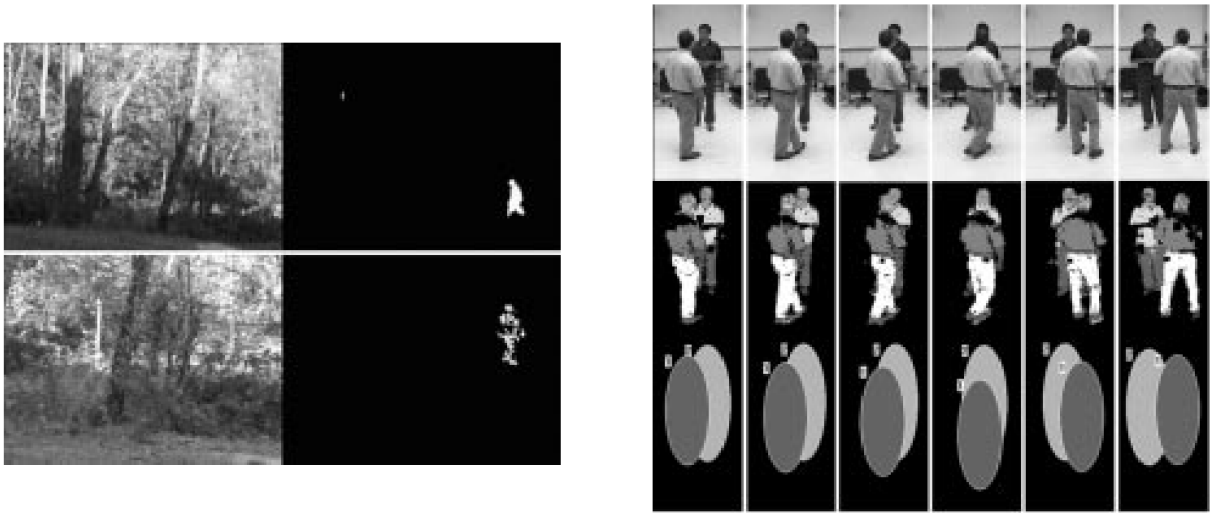
koja se definira kao skup susjednih točaka čije vrijednosti su promijenjene sjenom stvorenom od subjekta u pokretu, eng. *ghost* skup susjednih točaka detektirane kao pokret ali koje ne pripadaju niti jednom subjektu u pokretu. Sustav radi tako da najprije odjeli pozadinu od segmenata u pokretu. Potom među segmentima u pokretu traži one koji opisuju sjenu nasuprot onima koji opisuju subjekt. Među segmentima detektiranim kao subjekt u pokretu dio segmenta će pripadati sjeni. Nadalje, među segmentima detektiranim inicijalno kao sjena dio segmenta će biti stvarna sjena subjekta dok će jedan dio takvih segmenata biti sjena koja ne pripada niti jednom MOV-u. Primjer rada sustava dan je slikom 8 desno. Subjekt prođe ispred zatvorenih vrata ormara, otvori ih i krene dalje. Sustav u prvi mah detektira i subjekt i ormar kao objekta u pokretu, međutim samo je jedan segment stvarni, segment subjekta. Sustav uspješno eliminira lažne sjene i segmente pokreta lažnih objekata.



Slika 8 – preuzeto iz [32] – lijevo model sustava, desno primjer obrade sekvence

Rad [34] i [35] konstruira eng. *probability density function* (PDF) koja je neparametarska zato što se sadržaj videa bitno mijenja kroz vrijeme te se ne može uklopiti u nekakav generalni, parametarski opis. Sustav funkcionira na temelju dvije osnovne konstrukcije: statističkoj prezentaciji pozadine koja je veoma precizna u detekciji i najmanjeg pokreta, ali je robusna na pokrete koje nije načinio subjekt te statističkoj prezentaciji subjekta koja uspješno rješava problem preklapanja subjekta sa okolinom ili drugim subjektom. Slika 9 (lijevi dio slike) prikazuje primjer segmentacije subjekta i okoline koja ga djelom preklapa. Čak i u uvjetima gdje ljudsko oko ne vidi subjekt, kao što je slučaj na slici 9, sustav uspješno odrađuje prepoznavanje i praćenje. Također, desni dio slike 9 prikazuje primjer kako izgleda segmentacija i na koji način se prati preklapanja.

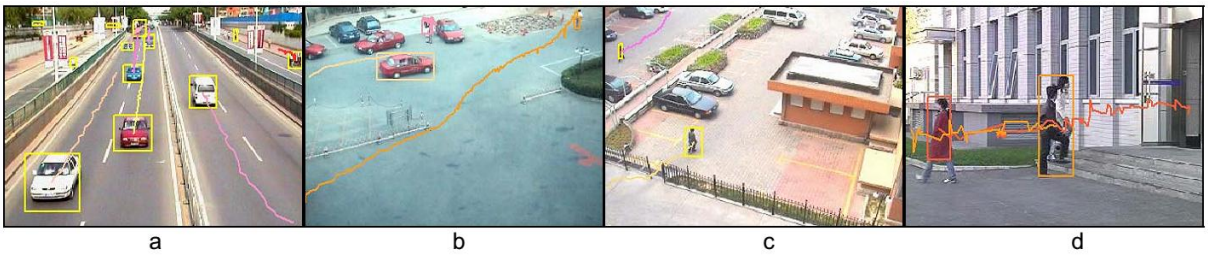
Rad [36] rješava problem preklapanja subjekta sa drugim subjektom ili sa okolišem tako da najprije konstruira značajku (eng. *feature*) koje se sastoje od otežanog histograma boja. Potom radi praćenje korištenjem tako konstruirane značajke i dobivene podatke pušta na algoritam za predviđanje pokreta koji se koristi eng. *mean shift* algoritmom.



Slika 9- preuzeto iz [34] – lijevo je primjer detekcije subjekta u izrazito pretrpanoj pozadini, desno je primjer bloba (2. sekvenca slika odozgo) i načina praćenja preklapanja (3. sekvenca odozgo)

Rad [37] je predložio adaptivnu metodu oduzimanja pozadine koja kombinira informaciju o boji i gradijentu kako bi eliminirala sjene i nepouzdane značajke boje. Boja se koristi kako bi se razriješilo preklapanje tako da se preko nje računa dubina subjekta na slici i pozicija u odnosu na drugi subjekt koji sudjeluje u preklapanju.

Rad [38] uspješno rješava potpuno, dugotrajno preklapanja subjekta bez prethodnog poznavanja oblika subjekta koji je preklopljen. Primjer praćena sustavom [38] dan je na slici 10 (ispod).



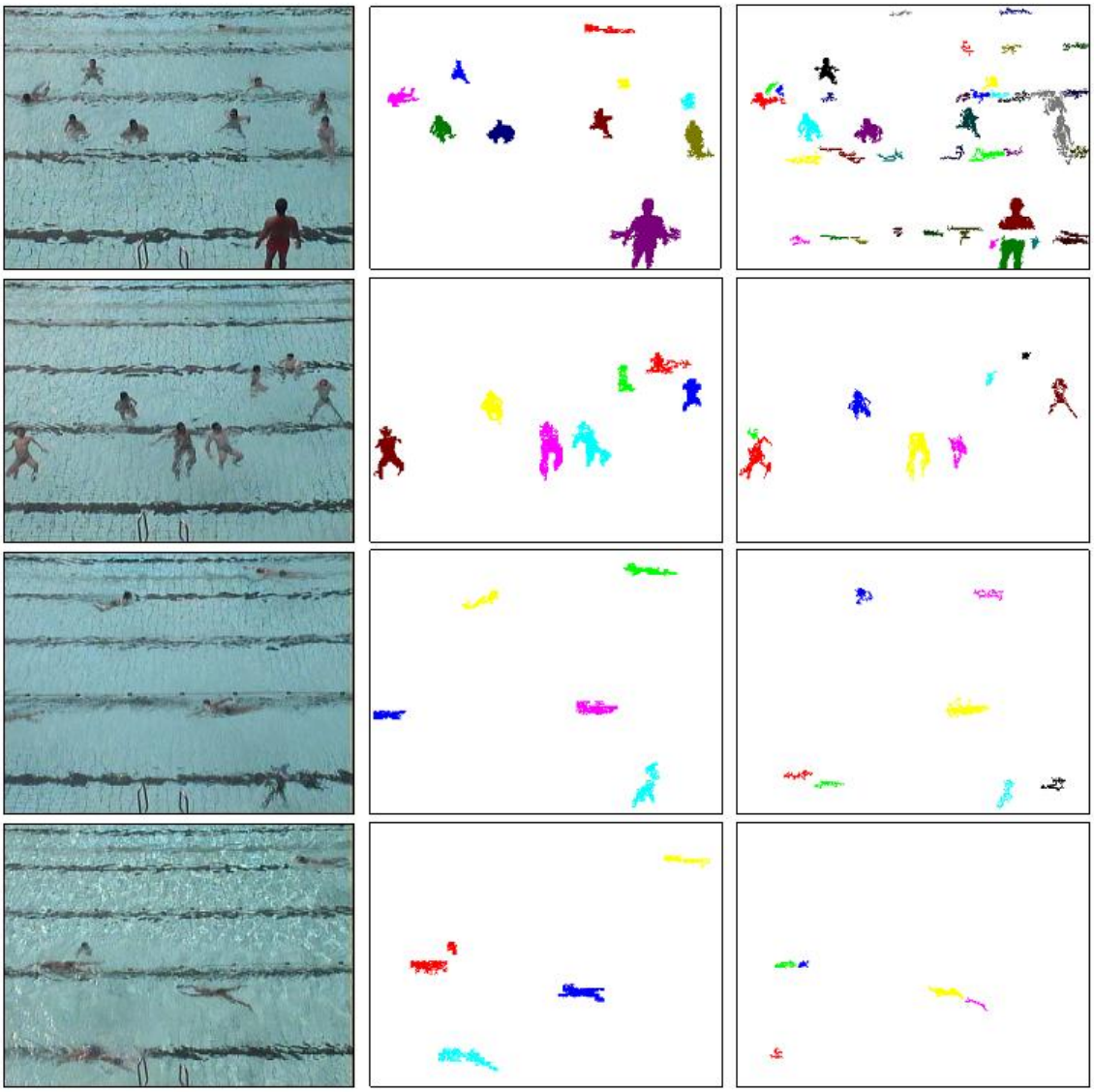
Slika 10 – preuzeto iz [38] – sustav trajektorijama ispisuju poziciju subjekta kroz sekvencu

Rad [39] koristi kinematski model ljudi, model kamere i pretpostavku da subjekti hodaju po ravnini.

Radovi [40] i [41] koriste proizvoljne Markovljeve lance u segmentaciji subjekta ili eng. *foreground-a*.

Rad [42] je predložio metodu za grupiranje piksela slike koja procesira podatke korištenjem dva kaskadna klasifikatora na način da se najprije klasificiraju pikseli u one koji čine pozadini i one koji to nisu. Potom se pikseli eng. *foreground-a* razvrstavaju u 4 grupe ili klastera: klaster pokreta, klaster sjene, klaster šuma i klaster piksela čija vrijednost je promijenjena uslijed automatskog namještanje ekspozicije kamere. Rad je pokazao dobre

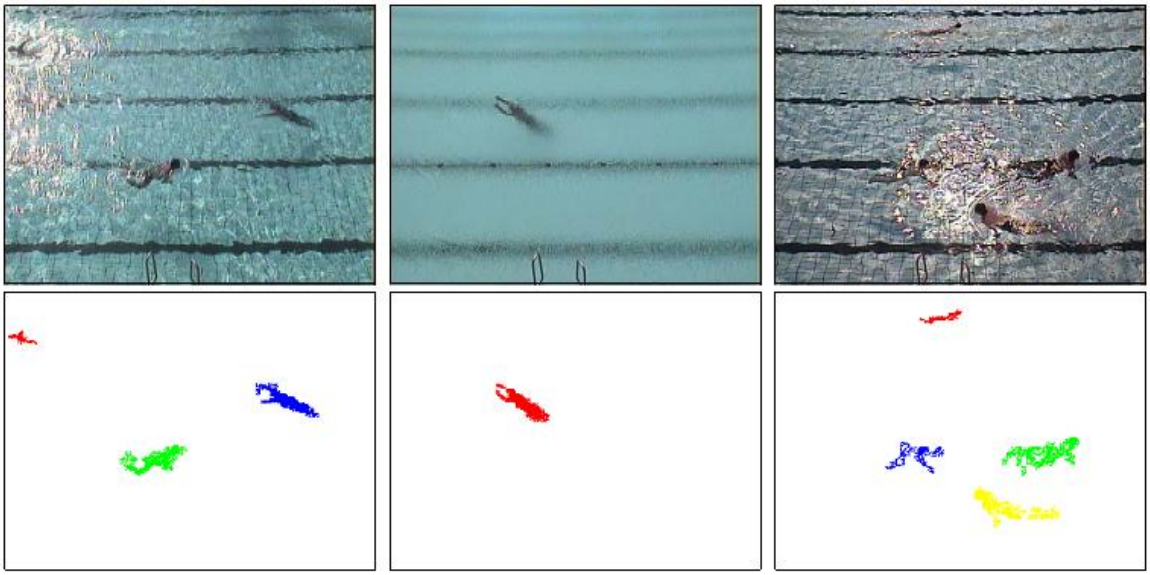
rezultate kod praćenje subjekata dugoročno uslijed kompleksnih scena (preklapanje, nepoznat okoliš itd.).



Slika 11 [43]– prvi stupac je snimljeni video sadržaj, drugi stupac je segmentacija korištenjem sustava [43], a treći stupac je segmentiranje korištenjem W4 sustava (pregled W4 kako ga definira rad [44] dan u poglavlju 2.3.1.4).

Rad [43] ima za cilj prepoznavanje kriznih situacija u vodi. Točnije, segmentira se i prati subjekt u otvorenom olimpijskom bazenu u uvjetima prskanja vode, refleksije sa površine vode te nagle promjene osvjetljenja (različito doba dana) s ciljem prepoznavanja opasnih situacija, poput utapanja. To se postiže korištenjem posebno konstruiranog algoritma segmentacije pozadine koji radi na blokovima piksela te korištenjem oduzimanja sadržaja sa dva praga (eng. *hysteresis*) tako da svi pikseli ispod donjeg praga ili iznad gornjeg bivaju obojani istom bojom (crnom obično). Parcijalna preklapanja se rješavaju prilagođenim Markovljevim lancima koji poboljšavaju praćenje sustava. Sustav koristi znanje eksperata kako bi definirao kriznu situaciju. Na osnovu tog znanja definira se set deskriptora plivača

koji se šalju posebnoj vrsti neuralnih mreža naziva eng. *functional link network* na procesiranje. Sustav sa velikom preciznošću i velikom otpornošću na promjene uvjeta scena detektira i prati subjekte. Slika 11 (gore) prikazuje usporedbu rezultata sustava i W4 algoritma, a slika 12 (dolje) funkcioniranje sustava uslijed velikih promjena uvjeta scene. Sa slike11 se uočava da sustav [43] radi bolju segmentaciju od W4 za vodene površine.



Slika 12 preuzeto iz [43]– primjeri segmentacija uslijed različitih uvjeta scene: rano jutro, kišni dan i večer

2.3.1.3 Ažuriranje pozadine

U scenama snimljenim na otvorenom prostoru vrijednosti piksela pozadine će se promijeniti nakon određenog vremena uslijed niza faktora te ja zato potreban mehanizam ažuriranje njihove vrijednosti na trenutnu, stvarnu vrijednost. Općenito vrijedi da se treba ažurirati samo one piksela pozadine čija se vrijednost nije promijenila kroz određeni period.

U modernim sustavima ovaj korak je spojen sa inicijalizacijom pozadine (poglavlje 2.3.1.4. – sljedeće poglavlje) tako da postoji mali broj trenutno aktualnih radova koji su vrijedni isticanja.

Rad [47] koristi adaptivnu metodu učenja na MoG-u kako bi ažurirala trenutnu vrijednost piksela. To postiže modifikacijom EM algoritma (eng. *Expectation Maximization*) na način da se smanji kompleksnost računanja EM-a.

2.3.1.4 Inicijalizacija pozadine

Inicijalizacija pozadine ovaj korak treba razlikovat od generalne inicijalizacije sustava prepoznavanja pokreta po tome što se u ovom koraku nastoji samo eliminirati subjekt kako bi se dobila osnovna pozadina, bez kretanja. Kod generalnog koraka inicijalizacije cilj je staviti sustav u početnu točku.

Stariji sustavi su koristili pretpostavku da se u inicijalnim slikama sekvence neće kriti niti jedan subjekt pa su inicijalnu pozadinu računali na takvim slikama sekvence. Međutim, moderni sustavi su u stanju inicijalizirati sustav i uz postojanje subjekta u početnim scenama čak i onda kada se on kreće te konstruirati model kojemu to neće bitno poremetiti stvarnu vrijednost parametara obrade zato što nekada učenje i treniranje na inicijalnim scenama bez subjekta neće biti moguće.

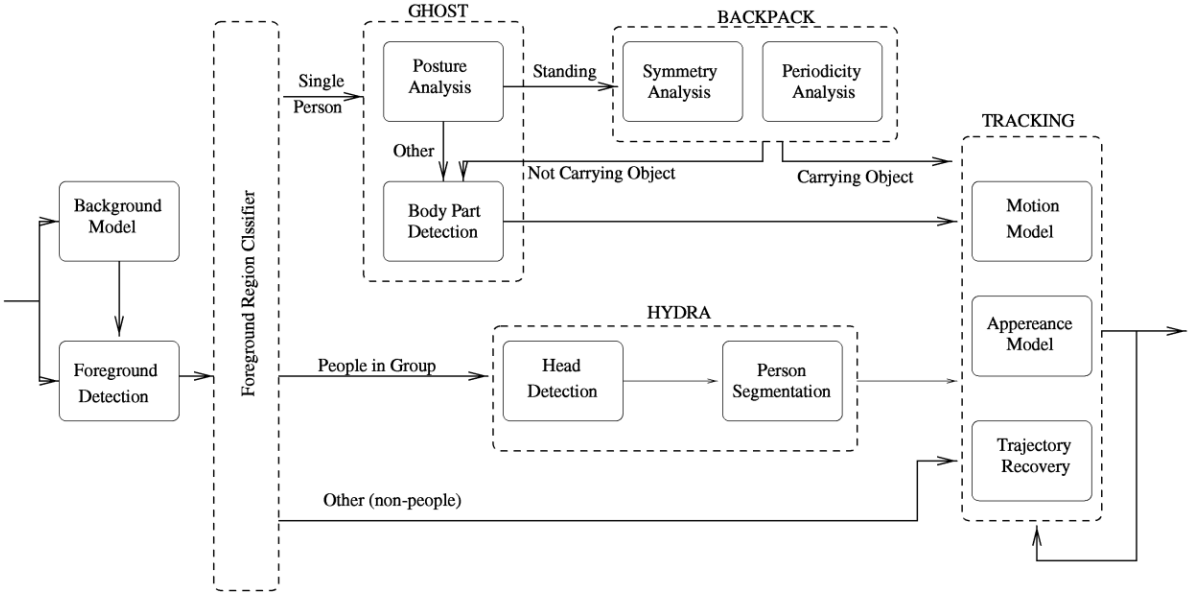
Kod prezentacija korištenjem MoG-a kretanje subjekta u fazi inicijalizacije izazvat će puno krivih klasifikacija u fazi klasifikacije pozadine.

Autori [43] rješavaju problem kretanja u inicijalnim scenama tako da koriste posebno konstruiran detektor ljudske kože te izrežu dio slike koji predstavlja čovjeka.

Autori [45] podjele piksele iz faze inicijalizacije u privremene, kratke intervale sa međusobno najsličnijim vrijednostima piksela. Potom na osnovu optičkog toka računaju koji interval koji pripada pozadini ima najmanju razliku međusobnih vrijednosti te ga uzmu kao pozadinu jer su u njemu našli minimalnu količinu promjene uslijed gibanja.

Autori [44] koriste model varijacije pozadine dobiven bimodalnom distribucijom nad uređenom (eng. *order*) statistikom pozadine kako bi poništili efekt kretanja u inicijalnim scenama. Na slici je prikazana arhitektura W4 sustava. U prvoj fazi W4 detektira piksele eng. *foreground-a* korištenjem statističkog modela pozadine te ih stavlja u nerazvrstane blobove, dakle ne zna se da li blob predstavlja čovjeka, grupu ili šum, samo se zaokruže pikseli koji nisu pozadina. Na tako definiranim blobovima računaju se potom globalne i lokalne značajke svakog od njih. Ovaj dio sustava prikazan je isprekidanim oblim pravokutnikom naziva eng. *Foreground Region Classifier.* Druga faza sustava svaki blob klasificira u jedan od tri zadane klase na osnovu analize statičke siluete i detektora harmonijskih pokreta. Klase su vidljive na slici i zovu se: osoba, grupa ljudi te ostali objekti. Ako je blob klasificiran u klasu osobe onda se računa poza osobe opet preko siluete. Ukoliko je osoba uspravna primjenjuje se detektor harmonijskih pokreta i analiza simetričnosti kako bi se utvrdilo da li osoba nosi neki predmet. Ako pak osoba ne nosi objekt ili nije u uspravnoj pozi koristi se rub siluete kako bi se identificirali svi dijelovi tijela. Ako blob iz prve faze obrade predstavlja grupu osoba onda se ne može detektirati dijelove tijela, poza i da li osoba nosi objekt ili ne. W4 u tim slučajevima pretpostavlja da su sve osobe u grupi uspravne i segmentira svaku jedinko za sebe. Ako blob pripada trećoj klasi (ostali objekti) W4 ne radi nikakvu daljnju analizu osim što nastaviti pratiti blob kroz sekvence. Nakon što ova druga faza obrade obavi svoj dio posla kreće treća faza, prikazan na slici kao praćenje. Praćenje najprije računa u kolikoj su mjeri prethodno

praćeni blobovi i trenutno otkriveni blobovi slični. Pomoću tog podatka konstruira trajektorije blobova kroz sekvence. Posljednja faza konstruira još izgled i model pokreta bloba.



Slika 13 – preuzeto iz [44] – arhitektura W4 sustava

Autori [46] su smislili poboljšanje metode mediana koja se koristila za inicijalizaciju početne scene na temeljnu pozadinu i *foreground.* Kako bi medijan bio uspješan potrebno je da više od polovice piksela slike pripada pozadini što nije uvijek slučaj.

2.3.2 Segmentacija temeljena na pokretu subjekta

Segmentacija putem pokreta pretpostavlja da je jedini izvor promjene u slici u odnosu na prošlu pokret (točnije pokret jednog ili više ljudi) te otkrivanjem pokreta otkrivate poziciju čovjeka. Pokret se mjeri putem optičkog toka ili razlike nastale oduzimanjem slika.

Rad [48] segmentira ljude na osnovu hoda. To postiže tako da najprije formira robustan i gust optički tok definiran vektorom:

$$\boldsymbol{u} = [u, v]$$

gdje je komponenta vektora *u* jednaka horizontalnom toku, a *v* je vertikalni optički tok između para slika koje su izabrane tako da budu jedna iza druge u sekvenci. Autori ručno izabiru uzorke toka hoda ljudi i uzorke toka koji nisu hod te tako izabrane uzorke predaju sustavu kao materijal za učenje. Zbog višedimenzionalnog prostora mogućih stanja i malog broja uzoraka za učenje izabran je SVM kao algoritam koje će klasificirati tokove u klase onih koji jesu ljudski hod i onih koji to nisu. Sustav podržava skalirani ulaz kao i različite kutove kamere i subjekta.

Optički tok funkcionira uspješno i rezultat se garantira samo ako nema brzih promjena scene te ako je osvjetljenje konzistentno. U slučaju kada jedan ili oba navedena uvjeta nisu zadovoljeni rješenje je konstruirati kompleksnije entitet iz normalnih, ulaznih podataka koji nisu direktno podložni analizi.

Rad [49] povećava preciznost praćenja pokreta u otežanim uvjetima povezivanjem susjednih točaka praćenja. Klasični Shi–Tomashi–Kanade treker definira zapis inteziteta sa

$$I(\boldsymbol{x}, t)\ gdje\ je\ vektor\ \boldsymbol{x}\ jednak\ [x_1, x_2]^T, a\ x_1\ i\ x_2\ su\ koordinate\ piksela\ slike$$

Jednadžba koja definira intezitet kroz vrijeme glasi:

$$I(\boldsymbol{x}, t) = I(\delta(x), t + \tau)\ gdje\ je\ \delta(*)\ polje\ pokreta \qquad (1)$$

Vrijednost polja pokreta u koordinati $\boldsymbol{x}$ aproksimira se s:

$$\delta(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{d}\ gdje\ je\ \boldsymbol{d}\ vektor\ pokreta$$

Kako jednadžba (1) nije u potpunosti točna u uvjetima šuma na slici, koristi se metoda najmanjih kvadrata kako bi se pronašao vektor pokreta koji minimizira sumu kvadrata udaljenosti stare pozicije uzorka *W* i nove pozicije uzorka *W (W* je mali uzorak slike sa središtem u hvatištu vektora ***d) :***

$$\varepsilon = \sum_W [I(\boldsymbol{x} + \boldsymbol{d}, t + \tau) - I(\boldsymbol{x}, t)]^2 \qquad (2)$$

Ekspanzijom izraza

$$I(\boldsymbol{x} + \boldsymbol{d}, t + \tau) - I(\boldsymbol{x}, t)$$

korištenjem Taylora prvog reda i postavljanjem uvjeta da je derivacija inteziteta po vektoru pomaka jednaka 0, tj. da nema promjene inteziteta dobije se linearni sustav oblika:

$$\boldsymbol{Gd} = \boldsymbol{e} \qquad (3)$$

gdje je

$$\boldsymbol{G} = \sum_W \begin{bmatrix} g_1^2 & g_1 g_2 \\ g_1 g_2 & g_2^2 \end{bmatrix}, \boldsymbol{e} = \sum_W h[g_1, g_2]^T$$

a g_i je oznaka za

$$g_i = \frac{\partial I}{\partial x_i}$$

Rješavanjem jednadžbe (3) po **d** dobijemo:

$$\widehat{\boldsymbol{d}} = \boldsymbol{G}^{-1}\boldsymbol{e}$$

Jednadžba se iterira metodom Newton-Raphson dok ne konvergira. Ovako definiran sustav je poznati Shi–Tomashi–Kanade treker. Međutim, ima jednu manu prikazanu slikom 14 a) koja se očituje povremenim praćenjem točaka koje ne pripadaju subjektu. Autori [49] su nadogradili poznati treker tako da su postavili hipotezu koja kaže da se značajne točke na površini istog čvrstog tijela ne mogu kretati neovisno i proizvoljno, nego zajednički. Zato vrijednost vektora $\widehat{\boldsymbol{d}}$ računaju drugačije:

$$\frac{(\widehat{\boldsymbol{d}} + w\boldsymbol{v})}{(1+w)} \; gdje\ je\ \boldsymbol{v}\ vektor\ zajedničkog\ pokreta\ značajnih\ točki$$

Autori [49] definiraju skup od *n* značajnih točaka oznake $\boldsymbol{p_i}$ te pripadajućih vektora pomaka $\widehat{\boldsymbol{d}}_{\boldsymbol{i}}$**,** za svaku značajnu točku po jedan. Potrebno je konstruirati funkciju koja povezuje značajne točke sa vektorom pokreta značajne točke, tj. računa za svaku značajnu točku njen vektor pomaka. Takva funkcija nosi oznaku $T: R^2 \longrightarrow R^2$,a preslikavanja zapisujemo:

$$\widehat{\boldsymbol{d}}_{\boldsymbol{i}} \approx T(\boldsymbol{p_i}), i = 1, 2, 3, \dots, n \qquad (4)$$

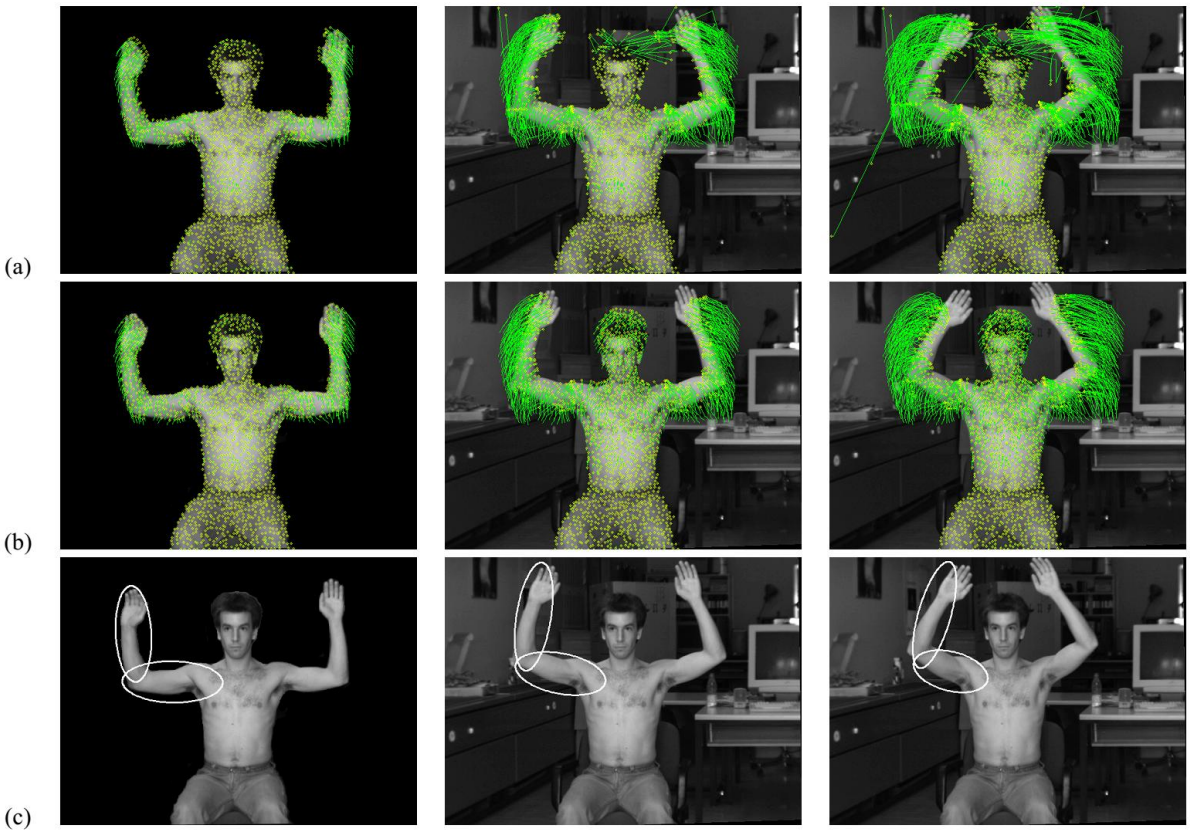
Pri čemu je moguće postojanje n funkcija **T,** a to znači za svaku značajnu točku po jedna pa to pišemo kao T_k. Ako ubacimo jednadžbu (4) u vektorski oblik imao traženi izraz za vektor $\boldsymbol{v}$:

$$\boldsymbol{v} = T(\boldsymbol{p})$$

Za računanje T-a koriste se eng. *radial basis functions* (RBF) koje funkcioniraju na istom principu kao neuralne mreže sačinjene od Gaussa. Izraz kojim se dobije jedna takva funkcija glasi:

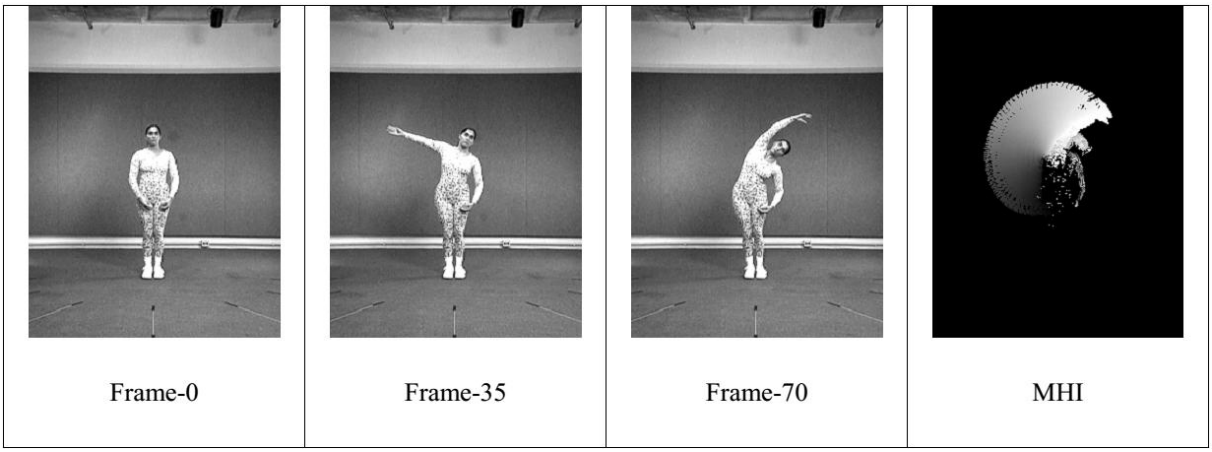
$$f(x) = \sum_i w_i^{\left(-\frac{\|x-c_i\|^2}{\sigma_i^2}\right)} \qquad (5)$$

Jedino što je potrebno za izračunati svaku funkciju (5) je odrediti trojke $\{(w_i, \sigma_i, c_i)\}$, što se postiže hijerarhijskim RBF mrežama.



Slika 14 – preuzeto i [49] – a) Praćenje originalnim Shi–Tomashi–Kanade algoritmom, b) praćenje sustavom [49] i c) blobovi koji se dobiju iz b)

Rad [50] konstruira vektor pokreta razmještanjem blokova piksela.



Slika 15 – preuzeto iz [51] – primjer kako nastaje eng. *motion history image* (MHI). Tamno sivi dio polukruga na posljednjoj slici sekvence predstavlja dio pokreta koji se dogodio vremenski ranije dok je svijetlo sivi dio vremenski mlađi dio pokreta.

Rad [51] je na temelju eng. *motion history image*-a (MHI) – slika 15 - napravio veoma robustan i precizan sustav koji se može koristi za procjenu poze u bilo kojem tipu ljudske aktivnosti kao i za računanje statistike pokreta (npr. koliko puta je napravljen neki pokret) te segmentaciju pokreta. MHI je dosta zahvalna kompleksna konstrukcija ili entitet nastao na osnovu ulaznih podataka zato što daje podatak o pozi i podatak o smjeru kretanja i puno je

robusniji na šum, praćenje itd. od npr. optičkog toka. Postoji jedna mana ovakve konstrukcije, a to je što svi pokreti unutar površine tijela ostaju ne zapaženi. Ovo se može ispraviti tako da se koriste 2 kamere ili da se različito boja dijelove tijela prekrivene odjećom s obzirom na dio koji je ostao neprekriven odjećom (npr. mahanje rukom šaku oboja drugačije nego nadlakticu i podlakticu prekrivenu rukavom). Slika 16 prikazuje skicu arhitekture sustava [51]. Korakom 1 označena je faza izvlačenja siluete iz svake slike sekvence. Za izvlačenje siluete koristi se jednostavan postupak koji najprije označi kao *foreground* sve piksele slike koji su za određenu vrijednost standardne devijacije odmaknuti od *mean* vrijednosti RGB pozadine. Potom se radi dilatacija i eng *region growning* kako bi se eliminiralo sav šum na slici (piksele koji su inicijalno bili proglašeni *foreground,* ali tu ne pripadaju jer ne čine površinu subjekta nego nečega drugog). Korak označen sa 2 radi procjenu poze korištenjem Hu momenata definiranih u [53] tako da oni generiraju deskriptore otporne na translaciju i skaliranje. Kako su Hu momenti različitih redova koristi se Mahalanobis udaljenost definirana u radu [54] kako bi se na sve deskriptore mogla primijeniti ista statistička mjera udaljenosti. Udaljenost je definirana jednadžbom (6):

$$mahal(x) = (x - m)^T K^{-1}(x - m) \qquad (6)$$

gdje je x vektor momenta značajke, m je *mean* vektora momenta, K^{-1} je inverzna, transponirana matrica vektora momenta. Rezultat koraka 2 je prepoznata poza. Korak 3 formira tMHI tako da na svaku siluetu zalijepi točan vremenski žig trenutka u formatu „sekunda.milisekunda“ prema sljedećoj jednakosti:

$$tMHI_{\delta(x,y)} = \begin{cases} \tau \; ako \; je \; trenutna \; pozicija \; siluete \; (x,y) \\ 0 \; ako \; je \; tMHI_{\delta(x,y)} < (\tau - \delta) \end{cases}$$

gdje je τ trenutno vrijeme, a δ je vremenska konstanta (oko nekoliko sekundi) koja kaže maksimalnog trajanja siluete. Ovakav način označavanja siluete je potrebna jer omogućava izračun jednakog MHI-a bez obzira na broj sličica u sekundi. Korak 4 računa gradijent promjene tako da napravi gradijent tMHI-a. Gradijent se računa kompozicijom zasebnih Sobelovih filtra u *x i y* smjeru čime se dobiju prostorne derivacije $F_x(x,y)$ i $F_y(x,y)$. Orijentacija gradijenta u svakom pikselu tHMI-a može se onda zapisati kao omjer:

$$\phi(x,y) = \tan^{-1}\frac{F_y(x,y)}{F_x(x,y)}$$

pri čemu u obzir ne dolaze vrijednosti pikseli sa prevelikim kontrastom s obzirom na susjedne jer bi bitno poremetili izračun gradijenta. Nakon računanja gradijenta izračunati su svi podaci

potrebni za posljednji korak 5, a to je korak prepoznavanja aktivnosti. Na osnovu izračunatih gradijenata izračuna se globalni smjer aktivnosti prema jednakosti:

$$\bar{\phi} = \phi_{ref} + \frac{\sum_{x,y} kutnaRalzika\big(\phi(x,y), \phi_{ref}\big) * norma(\tau, \delta, t, MHI_{\delta(x,y)})}{\sum_{x,y} norma(\tau, \delta, MHI_{\delta(x,y)})}$$

gdje je ϕ_{ref} maksimalna vrijednost iz histograma svih orijentacija, $\varphi(x, y)$ je mapa orijentacija izračunata operacijom konvolucije gradijenta, $norma(\tau, \delta, MHI_{\delta(x,y)})$ je noramlizirani tMHI korištenjem vremenskog žiga τ i konstante δ te $kutnaRalzika\big(\phi(x,y), \phi_{ref}\big)$ označava kutnu razliku trenutne orijentacije od referente. Najvažniji korak ovog sustava je korak 6 sa slike 17 koji određuje segmente subjekta i pozadine. To čini algoritmom koji se može sažete u korake:

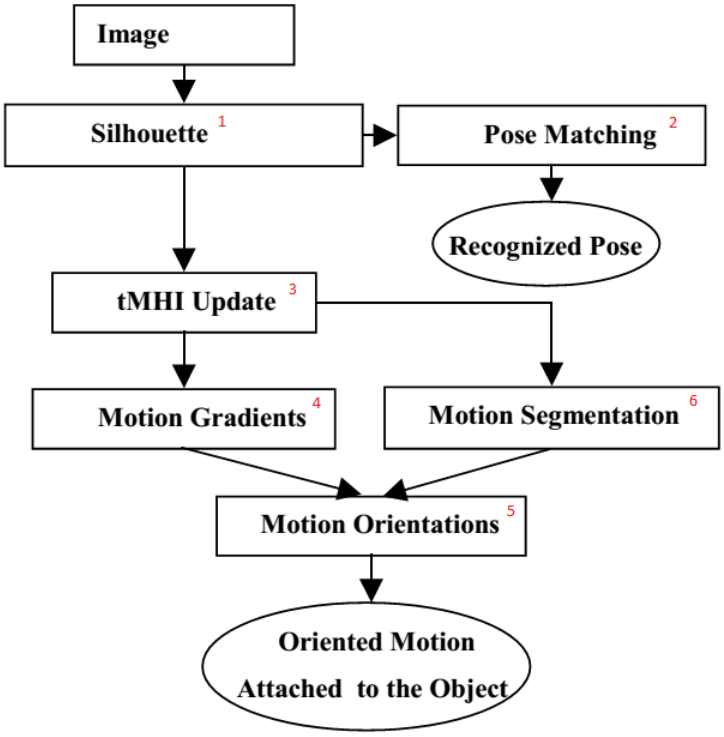
I. Pretraži sve tHMI-e dok ne nađeš piksele pridjeljenje zadnjoj silueti (slika 17 b)).

II. Hodaj po rubu i traži sve piksele koji su dio pokreta ali su izvan trenutne siluete jer su se dogodili vrlo malo vremena prije (npr. vrijeme između dvije slike sekvence.). Sve piksela nađene takvim pristupom obojaja eng. *downfill* algoritmom (slika 17 b)).

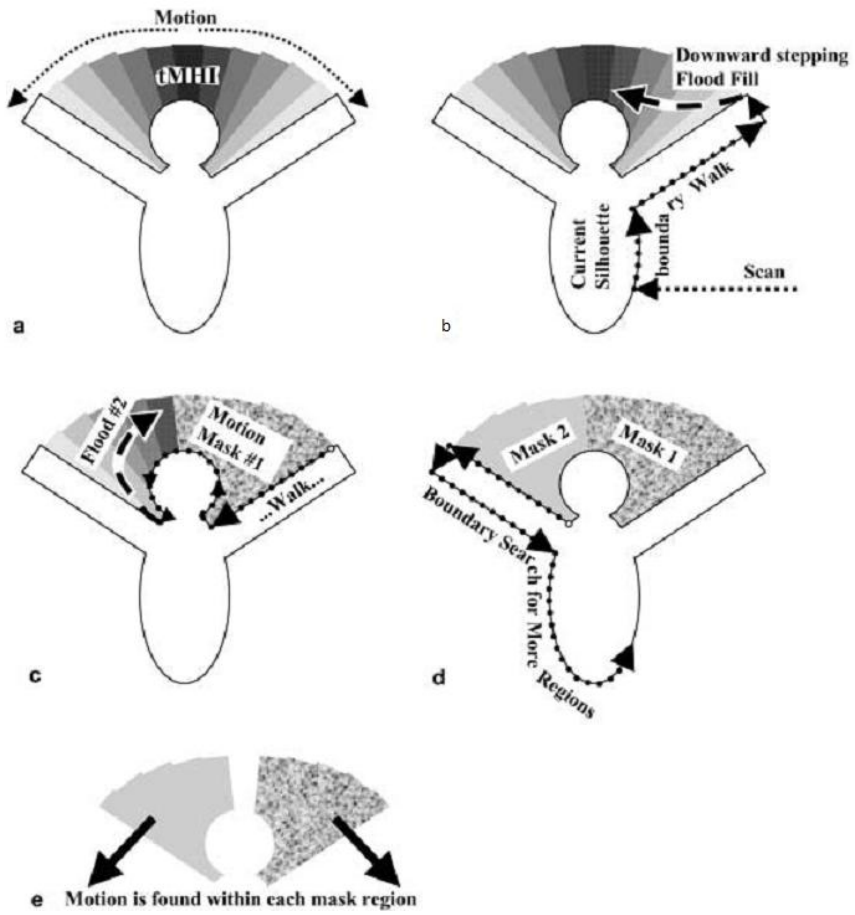
III. Spremi segmente već prebrisane korakom II (slika 17 c i d)

IV. Ponavljaj korake I – III dok ne obiđeš cijelu trenutnu siluetu.

V. Izračunaj orijentaciju pokreta svakog pronađenog segmenta (slika 17 e)).



Slika 16 – preuzeto iz [52] – skica arhitekture sustava [52]



Slika 17 . preuzeto iz [52] – segmentacija subjekta

Kako je moguće da se u sceni pojavi subjekt koji stoji ili sjedi, on neće biti segmentiran kao dio *foreground-a.* Također, moguće je i da subjekt dosta sliči na *background* po nekom odabranom kriteriju te on neće biti segmentiran kao dio *foreground-a*. Zbog toga neki autori koriste 3 slike kod računanja osnovne pozadine.

Autori [54] bili su jedni od prvih koji su uspješnim korištenjem i izgleda i pokreta subjekta postigli puno bolje rezultate prepoznavanja nego korištenjem samo izgleda ili samo pokreta. Njihov sustav opisan u [54] ima mogućnost prepoznavanja hoda i u slučajevima kada je rezolucija čovjeka samo 20 x 15 piksela te se odlikuje vrlo niskom stopom krive klasifikacije. Sposoban je prepoznati hod i u uvjetima kiše ili snijega na slici. Međutim, mana mu je da zahtjeva punu figuru čovjeka, bez preklapanja ili zaklonjenosti nekih dijelova tijela. Rad najprije definira 5 karakterističnih slika pomoću sljedećih jednakosti:

$$\Delta = abs(I_t - I_{t+1})$$

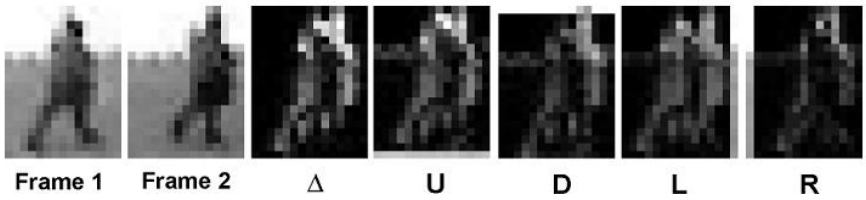
$$U = abs(I_t - I_{t+1} \uparrow)$$

$$L = abs(I_t - I_{t+1} \leftarrow)$$

$$R = abs(I_t - I_{t+1} \rightarrow)$$

$$D = abs(I_t - I_{t+1} \downarrow)$$

Slova su prvo slovo engleskih riječi za smjerove. Svaka od pet slika se dobije tako da se oduzmu dvije uzastopne sive slike jedna od druge, ali druga slika u izračunu se pomakne za jedan piksel u smjeru za koji se izračun radi (osim prve – nema smjer u izrazu). Primjer kako slika izgleda nakon primjene svakog od filtera dan je na slici 18:

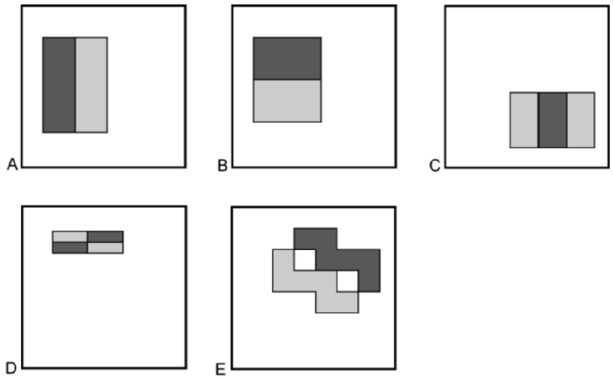


Slika 18 – preuzeto iz [54] – rezultat primjene filtra na sliku.

Potom se definiraju 3 filtra. Prvi računa zbroj razlika između slike Δ i svih ostalih:

$$f_i = r_i(\Delta) - r_i(S) \qquad (7)$$

tako da definira funkciju $r_i()$ preko pravokutnih filtera (slika 19) na način da se oduzima zbroj piksela unutar tamnih pravokutnika od zbroja piksela unutar svijetlih pravokutnika. dakle, ne ulaze svi pikseli iz prozora promatranja u izračun, samo oni prekriveni pravokutnicima i to na način upravo opisan. Filter (7) je potreban kako bi se utvrdio smjer promjene na dvije uzastopne slike.



Slika 19 [54]– pravokutni filteri – kod tri pravokutnika tami se pomnoži sa faktorom 2 kako bi se kompenziralo oduzimanje sa dva svijetla pravokutnika

Filter (8) uspoređuje sume dobivene jednim od pravokutnih filtera sa slike 19, ali unutar iste slike sekvence:

$$f_j = \phi_j(S) \qquad (8)$$

Zadatak ovog filtera je mjerenje u kojem smjeru subjekt ima tendenciju izvršiti pokret.

Posljednji filter pokreta je (9). Njegov zadatak je mjerenje veličine pokreta u jednoj od pomaknutih slika:

$$f_k = r_k(S) \qquad (9)$$

gdje je $r_k()$ jedan od pravokutnih filtera sa slike 19.

Definira se i filter izgleda koji ima oblik:

$$f_m = \phi(I_t) \qquad (10)$$

Pravokutni filteri nisu unaprijed dani niti definirani nekim izrazom nego se primjenjuje algoritme strojnog učenja kako bi oni konstruirali optimalne klasifikatore koji najbolje razdvajaju pozitivne i negativne primjere za učenje. Klasifikatori se definiraju kao binarna funcija:

$$C(I_t, I_{t+1}) = \begin{cases} 1 \; ako \; je \; \sum_{i=1}^{N} F_i(I_t, \Delta, U, L, R, D) > 0 \\ 0 \; inače \end{cases}$$

Funkcija F je također binarna funkcija oblika:

$$F_i(I_t, I_{t+1}) = \begin{cases} \alpha \; ako \; je \; f_i(I_t, \Delta, U, L, R, D) > t_i \\ \beta \; inače \end{cases}$$

gdje je f_i jedan od filtera pokreta ili izgleda definiranih jednadžbama (7) – (10), a t_i je realan broj koji predstavlja prag.

2.3.3 Segmentacija temeljena na boji i obliku subjekta

Segmentacije temeljene na izgledu pretpostavljaju da se izgled subjekta razlikuje od izgleda pozadine. Također vrijedi i da se subjekti međusobno razlikuju po izgledu. Većina pristupa temeljenih na izgledu konstruira model svakog čovjeka na slici i potom radi jedno od sljedeća dva izbora:

I. konstruiraju modele segmentiranih *foreground* objekta u trenutnoj sceni i uspoređuju ih sa modelima ljudi

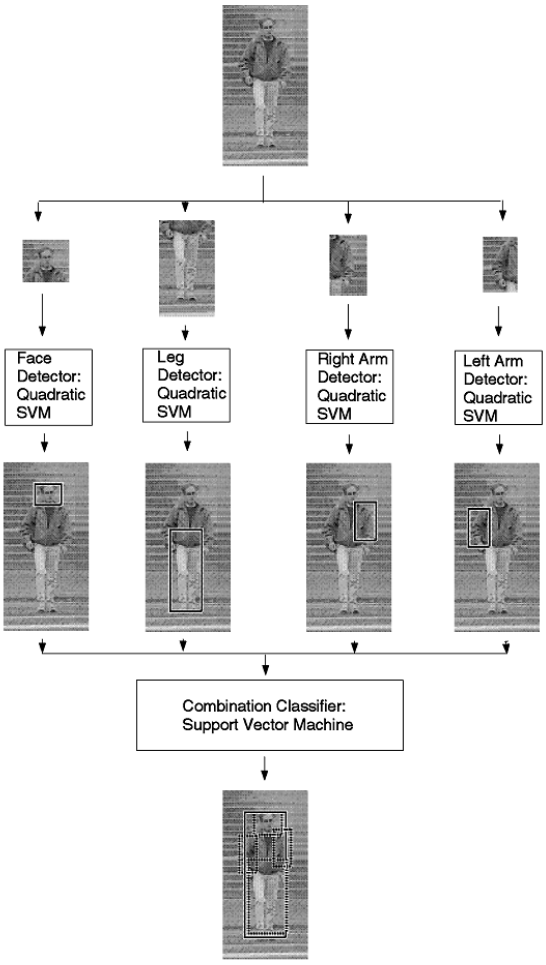
II. direktno na osnovu modela ljudi segmentiraju piksele trenutne scene

Neki od ovako konstruiranih sustava obrade su neovisni o trenutnom kontekstu obrade što znači da metode obrade koriste generalni model ljudi za razliku od sustava gdje se koristi model ljudi naučen direktno na sekvencama koje se analiziraju. Zbog toga se segmentacija temeljena na izgledu subjekta dijeli na dvije grupe: sustavi koji koriste trenutni kontekst i one koji funkcioniraju bez korištenja informacija iz konteksta.

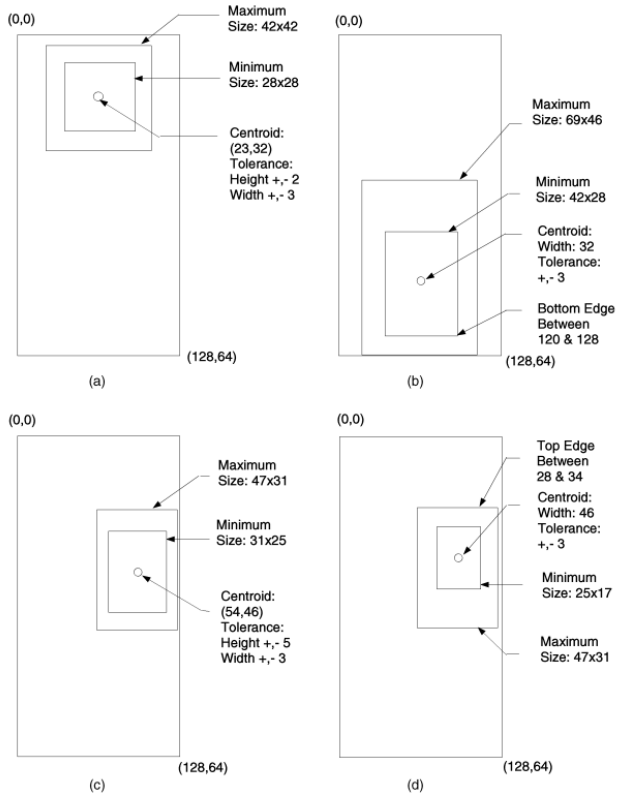
2.3.3.1 Segmentacija temeljena na izgledu bez korištenja konteksta

Rad [55] detektira ljude na osnovu samo jedne slike. Sustav, prikazan slikom 20, ima ishodište koordinatnog sustava slike u gornjem lijevom kutu te koordinate rastu prema dolje i prema desno. Sustav najprije napravi prozor obrade veličine 128 x 64 tako da mu je gornji lijevi vrh u ishodištu upravo opisanog koordinatnog sustava. Potom se donosi odluka o klasifikaciji prozora kao pozitivnog ili negativnog ovisno da li sadrži osobu, odnosno dio osobe ili ne. Proces koji donosi odluku sastoji se od detektora komponenti ljudskog tijela i to: detektor glave, detektora nogu, detektora lijeve i detektora desne ruke. Detektori obrađuju sliku prozora tako da primjene Haar *valiće* (eng. *wavelet*) te klasificiraju rezultanti vektore podataka. Klasifikacija vektora podataka se radi korištenjem eng. *quadratic Support Vector Machines*-a (SVM) koji se uče u fazi treninge na sve moguće oblike spomenutih dijelova tijela. Proces klasifikacije odvija se tako da svaki od detektora izračuna svoj eng.*score* (kako ga je označen u radu) te onaj sa najvećom vrijednošću kaže o kojem djelu tijela se radi u prozoru obrade. Potom se svi prepoznati dijelovi iz svih prozora obrade predaju kombinacijskom klasifikatoru koji je sastavljen od linearnog SVM-a. On je zadužen da spoji sve dijelove u jedan. Sve navedeno prikazano je slikom 20. Opisani proces se ponavlja za svaki prozor obrade tako da se prozor translatira u desno i prema dolje te se sadržaj prozora skalira u rasponu vrijednosti od 0.2 do 1.5. Ovo se radi kako bise omogućilo detektiranje osoba različitih veličina u bilo kojem djelu slike. Kod prozora obrade dijelovi tijela ne mogu biti bilo gdje unutar prozora. Npr. ograničeni su nekim hipotezama poput one da se snimaju samo uspravne osobe te noge nikada neće biti vodoravno položene. Ta informacija se koristi tako da se definiraju unutarnje prozori unutar prozora obrade sa unutarnjim i vanjskim margine unutarnjeg prozora (prikazane slikom 21) i odstupanje od centra unutarnjeg prozora kako bi se ograničilo 2D prostor u kojem se potencijalno kriju dijelovi tijela.

Rad [56] uspješno segmentira i prati više subjekata odjednom na dinamičkoj pozadini, npr. igrači na utakmici. Rad operira u HSV modelu boja zato što ovaj model boja izdvaja osvjetljenje u zasebnu komponentu. Rad koristi filtere čestica tako da ih navodi, korištenjem modificiranog Adaboost algoritma, na regije od interesa (*foreground* koji bi mogao predstavljati subjekte). Međutim, multimodalne funkcije distribucije vjerojatnosti nastaju praćenjem više subjekata odjednom ne rade dobro sa filterima čestica pa autori koriste princip konstruiranja po jednog filtera za svaki subjekt. Filteri se konstruiraju neovisno osim u trenutku pridjeljivanja težinskih faktora svakome od njih. Ova tehnika poznata je kao eng. *mixture particle filter* (MPF). Autori ovako konstruiran algoritam nazivaju eng. *Boosted Particle Filter* (BPF). Nadalje, polje stanja sustava modelira se Markovljevim procesima.

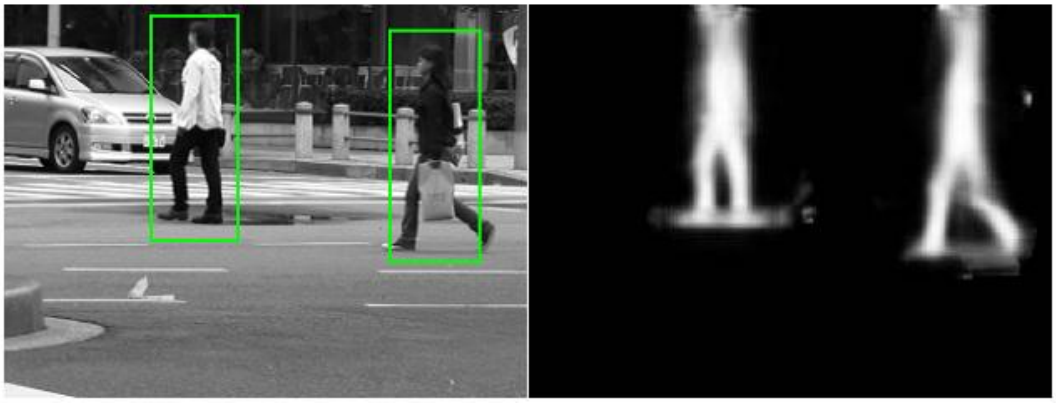


Slika 20 [55] – opis sustava [55] preko dijagrama toka obrade

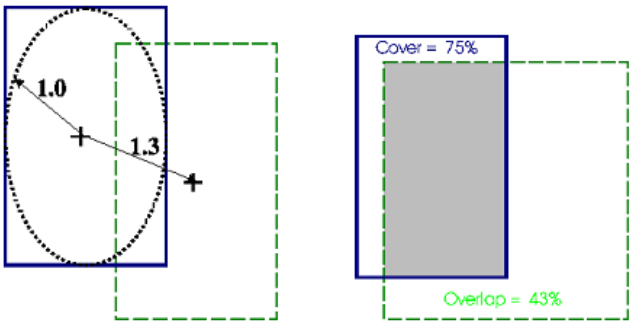


Slika 21 [56] – slike prikazuju dozvoljenu površinu lokacije: a) glave, b) nogu c)desne ruke, d) lijeve ruke

Rad [64] radi praćenje velikog broja subjekta koji se međusobno preklapaju u iznimno dinamičnim scenama sa pozadinom koja je izrazito pretrpana sadržajem (ulice velegrada sa velikom količinom automobila, uličnih rekvizita itd.). Autori [64] kreću od hipoteze da ne postoji niti jedna metoda ili značajka objekta koja bi pojedinačno riješila problem praćenja u navedenim uvjetima. Stoga su konstruirali novi algoritam koji kombinira lokalne i globalne značajke sa tzv. eng. *top-down* segmentacijom (segmentacija koja ne radi klasično oduzimanje pozadine i segmentiranje *foreground-a* nego nastoji prepoznati na slici zadani oblik i boju subjekta). Primjer takve segmentacije dan je na slici [57]. Zahvaljujući opisanom algoritmu autori su u stanju spustiti se do nivoa piksela tijekom obrade. Način kako određuju preklapanje površina zelenih pravokutnika sa slike 57 dan je na slici 58.



Slika 21 – preuzeto iz [64] – primjer top-down segmentacije kakvu koristi rad [64]



Slika 22 – preuzeto iz [64] – lijevo je relativna udaljenost dva pravokutnika praćenja, desno preklopljena površina

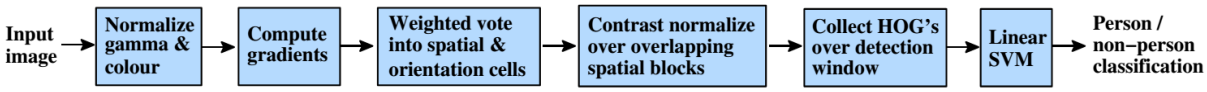
Rad [65] definira najprije *apriori* model oblika ljudske siluete te je prikaže kao Boltzmann distribuciju u Markovljevom polju. Rad uspješno rješava preklapanje subjekata.

Rad [66] je dao iznimno velik doprinospolju praćenja tako da je eksperimentalno dokazao koje značajki oblika ili boja najbolje rješavaju problem robusnog praćenja ljudi, a za testiranje performansi pojedinih značajki koristili su linearni SVM treniran da prepozna ljudsku konturu. Zaključili su da je najbolja značajka eng. *Histograms of Oriented Gradient* poznat bolje pod skraćenicom HOG te su predložili i vlastiti sustav temeljen na HOG

deskriptorima, također dan u radu [66]. Osnovna hipoteza je da se lokalni izgled ili oblik objekta može najbolje (najtočnije) predstaviti distribucijom lokalnog inteziteta gradijenta, tj. smjera ruba, čak i u slučajevima kada se ne zna točna pozicija ruba ili ishodišta gradijenta. Navedena hipoteza implementira se tako da se originalna slika podijeli na male površine, reda veličine nekoliko piksela. Takve površine nazvana su *ćelije* te se za svaku računa jednodimenzionalni histogram ili gradijenta smjera ili orijentacije ruba. Da bi se postigla robusnost s obzirom na iluminaciju, sjenu i slične umjetne artefakte koristi se korak normalizacije vrijednosti unutar ćelija nakon samog računanja gradijenta. Normalizacija se obično provodi tako da se izračuna ukupna energija svih histograma ćelija na površini sastavljenoj od više ćelija te se sa tako dobivenom vrijednošću radi normalizacija svake vrijednosti ćelije. Površine sastavljene od više ćelije zajedno koje su prošle korak normalizacije autori nazivaju *Histogram of Oriented Gradient* (skraćeno HOG) deskriptori. Ovako dobivenim HOG-ovima popoča se prozor detekcije tako da se rubovi HOG-ova preklapaju te se na njih pusti klasični SVM. HOG je veliko praktično značenje dobio kao dio SIFT algoritma (eng. *Scale Invariant Feature Transformation*) u sklopu kojeg se koristi za dobivanje deskriptora uzoraka koji služe povezivanju ključnih točaka invarijantnih na skaliranje. Rad [67] je pokazao da se relativno jednostavnim deskriptorima (a jednostavni implicira i veliku brzinu obrade te manju potrošnu resursa) naziva eng. *keypoint* deskriptori može dobiti solidne rezultate praćenja. Međutim, radom [66] dokazano je da su jednostavni deskriptori, iako lakši za računanje 1-2 reda veličine lošiji klasifikatori od tzv. teških ili složenih deskriptora zbog toga što ne postoji niti jedan *keypoint* deskriptor koji je u stanju sa velikom pouzdanošću detektirati strukturu ljudskog tijela. **Važan** zaključak proizašao iz uspješnosti HOG/SIFT kombinacije kaže da je za uspješnu detekciju, a potom praćenje ljudi na slici bitno sljedeće:

1. Grubo prostorno uzorkovanje na nivou slike
2. Fino ili detaljno, precizno uzorkovanje orijentacije subjekta
3. Značajna ili jaka fotometrijska normalizacija

Gornja tri uvjeta garantiraju praćenje udova i dijelova tijela u njihovom normalnom kretanju i promjeni boje ili oblika istih uslijed različitih promjena na slici. Sve navedeno za rad [66] vrijedi samo ako subjekt vrši radnje u okomitom položaju.



Slika 23 – preuzeto iz [66] – tijek obrade podatak u sustavu [64]

Autori [66] su objavili unaprjeđenje rada [66], dano radom [68], a unaprjeđenje se sastojalo od definiranja eng. *motion* HOG-a ili pokretnog HOG deskriptora. Pokretni HOG radi praćenje na videu snimljenom kamerom koja nije statička ili u uvjetima pozadine koja se miče.

2.3.3.2 Segmentacija temeljena na izgledu i obliku sa korištenjem konteksta

Metode temeljene na kontekstu obrađuju podatke tako da na skupu slika nauče model ljudi koji onda koriste na sljedećim slikama te iste sekvence kako bi ili otkrili ili segmentirali piksele *foreground-a.* Metode rade na nivou piksela ili na nivou regije. Ukoliko operiraju na nivou piksela onda je njihov zadatak računanje vjerojatnosti da detektirani piksel *foreground-a* pripada modelu čovjeka koji je već postavljen. Ako operiraju na nivou regije onda računaju vjerojatnost da regija unutar prozora promatranja odgovara odabranom ili postavljenom modelu ljudi.

Rad [57] predstavlja izniman doprinos polju te je prekretnice u metodama praćenja. Rad najprije definira prostor ili polje iz kojeg će se uzimati karakteristične vrijednosti (eng. *features*). Potom rad definira dvije funkcije: ciljani model subjekta i ciljani subjekt. Ciljani model je funkcija razdiobe vjerojatnosti ili PDF u izabranom prostoru stanja (npr. vjerojatnost razdiobe boje subjekta) sa središtem razdiobe u točci slike koja ima koordinate 0,0 (ishodište), a ciljani subjekt je također PDF kojemu je središte na prostornoj lokaciji y. Oba PDF-a se procjenjuju iz uzoraka za učenje. Da bi se trajanje računanja PDF-ova minimiziralo radi se sa histogramima koji imaju m binova. Iz navedenog proizlazi:

$$ciljani\ model: \quad \hat{q} = \{\hat{q}_u\}_{u=1\ldots m} \qquad \sum_{u=1}^{m} \hat{q}_u = 1$$

$$ciljani\ subjekt: \quad \hat{p}(y) = \{\hat{p}_u(y)\}_{u=1\ldots m} \quad \sum_{u=1}^{m} \hat{p}_u = 1$$

Cilj je pronaći funkciju sličnosti između $\hat{q}$ i $\hat{p}$:

$$\hat{\rho}(y) \equiv \rho[\hat{p}(y), \hat{q}]$$

čiji maksimum signalizira poklapanje ciljanog modela i subjekta, a to pak znači da je pronađen PDF koji najbolje modelira subjekt. Traženje maksimuma nekom od metoda gradijenta je računalno dosta teško implementirati te se zato površina subjekta obradi izotropnim *kernel-om,* a potom se težinski faktor takvog *kernel-a* koristi u definiranju prezentacije značajnih točki te funkcija $\hat{\rho}(y)$ postane glatke na domeni y. Na ovakav način lakše je računati gradijent, a samim time i maksimum funkcije.

Rad [58] radi praćenje temeljeno na filteru čestica. Rad koristi boju i histograme orijentacije ruba kao značajke praćenja. Također koristi i integralne slike, višestruke značajke te pseudo proizvoljno uzorkovanje čestica kako bi ubrzao proces praćenja i povećao otpornost na grešku.

Kako bojanje subjekta samo jednom bojom nije dovoljno za robusno praćenje, koriste se i dodatne informacije uz samu boju ili oblik subjekta. Npr. rad [59] koristi entitet naziva eng. *co-occurrence matrix*. Matrica je matematički definirana jednadžbom (11):

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & ako\ je\ I(p,q) = i\ te\ I(p+\Delta x, q+\Delta y) = j \\ & 0\ inače\end{cases} \tag{11}$$

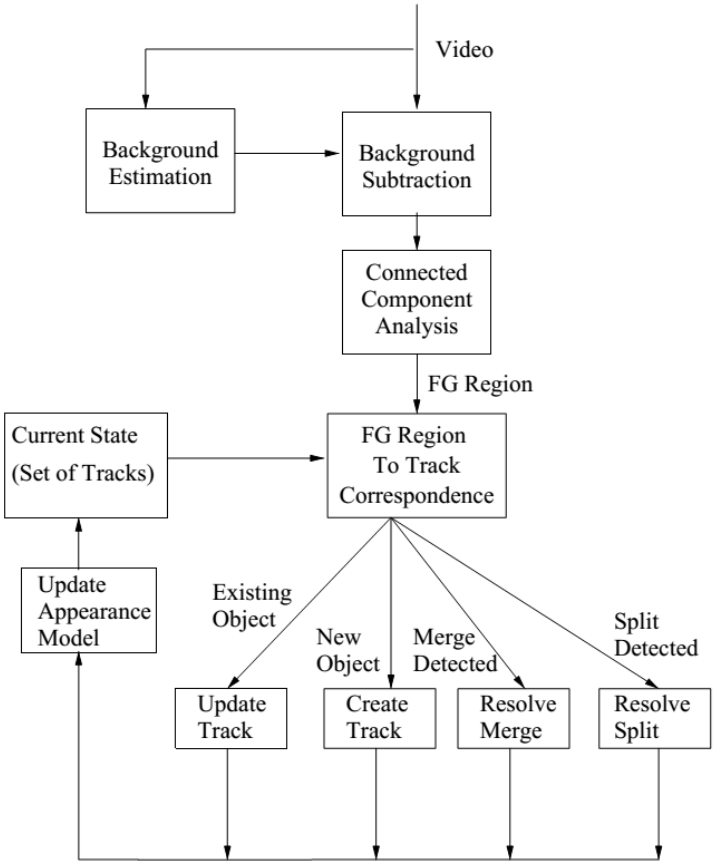
gdje je Δx pomak po širini slike, a Δy je pomak po visini slike I za koju se matrica definira, a (p,q) koordinate matrice, odnosno slike. Dakle, matrica kompaktno zapisuje vrijednosti piksela (nijanse sive, boja ili nešto treće) slike koje se jednake, ali prostorno međusobno dislocirane.

Rad [60] koriste drugačiju metodu postizanja prostornih informacija o subjektu, uz boju i oblik kao primarne izvore informacija za praćenje subjekta. Rad definira poznati M_2 treker, a prostornu informaciju dobije podjelom tijela subjekta na regije te potom regije predstavlja histogram boja ili mješavinom Gaussa. Primjer kako to izgleda dan je na slici [59]:

Slika 23 – preuzeto iz [60] – lijevo je 6 slika snimljenih iz 6 različitih perspektiva, desno su 4 slike kao primjer različitog bojanja dijelova tijela kod subjekata na slikama (0,0), (0,1), (1,1) te (2,0) desno.

Rad [61] definira modele izgleda korištenjem kojih uspješno prati više subjekata istovremeno uz prisustvo parcijalnih ili potpunih preklapanja. Sustav najprije oduzima

pozadinu korištenjem konstruiranih modela pozadine. Ono što ostane procesom oduzimanje pozadine predstavlja podatke za sustav praćenja koji funkcionira dvoslojno (sustav dan na slici 60). Prvi sloj radi trajektorije koje povezuju segmente istog subjekta kroz sekvencu, a drugi sloj na osnovu modela izgleda razrješuje preklapanje trajektorija tako da je rezultat drugog koraka skup trajektorija koje, svaka za sebe, neovisno opisuju putanju subjekata kroz sekvence. Posljednja operacija sustava je odbacivanje trajektorija koje ne prate subjekte (nego npr. šum slika) te spremanje informacija koja opisuju trajektorije (centar segmenta čiju putanju trajektorija opisuje) u PETS XML formatu.

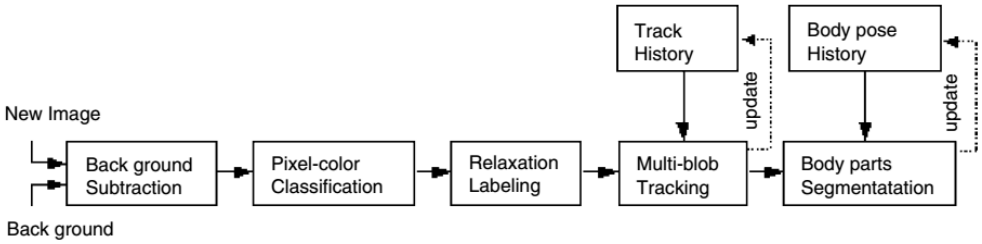


Slika 24 – preuzeto iz [61] – arhitektura sustava

Rad [62] koristi EM algoritam kako bi segmentirao subjekt u površine sličnih boja te svaka boju svrsta u zasebnu klasu. Potom radi praćenje tako da prati klase boja na razini slike korištenjem principa maksimalne *posteriori* vjerojatnosti.

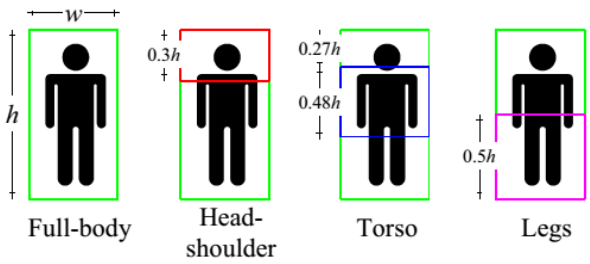
Rad [63] koristi boju na tri različita nivou obrade podataka u sustavu: nivo jednog piksela, nivo *blob*-a i novo subjekta ili objekta. Na nivou piksela koristi se Gaussova mješavina kako bi se treniralo sustav i klasificiralo svaki piksel na osnovu boje. Na nivou bloba koristi se eng. *attribute relational graph.* ARG je struktura koja predstavlja sliku, baš kao što to rade: matrice, integralne slike, matrice susjedstva itd.. Međutim ARG bilježi odnose između objekata slike i same objekte. To znači sa se slika najprije segmentira te se ARG

konstruira na način da je informacija koja opisuje segmentirani objekt pridružena vrhu ARG-a (eng. *vertex*), a rub ili eng. *edge* nosi informaciju u kakvom su odnosu segmentirani objekti međusobno. Npr. ljudsko lice se predstavi ARG-om tako da oči, usne, nos čine segmentirane površine, a kut koji svaka od površina zatvara sa centrom mase lica predstavlja rub ili vezu. Rad koristi ARG kako bi spojio piksela u koherentne blobove te predstavio veze između blobova. Sustav radi praćenje temeljem dvoslojnog procesa koji najprije spaja susjedne blobove korištenjem više slika sekvence te potom pridjeljivanje ili povezivanje blobova sa dijelovima tijela koji bi oni trebali predstavljati unutar iste slike. Kako ovakav način funkcioniranja podsjeća na eng. *multi-target, multiassociation tracking,* autori su svoj sustav nazvali „ARG–MMT“. Sustav sa velikom vjerojatnošću ispravno klasificira radnje poput: 'udaranja šakom', 'rukovanja', 'nasilno guranje drugog subjekta', 'grljenja drugog subjekta'.

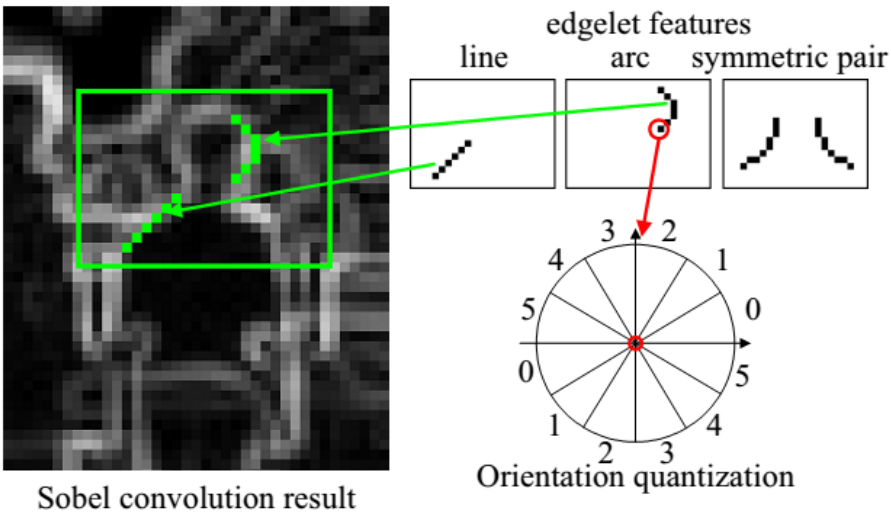


Slika 25 – preuzeto iz [63] - arhitektura sustava

Rad [69] ima sposobnost praćenja više subjekta iz statičke slike koji su međusobno parcijalno preklopljeni. Pretpostavka je da su uspravni i da njihova silueta ne izlazi više od 30° iz ravnine snimanja (subjekti ne smiju biti zakrenuti više od tog kuta naprijed ili nazad u odnosu na ravninu snimanja) te da kamera mora biti unutar 45° u odnosu na ravninu površine po kojoj hodaju subjekti snimanja (tlo). Rad koristi prezentaciju subjekta temeljenu na dijelovima tijela kako su definirani slikom 26. Sustav se temelji na metodi parcijalnog praćenja koja ima dva koraka: detekcija samo nekih dijelova te sastavljanje, tj. zaključivanju o izgledu cijele osobe na osnovu detektiranih dijelova. Detektori dijelova se uče korištenjem Adaboost metode uz pomoć nove vrste značajki tzv. eng. *edgelet* čiji primjer je dan slikom 27 te se rezultati parcijalnih detektora kombiniraju kako bi se konstruirao model vjerojatnosti. Zato ovaj rad spada u MAP ili eng. *maximum a posteriori estimation* grupu radova.



Slika 26 – preuzeto iz [69] – dijelovi tijela kako ih definiraju autori [69]



Slika 27 – preuzeto iz [69] – primjer eng. *edgelet* značajke. Dozvoljeni oblici kod ove vrste značajke su: linija, krug, polovica, četvrtina i osmina kruga te sve njihove simetrične kombinacije.

2.3.4 Segmentacija temeljena na dubini slike

Odjeljivanje pozadine od *foreground-a* korištenjem podataka o dubini sadržaja na slici temelji se na pretpostavci da se ljudi ističu u 3D prostoru slike od ostalog sadržaja u istom prostoru. Metode ovakvog pristupa segmentaciji mogu se većinom svrstati u dva skupa: direktni rad sa 3D podacima te na indirektni rad sa 3D podacima na način da se kombinira više kamera istovremeno nakon faze izvlačenja značajki. Napredak u ovom polju ostvaren je većinom zbog sve veće brzine današnjih računala u odnosu na ona prije 10-ak godina jer smo danas u stanju paralelno (istovremeno) obrađivati više video sekvenci iste scene iz kojih se dobivaju 3D podaci.

Kako je u uvodu ovog poglavlja 2. kazano da se neće raditi recenzija radova koji nisu temeljeni na monokularnom video 2.3.4 se preskače.

2.3.5 Temporal Correspondences

Praćenje u vremenu ili eng. *temporal correspondences* definira se kao pronalaženje subjekta prepoznatog u prošloj slici sekvence unutar trenutne slike sekvence pri čemu subjekt praćenja može biti opisan bilo kojom reprezentacijom iz tablice 2. Tj., tražimo trenutno stanje subjekta u trenutnoj slici sekvence na osnovu prošlog, zadnje poznatog stanja pri čemu se pri stanju većinom misli na njegov položaj, međutim u nekim radovima podrazumijevaju se i drugi atributi poput boje, oblika itd. kao stanje koje se prati kroz sekvencu.

Kod većine radove pronalaženje stanja se modelira raznim modelima predikcije tako da se trenutno stanje predvidi i uspoređuje, korištenjem izabrane metrike, sa stvarnim stanjem na trenutnoj slici. Predikcija koristi regije interesa ili eng. *region od interest* u polju mogućih stanja kao i na samoj slici sekvence čime smanjiva ukupnu količinu podataka koje treba

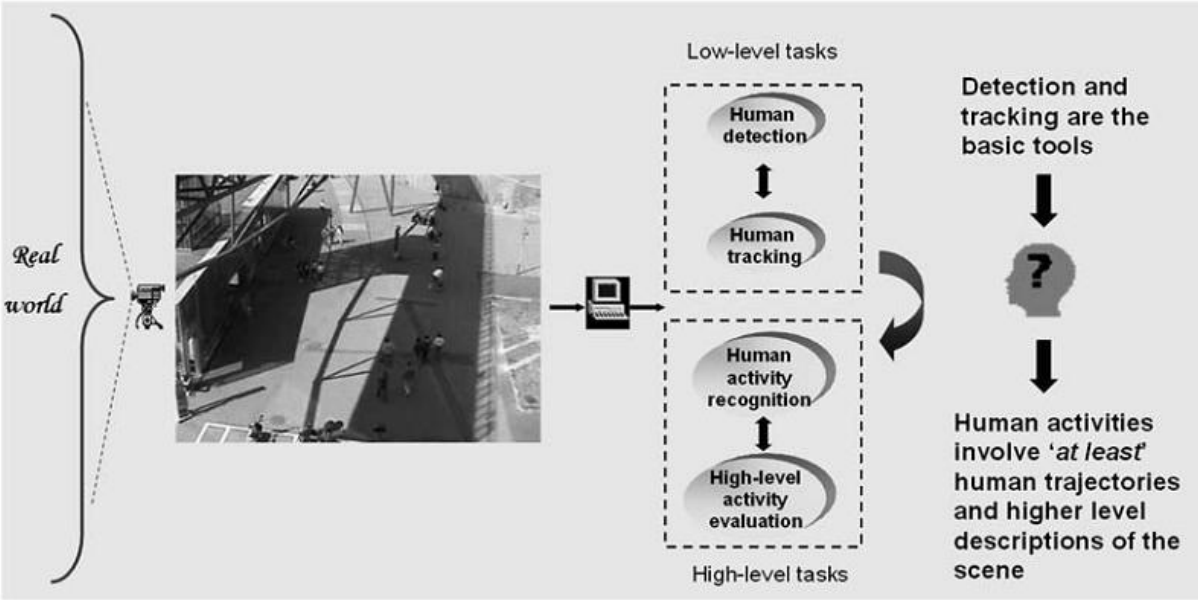
procesirati što pridonosi brzini sustava. Najpoznatija metoda predviđanja za unimodalnu *aposteriori* distribuciju je Kalmanov filter koji ima sposobnost ocjenjivanja greške predviđanja. Neki autori koriste grešku predviđanja Kalmanovog filtera kako bi locirali regiju interesa. U slučajevima multimodalne distribucije najboljim se pokazao Kondenzacijski algoritam (već spominjan i objašnjen u uvodu u 2.3 poglavlje).

Bitna razlika postoji među procesima *temporal correspondences* s obzirom da li je subjekt praćenja preklopljen s drugim subjektom ili ne te se stoga radovi dijele u dvije osnovne grupe: *temporal correspondences* prije i poslije preklapanja te *temporal correspondences* tijekom preklapanja.

2.3.5.1 Temporal correspondences prije i poslije preklapanja

U radovima koji rade vremensko praćenje prije i poslije preklapanja potrebno je napravit model svakog subjekta prije samog praćenja subjekta.

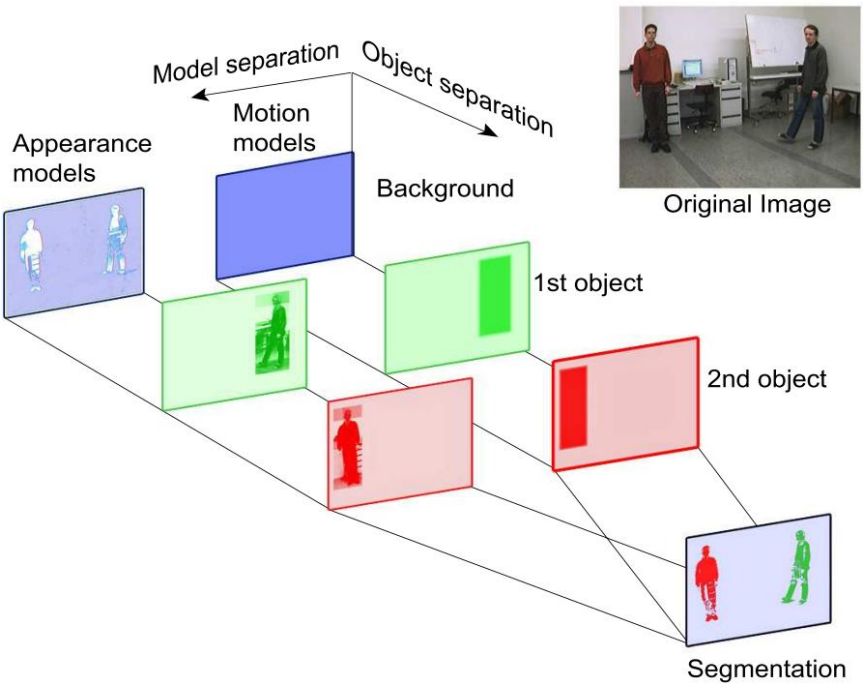
Rad [70] definira model ponašanja pješaka koji se temelji na eng. *Discrete Choice Models* ili DCM modelima, a koji kombinira sa metodama obrade slike kako bi se postiglo najprije prepoznavanje subjekta na slici, a potom i praćenja. Slika 28 daje imena globalnih modula sustava. Autori [70] su došli do zaključka da se ne može napraviti složeni sustav koji samo na osnovu podataka sa slike zaključuje o vrsti aktivnosti nego da je potrebno i dodatno znanje o promatranom subjektu ili skupu subjekta, a koje se mora ugraditi posebno u sustav. Stoga je njihova ideja i doprinos bila napraviti matematički okvir koji najbolje opisuje ponašanje pješaka, a koji se onda može ugraditi u bilo koji sustav koji rade praćenje i analizu radnji pješaka.



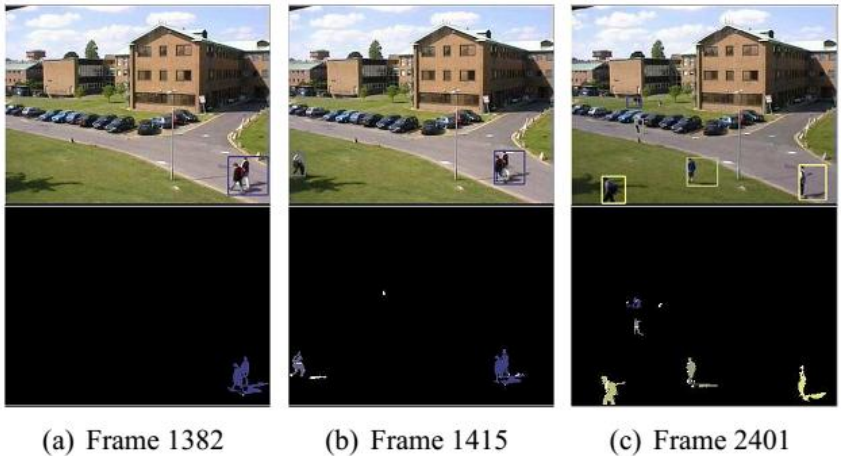
Slika 28 – preuzeto iz [70] – moduli sustava [70]

Rad [71] koristi Bayesovu klasifikaciju na nivou piksela kako bi sagradio *framework* za praćenje koji segmentira sliku na osnovu izgleda i kretanja subjekta. Također , koriste i mješavinu Gaussa kako bi izgradio model boje subjekta. Sustav radi u realnom vremenu, uspješno rješava preklapanja i robustan je na promjenu osvjetljenja. Slika 29 prikazuje skicu sustava, a slika 30 rezultate ovakvog sustava. Sustav procesira podatke u sljedećim koracima:

a) Izračuna vjerojatnost *apriori* svih modela koje koristi
b) Segmentira sliku uz pomoć Bayesove klasifikacije na razini piksela
c) Pronalazi povezane segmente i pridjeljuje ih jednom subjektu
d) Dodaje, briše ili rastavlja subjekta sa slike
e) Razrješava preklapanje
f) Ažurira sve korištene modele



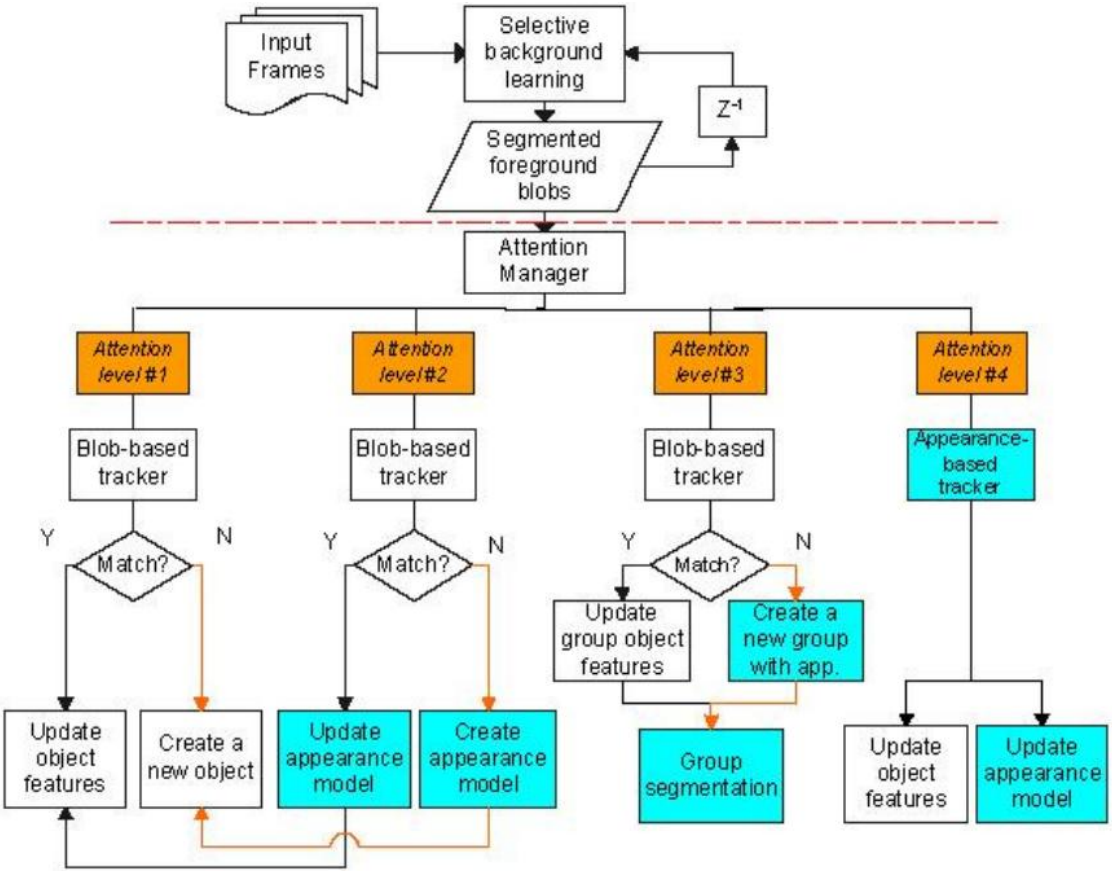
Slika 29 – preuzeto iz [71] – skica funkcioniranja sustava [71]



Slika 30 – preuzeto iz [71] – rezultati obrade sustava [71]

2.3.5.2 *Temporal correspondences tijekom preklapanja*

Rad [72] predlaže hibridni pristup problemu rješavanja preklapanja tako da konstruira sustav koji se oslanja na informacije iz *blob-a* kao i na informacije iz modela izgleda subjekta. U standardnu funkciju vjerojatnosti uvode nešto što ima kraticu SDAM, a znači eng. *spatial-depth affinity metric*. SDAM metrika se računa tako da se najprije dobave podaci o prostornoj poziciji piksela i dubinskom redoslijedu komponenta na slici koja sadrži taj piksel u trenutku preklapanja subjekta sa okolinom ili drugim subjektom. Redoslijed komponenti dobije se kombiniranjem *top-down* i *bottom-up* pristupa segmentiranja. Slika 31 prikazuje *workflow* sustava [72]. *Workflow* se može opisati sljedećim koracima: a) u RGB modelu boja konstruira se model pozadine korištenjem mješavine Gaussa, b) pikseli se klasificiraju korištenjem distorzije osvjetljenja ili eng. *brightness distortion* (BD) tako da se mjeri razlika između stvarne vrijednosti piksela i predviđene vrijednosti piksela pozadine, c) pragovi koji odlučuju u koju od 4 kategorije (*foreground,* sjena, veoma osvjetljene regije, pozadina) se odrađuju automatski u fazi inicijalizacije sustava te se ažuriraju ciklički tijekom unaprijed definiranog perioda kako bi se uvažilo promjene osvjetljenja na slici, d) na kraju jednog ciklusa obrade ažurira se pozadina, ali selektivno tako da se ne dogodi situacija da se ažurira i objekt koji je zastao ili se sporo kreće kroz sekvencu.



Slika 31 – preuzeto iz [72] - *workflow* sustava [72]

2.3.6 Trenutno otvoreni problemi u praćenju

Trenutno ne postoji niti jedan autonomni sistem koji radi neograničeno dugo u realnom vremenu, a koji koristi model segmentacije pozadine na dinamičkoj sceni koji bi sa velikom preciznošću i točnošću odjeljivao *foreground* objekte od *backgrounda* korištenjem samo jedne kamere normalne rezolucije.

Nadalje, za većinu sustava promjena u osvjetljenu zahtjeva dodatni posao obrade.

Zaključno, u teoriji postoji mogućnost istovremenog i segmentiranja i praćenja objekta korištenjem stohastičkog uzorkovanja, ali je u praksi teško naći radove koji uspješno implementiraju takav pristup.

2.4 ESTIMACIJA POZE – TREĆI KORAK

Estimacijom poze se nastoji otkriti na koji način je tijelo subjekta i njegovi udovi raspoređeno u prostoru scene. Odnosno, estimacije poze definira se kao metodologija procjene kinematičke ili skeletalne artikulacije subjekata ili individue. Estimacija može biti dio procesa praćenja, ali se može primijeniti i kao neovisan korak obrade nakon što je praćenje dalo svoj rezultat. Različite razine preciznosti se mogu zahtijevati od koraka estimacije. Npr. u sustavima nadzora ili sustavima interakcije sa računalom može se zahtijevati da se izračuna samo poza ruku i glave ili samo centar mase subjekta koji se giba. U nekim se pak sustavima kao što su medicinski zahtjeva točna pozicija, orijentacija i širina svakog uda subjekta.

Prema [2] i [4] radovi estimacije poze se mogu grupirati na sljedeća 3 polja ovisno o modelu subjekta koji sustav koristi:

I. Estimacija poze bez korištenja modela subjekta – ne postoji *apriori* model subjekta (oblika, kretnji ili nečeg trećeg) koji bi se koristio kao reprezentacija subjekta u fazi estimacije poze. Međutim, radovi ove grupe ipak konstruiraju određene modele poput točaka, jednostavnih geometrijskih oblika te pojednostavljene figure ljudi. Modeli točaka koriste se kod sustava koji rade sa markerima tako da je svaki marker povezan sa odgovarajućom točkom modela. Međutim, i kod sustava koji nemaju markera moguća je prezentacija točkom tako da svaki ud dobije svoju točku. Ova vrlo kompaktna prezentacija subjekta dovoljna je za preciznu estimaciju poze. Točke koje odgovaraju udovima bez korištenja markera pronađu se obično segmentacijom bojom ili segmentacijom *blob*-om. Što se tiče jednostavnih geometrijskih likova obično se koristi krivulje koje prate oblike ljudskih tijela poput elipse dok se pravokutnici koji ne prate precizno oblik tijela koriste samo u međukoracima obrade.

II. Estimacija poze sa indirektnim korištenja modela subjekta – model se koristi indirektno na način da služi samo kao referentna točka u procesu estimacije tako da se iz njega mogu izvlačiti informacije koje navode proces estimacije ka pravilnom zaključivanju o pozi subjekta.

III. Estimacija poze sa direktnim korištenja modela subjekta – ovi radovi koriste direktno 3D geometrijsku prezentaciju oblika i kinematičke strukture kako bi rekonstruirali pozu iz videa. Većina radova primjenjuje eng. *analysis-by-*

synthesis metodologiju kako bi optimizirala razliku između uočene poze i one opisane modelom.

2.4.1 Estimacija poze bez korištenja modela subjekta

Rad [73] predlaže algoritam koji efikasno pronalazi globalnu slikovnu strukture (eng. *pictorial structures*) scene. Kvalitativno opisivanje scene slikovnim strukturama se potom koristi dalje u sustavu [73] kako bi se takva reprezentacija sekvence koristila za prepoznavanje aktivnosti.

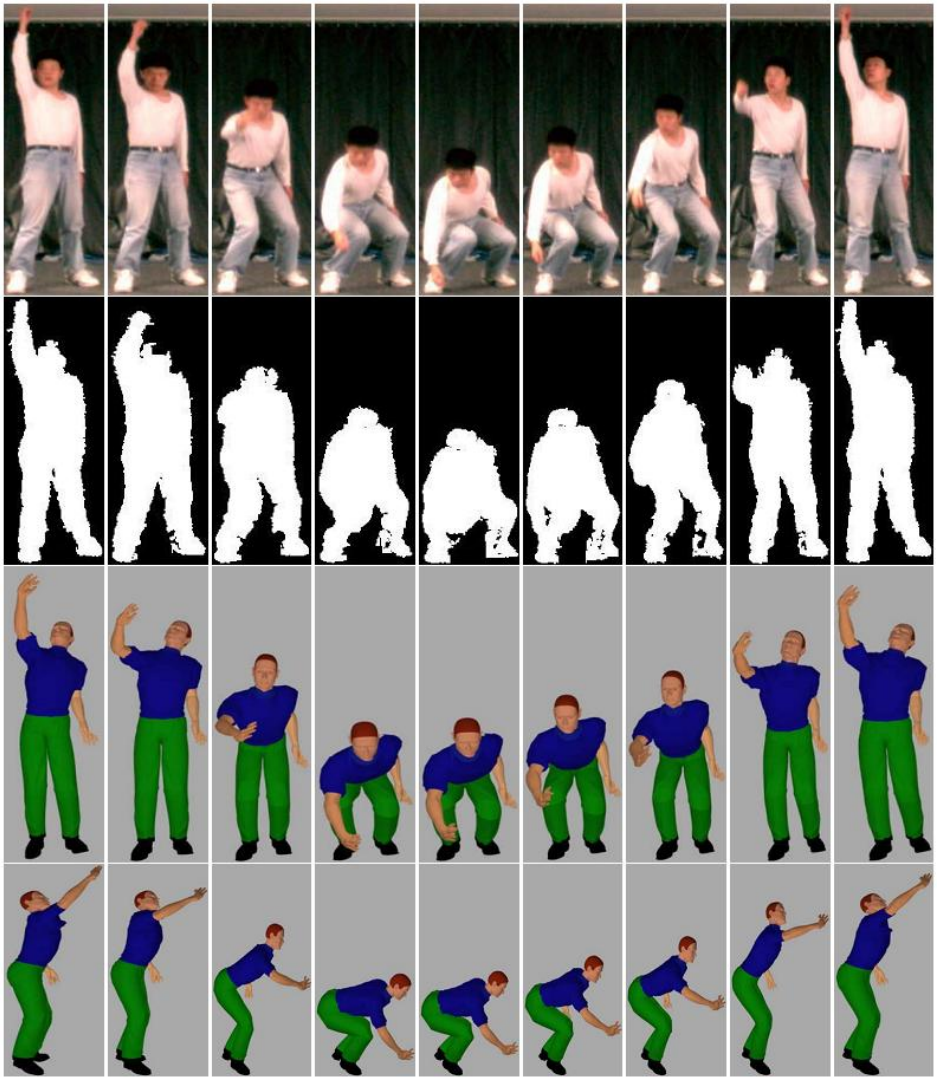
Rad [74] predlaže metodu učenja 3D poze iz jedne slike monokularnog videa. Metoda rekonstruira pozu na osnovu direktne, nelinearne regresije korištenjem deskriptora oblika koji se formiraju automatski iz siluete oblika subjekta. Da bi se eliminiralo lokalne greške segmentacije siluete rub subjekta se enkodira, prije samog postupka segmentacije, sa deskriptorima koji se sastoje od histograma konteksta oblika. Regresija se postiže uporabom standardnih metoda najmanjih kvadrata u kombinaciji sa RVM klasifikatorim (eng. *Relevance Vector Machine*) na linearnoj bazi te na bazi određenoj *kernel-om*.

Rad [75] predlaže jednostavni, ali efektivni algoritam praćenja artikuliranih 3D objekata (kao što je ljudsko tijelo) na način da uspoređuje otkrivenu pozu sa pohranjenom kolekcijom poza. Algoritam se automatski inicijalizira i oporavlja od slučajeva greške.

Za kompleksne i višedimenzionalne probleme, kao što je problem estimacije poze, broj potrebnih primjera za učenje te složenost izračuna rastu eksponencijalno. Zbog toga rad [76] predlaže algoritam koji uči skup *hash* funkcija koje indeksiraju primjere učenja na način da bitne poze jedne akcije indeksiraju tako da su one povezane. Algoritam je nadogradnja na koncept eng. *locality-sensitive hashing* te ga autori nazivaju eng. *Parameter-Sensitive Hashing*. Izbor *hash* funkcija je kritičan za uspješnost ovog algoritma te autori [76] također rješavaju uspješno i taj problem tako da optimalno pronalaze set funkcija za svaku vrstu poze. Slika 32 daje primjer rada algoritma.

Rad [77] opisuje algoritam propagacije koji procjenjuje 3D pozu subjekta iz monokularnog videa na osnovu kodiranog izgleda siluete subjekta. Rad se temelji na diskriminativnim metodama pa ne zahtjeva obzervacijski model predviđanja. Umjesto toga koristi bazu podataka koja pohranjuje veliku količinu snimljenih pokreta ljudi te 3D model čovjeka kako bi generirao uzorke za učenje u obliku 2D silueta. Na osnovu ovako konstruiranih primjera sustav nauči predviđati uvjetovane distribucije stanja koje su potrebne za praćenje 3D poze kroz sekvencu. Primjer generiranje sekvence sustavom [77] dan je na slici 33.

Slika 32 – preuzeto iz [76] – gornji redak su slike ulazne sekvence, srednji redak su najsličnije poze iz baze znanja prema eng. *Parameter-Sensitive Hashing* algoritmu te donji redak prikazuje slike poze generirane na osnovu ulaznih korištenjem eng. *locally-weighted* regresije kako bi se mogli napraviti usporedba.



Slika 33 – preuzeto iz [77] – prvi redak je originalni video, drugi redak je silueta, treći redak je primjer generiran za učenje, a četvrti redak je isto što i treći, ali iz drugog kuta gledanja

Rad [78] predlaže statistički izračun 2D poze subjekta iz monokularnog videa. Rad modelira konfiguraciju ljudskog tijela korištenjem Markovljevih mreža te je stoga problem estimacije poze sveden na pronalaženje parametara poze na osnovu neke značajke subjekta poput izgleda, oblika, boje ili ruba. Sustav [78] najprije uči moguće 2D poze subjekta korištenjem skupa slika sekvence na kojima je ručno označen svaki dio poze. Iz ovog skupa slika za učenje dobiju se parametri poza. Potom se primjenjuje Monte Carlo algoritam koji koristi funkciju uzorkovanja konstruiranu iz vizualnih značajki slike eng. *bottom-up* pristupom kako bi se zaključilo na koji način se rasprostire aposteriori distribucija.

2.4.2 Estimacija poze sa indirektnim korištenja modela subjekta

Svi relevantni radovi ovog polja koriste više kamera za dobivanje ulaznih podataka u sustave pa stoga ovo polje preskačemo jer ne zadovoljava uvjet da ulazni video mora biti monokularan rečen u uvodu ovog poglavlja.

2.4.3 Estimacija poze sa direktnim korištenja modela subjekta

Rad [79] predlaže novu vrstu estimacije poze koja je sposobna prepoznati pozu izvedenu sa nekoliko stupnjeva slobode kretanja. Navedeno se postiže produktom eksponencijalnih mapa i momenta zakreta subjekta što rezultira u procjeni diferencijalnog pokreta subjekta. Rad pretpostavlja da su promjene u intezitetu unutar jedne slike uzrokovane premještanjem lokalnih inteziteta unutar same slike. Parametarska jednadžba inteziteta koja opisuje intezitet dvije uzastopne slike dana je jednadžbom (12) :

$$I\big(x + \boldsymbol{u}_x(x, y, \phi), y + \boldsymbol{u}_y(x, y, \phi), t + 1\big) = I(x, y, t) \qquad (12)$$

gdje je I(x,y,t) intezitet slike. Model pokreta $u(x, y, \phi) = \big[\boldsymbol{u}_x(x, y, \phi), \boldsymbol{u}_y(x, y, \phi)\big]^T$ opisuje premještanje piksela s obzirom na lokaciju (x,y) i parametar modela pokreta ϕ. Npr. 2D afini model pokreta sa parametrom $\phi = [a_1, a_2, a_3, a_4, d_x, d_y]^T$ izgleda ovako:

$$\boldsymbol{u}(x, y, \phi) = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \end{bmatrix} \qquad (13)$$

Razvijanjem u prvi Taylorov red jednadžbe (13) dobije se često korištena jednadžba (14):

$$I_t(x, y) + \big[I_x(x, y), I_y(x, y)\big] * \boldsymbol{u}(x, y, \phi) = 0 \qquad (14)$$

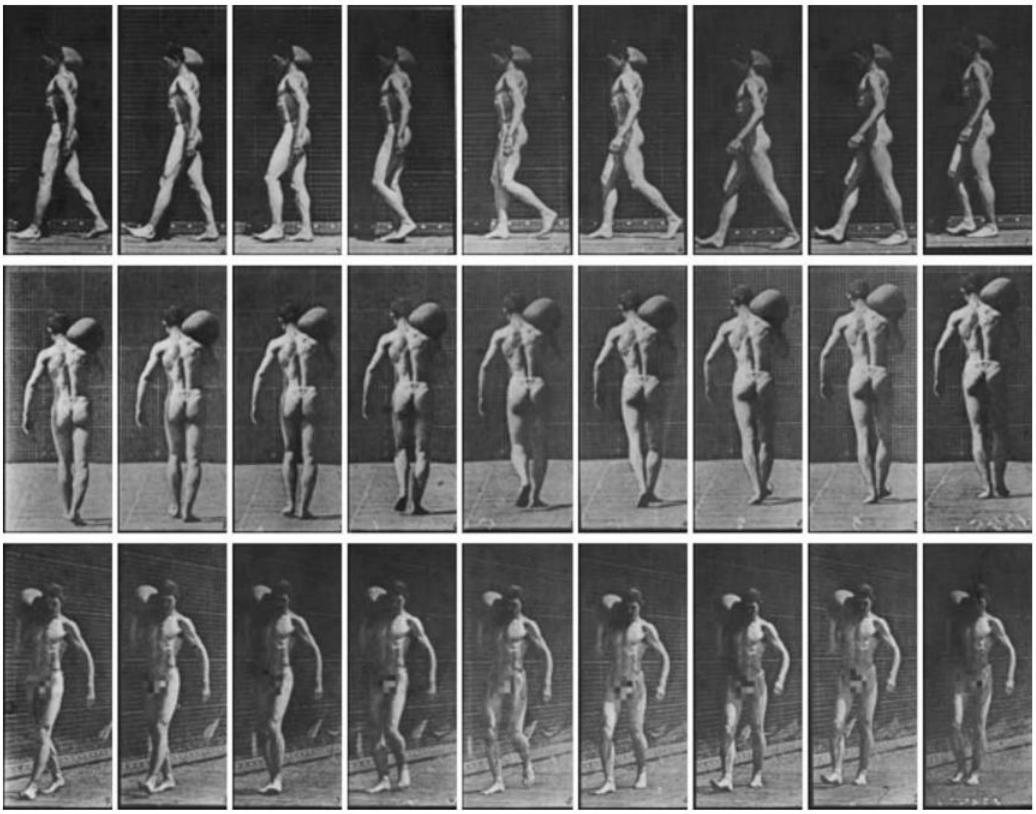
$I_t(x, y)$ je trenutni gradijent slike, a $I_x(x, y), I_y(x, y)$ je prostorni gradijent na lokaciji (x,y). Autori dalje pojednostavljuju jednadžbu (14) pretpostavkama da je afina transformacija linearna te dobiju jednadžbu (15):

$$\boldsymbol{H} * \phi + \vec{z} = \vec{0} \qquad (15)$$

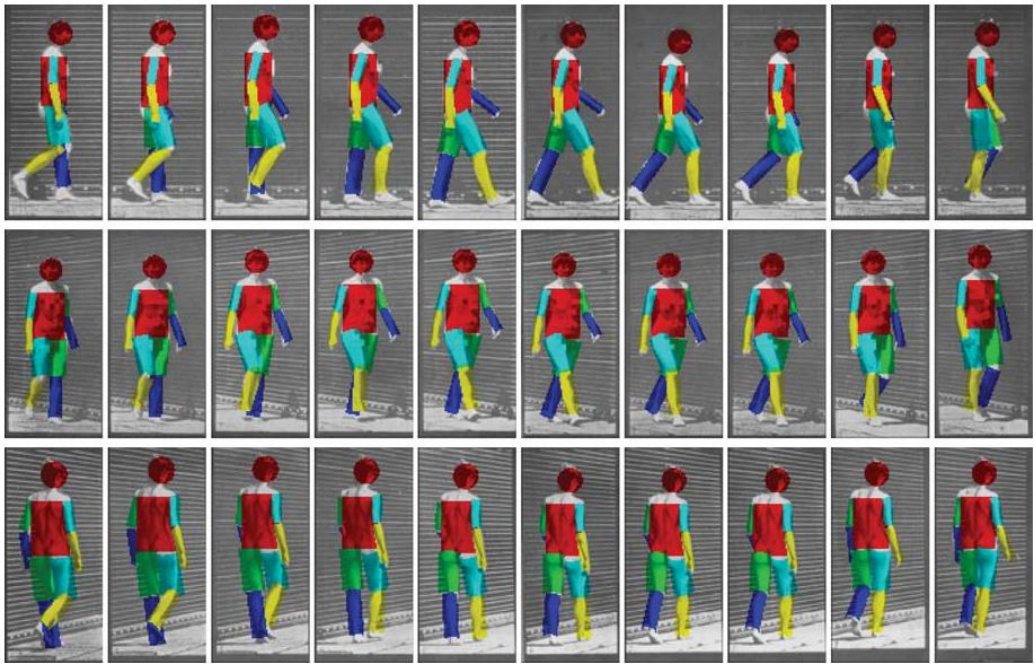
gdje je $\boldsymbol{H} \in \mathbb{R}^{NxK}$, a $\vec{z} \in \mathbb{R}^{N}$ pri čemu je N broj segmentiranih površina te K broj stupnjeva slobode kretanja. Rješenje metodom najmanjih kvadrata jednadžbe (14) jednako je:

$$\phi = -(\boldsymbol{H}^{T} * \boldsymbol{H})^{-1} * \boldsymbol{H}_{\vec{z}}^{T}$$

Slika 34 prikazuje sekvencu iz starog videa , slika 35 prikazuje praćenje subjekta radom [79]:



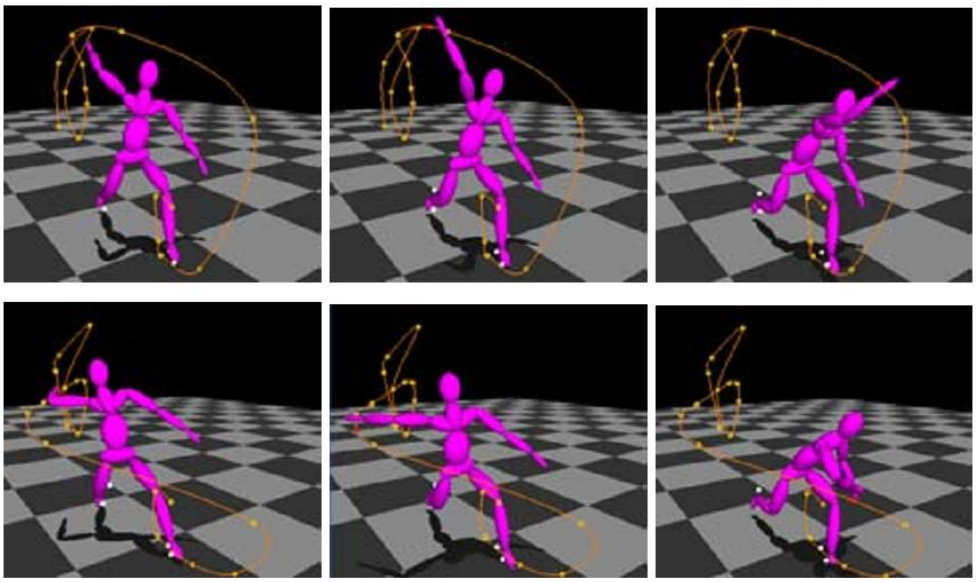
Slika 34 – preuzeto iz [79] – sekvenca hoda i nošenja predmeta golog čovjeka iz filma iz 1884. god.



Slika 35 – preuzeto iz [79] – ista sekvenca kao i na slici 34, ali segmentirana za praćenje i estimaciju.

Rad [80] i [81] definiraju skup pokreta uhvaćenih kamerom i prebačenih u animirani oblik tako da svaki autor koji radi nešto sa estimacijom pokreta može, uz novčanu naknadu, dobiti kolekciju svih animiranih pokreta u nekom od digitalnih formata (3D Studio Max, Maya, CharacterStudio, Lightwave, Life Forms, Poser itd.) kako bi ih dalje koristio kao dio vlastitog sustava, bilo da se radi o akademskim ili komercijalnim potrebama. Ovakvi setovi unaprijed obrađenih i dostupnih podataka o pozi, radnjama ili nečemu trećem vezanom za prepoznavanje doprinijeli su ubrzanju razvoja ove grane znanosti.

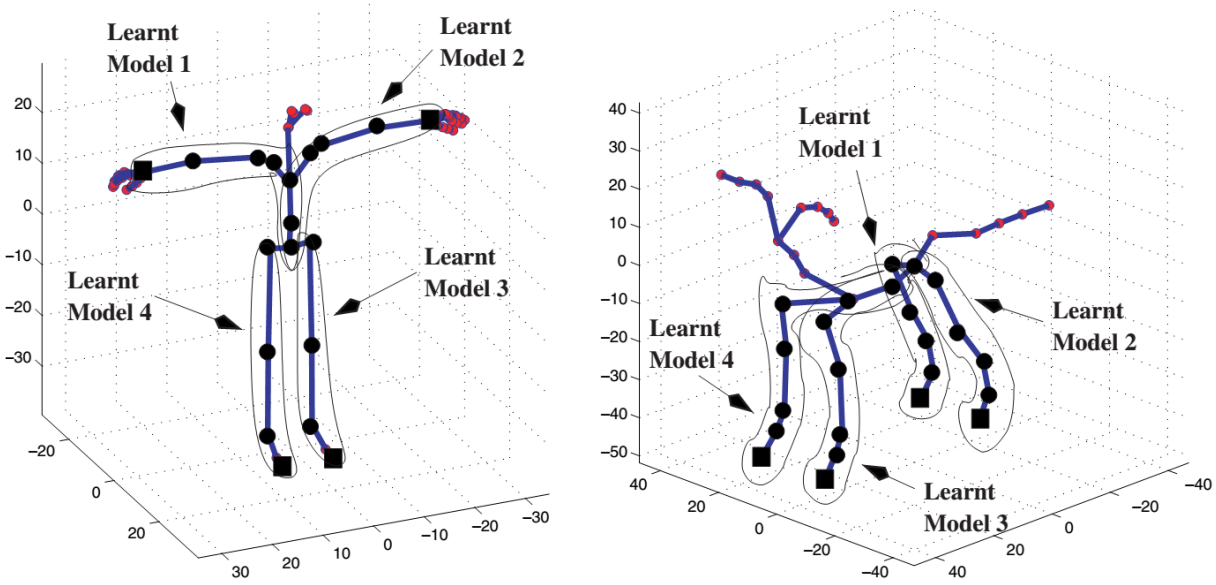
Rad [82] predstavlja inverzni kinematički sustav. Ideja je predstaviti funkciju poze kao PDF koji opisuje vjerojatnost tijekom izvođenja poze (da će se dogoditi određeno stanje poze). Uz pomoć poza za učenje mogu se utvrditi svi parametri PDF funkcije maksimiziranjem vjerojatnosti korištenih uzoraka za učenje kako bi se PDF funkcija mogla eksplicitno definirati. Prezentacija tako dobivene funkcije se radi vrstom prezentacije definiranom prvi puta u radu [83] imena eng. *Scaled Gaussian Process Latent Variable Model* (SGPLVM). Primjer kako izgleda prezentacija podataka akcije bacanje bejzbolske loptice dan je slikom 36:



Slika 36 – preuzeto iz [82] – gornji redak je vrsta bacanja u bejzbolu poznata kao eng. *baseball pitch,* a donji redak je vrsta bacanja poznata kao eng. *side-arm pitch.*

Rad [84] predstavlja novi način sinteze animacije radnji ljudi na osnovu minimalnog broja zglobova prema principu inverzne kinematike. Rad definira ograničenja kinematičkog modela tako da se model ne bi našao u fizički nemogućoj pozi. Ograničenja se definiraju korištenjem hijerarhijskog modela klastera treniranog da prepozna moguću pozu na bazi testnih uzoraka. Rad prepoznaje korelaciju među pozama zglobova na osnovu kuta kojeg oni

zatvaraju sa referentnom osi. Slikom 37 dan je primjer skeletnog modela kojeg rad [84] konstruira i koristi.



Slika 38 – preuzeto iz [84] – slika prikazuje model skeletona za čovjeka (lijevo) i psa (desno). Točke koje definiraju lokaciju ograničenja prikazane se kvadratom, a točke zglobova krugom. Modeli dijelova tijela se spajaju u jedan model u točkama zgloba.

2.4.4 Trenutno otvoreni problemi u estimaciji poze

Ne postoji propisani, zajednički ili jedinstveni set testnih uzoraka na kojima bi se mjerile performanse različitih sustava, a koji uvažavaju svi sustavi jednako.

2.5 PREPOZNAVANJE RADNJE – ČETVRTI KORAK

Korak prepoznavanja radnje može se gledati kao korak koji se pokreće kada je cijeli sustav praćenja i prepoznavanja poze završio sa obradom snimljenih podatka. On je cilj ili smisao cijelog sustava koji se gradi jer on daje konačnu odluku o vrsti radnje. Stoga se i pristup izgradnji sustava temelji na njemu, tj. na vrstama radnji koje je bitno prepoznati jer različite aplikacije imaju različite zahtjeve na vrstu i način klasifikacije radnji.

Ipak, postoje nekakva globalna pravila kojih se valja držati, a jedno kaže da prezentacija podatak u sustavu ne smije nikada zavisti o subjektu koji izvodi radnju.

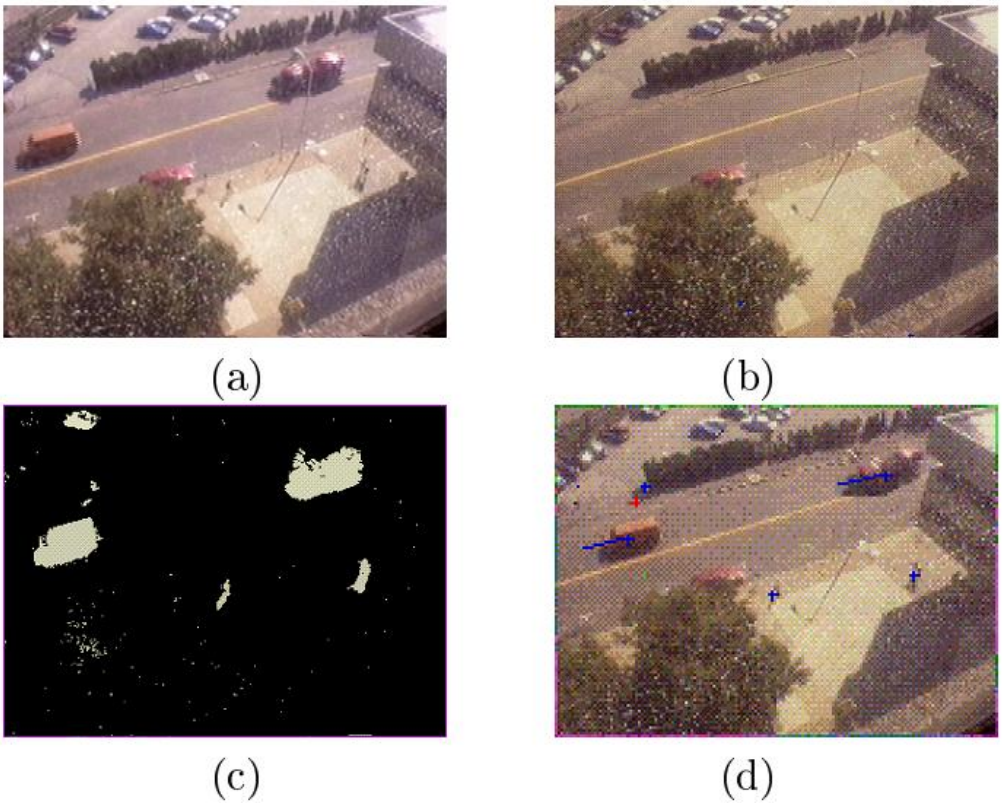
U ovom poglavlju radovi se grupiraju prema načinu na koji autori grade hijerarhijski sustav vizualnih abstrakcija. Tako imamo sustava koji interpretiraju cijelu scenu bez pravljenja razlike između subjekta i ostalih artefakata *foregrounda,* potom holističke ili cjelovite sustave koji koriste cijelo ljudsko tijelo ili samo njegove dijelove za prepoznavanje te akcijske primitive i gramatiku kod kojih predstavljanje akcije hijerarhijski po važnosti omogućava način semantičkog prikaza scene. U najnovije vrijeme neki autori primjenjuju i pristup temeljen na neuralnim mrežama. Slijedi kratak pregled radova po svakoj od navedenih kategorija.

2.5.1 Interpretacije scene

Neki radovi promatrajući općenito scenu nastoje naučiti i predvidjeti aktivnosti promatranjem pokreta subjekta bez da imaju neki podatak o kakav subjekt gledaju i analiziraju.

Rad [85] konstruira interpreter cijele scene koji uklanja iregularne pokrete (koji ne pripadaju objektu nego su posljedica šuma ili nečega sličnog). Sustav najprije izvlači značajke objekta iz scene (npr. 2D poziciju i brzinu subjekta, veličinu ili binarnu siluetu). Sustav je u stanju pratiti subjekt kroz sekvencu dugi period vremena bez ljudske intervencije, a za to ne koristi nikakve informacije o subjektu (poput modela kinematike itd.). U stanju je sam se oporavljati od grešaka praćenja te funkcionira i kod postepene izmjene osvjetljenja scene. Korišten je u različite svrhe poput praćenja subjekata na otvorenom, praćenje i prepoznavanje riba u pokretu unutar njihovih staništa, praćenja i prepoznavanje mravi na otvorenom i robota u zatvorenom prostoru. Slika 39 daje primjer kako izgleda praćenje sustavom [85].

Rad [43] koji je već spomenut u poglavlju 2.3.1.2 radi uspješno praćenje nakon prepoznavanja subjekta u vodenim uvjetima zatvorenih bazena sa velikom preciznošću bez korištenja modela subjekta ili nekog sličnog koncepta dodatnog izvora podataka o subjektu.



Slika 39 – preuzeto iz [85] – a) originalna slika iz sekvence, b) slika pozadine satkana od srednjih vrijednosti Gaussa, c) segmentacija *foregrounda,* d) slika a) sa uključenim informacijama o praćenju nakon što su subjekti prepoznati.

Rad [86] konstruira značajne točke subjekta u sceni tako da koriste prostorno – vremenske uzorke ili obrasce prethodno snimljenih slika sekvence te traži na trenutnoj slici sekvence iste te uzorke sa istom geometrijom, ali drugačijim rasporedom u prostoru. Pronađene točke predstavljaju lokalne deskriptore i omogućuju praćenje ponašanja subjekta.

Rad [87] predlaže metodu koja se fokusira na dinamičke podatke na trenutnoj sceni u svrhu interpretacije scene općenito. Sustav [87] je u stanju pratiti više subjekata koji izvode neovisne radnje odjednom u više različitih smjerova, a ukoliko je smjer radnje zajednički radnje se moraju razlikovati po prostorno-vremenskim atributima.

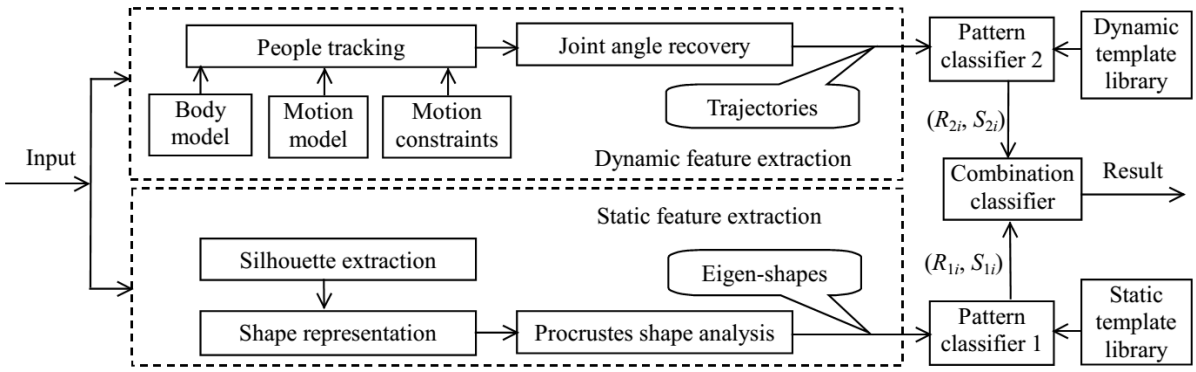
Rad [88] uvodi jedinstvenu metodu koja kombinira oba pristupa segmentiranja (bottom-up i top-down) subjekata koji izvodi kompleksnu radnju i preklapa se sa predmetima iz okoline ili sa ostalim subjektima. Sve ovo je postignuto bez poznavanja dodatnih izvora informacija o subjektima osim same scene.

2.5.2 Holistički pristupi

Većina sustava holističkog pristupa prepoznavanju koristi siluetu ili konturu subjekta.

U radu [89] računa se silueta čovjeka te poslije uzorkuje kontura objekta. Zatim se procesira udaljenost između svake točke konture i centra gravitacije tog dijela konture. Tako

dobiveni podaci puštaju se na PCA te je sustav u stanju postizati iznadprosječne rezultate sa malom količinom procesiranja podataka. Slika 40 prikazuje predloženi algoritam.



Slika 40 – preuzeto iz [89] – predloženi algoritam

Rad [90] kombinira skup metoda temeljenih na kovarijanci. Nakon normalizacije koja uzima u obzir skalu računa se mjera sličnosti koja funkcionira tako da stavlja u korelaciju slike siluete koje su posložene vremenski odmah jedna iza druge. Potom se koristi PCA da se umanji broj dimenzija podataka koji se obrađuju te eng. *k-Nearest Neighbor* algoritam u prostoru razapetom svojstvenim vektorima kako bi se napravilo prepoznavanje aktivnosti.

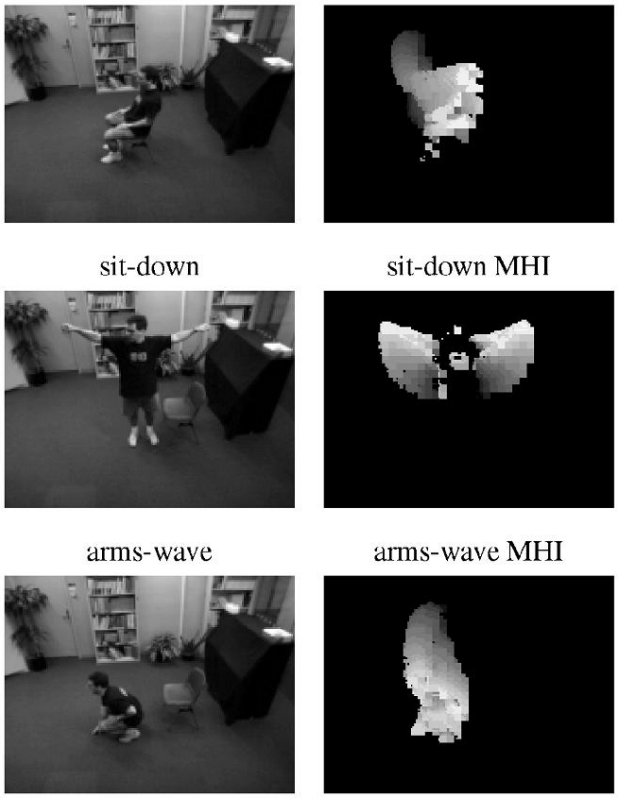
Rad [91] predlaže sustav koji ima mogućnost prepoznavanja aktivnosti samo jedne jedinke, ali u uvjetima vrlo niske rezolucija videa i subjekta koji je manji od 30 piksela po visini. Za postizanje takvih rezultata koriste se setom značajnih točki koji najbolje opisuju zamućeni optički tok. Individua se prati tako da se ju stabilizira usred prozora praćenja. Potom se računaju zamućeni optički tokovi na rezidualu pokreta deriviranom od pokreta svakog dijela tijela zasebno. Prostorno-vremenska eng. *cross-correlation* se koristi za povezivanje uzorka pokreta sa bazom.

U novije vrijeme postaje popularan pravac koji se temelji na prostorno-vremenskim značajkama (rad [92] i [93]) kod kojih se radnja povezuje sa trajektorijama pokreta.

Rad [94] je bio pionirski u korištenju vremenskih predložaka u svrhu prepoznavanja akcije. Autori [94] predlažu prezentaciju podataka i prepoznavanje radnji temeljeno na nečemu što su definirali kao eng. *Motion Energy Images* ili MEI. Poslije su isti autori napravili nadogradnju na inicijalni MHI pristup naziva tMHI tako da uključe još i podatak dobiven iz vremenske komponente u MHI sliku. Slika 41 prikazuje kako izgledaju određenw MHI konstrukcije zadanu radnju.

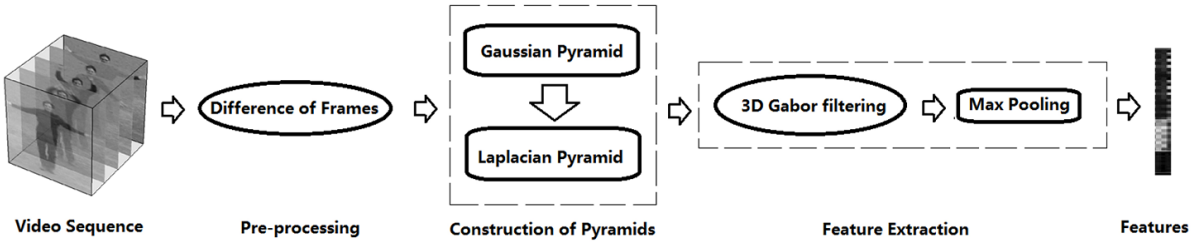
Rad [95] koristi eng. Spatio-Temporal Volumes ili STV za prepoznavanje aktivnosti. Najprije se iz 3D konture individue napravi 2D prijekcija na ravninu slike. Potome se bilježe

sve projekcije subjekta u nekom vremenu i upakiraju zajedno. Ovakvim pristupom dobije se rigidno 3D tijelo koje se onda uspoređuje sa subjektom.



Slika 41 – preuzeto iz [94] – radnje i primjer dobivenog MHI-a za svaku od njih

Rad [96] predlaže globalnog deskriptora za prepoznavanje cijelog tijela naziva eng. spatio-temporal Laplacian pyramid coding (STLPC). Ovaj deskriptor tretira vide okao cijelinu što pridonosi manjem gubitku informacije o pokretu. Potom se primjenjuje 3D Gabor filter kako bi se značajka napravila invarijantnom na šum i distorciju. Predloženi deskriptor testiran je na KTH, the multi-view IXMAS, UCF, HMDB51 setu video zapisa te se pokazao boljim nego bilo koja druga metoda holističkog pristupa do sada. Slika 42 daje prikaz procesa nastanka spomenute značajke.



Slika 42 – preuzeto iz [96] – proces dobivanja globalne značajke STLPC

Autori iz [96] predlaže u [97] 3 različita pristupa cjelovitom prepoznavanju subjekata,dakle daje više pristupa koji vode k istom rješenju. U prvom pristupu eksplicitno se izvlače značajke pokreta i strukture iz sekvence tako da se konvertira video prezentaciju u 2D

slikovnu prezentaciju problema. Drugi i treći pristup tretiraju video sekvencu kao 3D volumen pa onda na takvom volumenu primjene STLPC. Potom se primjenjuju Gabor filteri i filteri navođenja (eng. *steerable*) kako bi se postigao cjeloviti prikaz.

2.5.3 Akcijske primitivi i gramatika

Ova grupa radova uzima za hipotezu da je ljudska aktivnost direktno povezana sa motornom kontrolom u mozgu, tj. da se aktivnosti možemo modelirati tzv. akcijskim primitivima koji opisuju (svaki po jedan segment) akcije.

Radovi [98] i [99] daju znanstveno objašnjenje hipoteze o akcijskim primitivima i načinu kako su oni neuralno određeni te na koji način se akcije mogu modelirati akcijskim primitivima.

Rad [100] koristi HMM ili eng. Hidden Markov Models kako bi naučio posebne, nedvosmislene značajke pokreta. To radi tako da svakom zglobu dodijeli po jedan HMM. Potom se raznim distribucijama nastoji razbiti akciju na više akcijskih primitiva te se akcija opisana akcijskim primitivima može onda prezentirati alfabetom u kojem svaki primitiv ima svoje slovo.

Rad [101] konstruira dinamički *framework* za segmentiranje i klasifikaciju *foregrounda.* Definiraju se i tehnike identifikacije sustava kako bi se provocirala analitička greška analize i dobila procjena performansi sustava.

2.5.4 Neuralne mreže

Neuralne mreže kao dosta popularna tehnika eng. *Deep Learning-a* dobiva sve veći značaj u području prepoznavanja ljudskih aktivnosti.

Rad [102] koristi konvolucijsku neuralnu mrežu ili eng. Convolutional Neural Networks (CNNs) kojase može primijeniti direktno na video ulaz čime se automatizira korak izgradnjeznačajki. Do objave rada [102] ovakva uporaba CNN-a bila je mogućasamo za 2D ulazne podatke. Međutim, autori [102] su omogućili primjenu CNN-a i na 3D ulazne podatke. Njihov rad izvlači značajke i iz prostornih i vremenskih informacija o pokretu tako da radi 3D konvoluciju te time postiže efekt hvatanja informacije o pokretu iz susjednih slika sekvence. Sustav generira višestruke izvore informacija iz originalnog videa, a konačna prezentacija značajki dobije se tako da se kombinira sve izvore u jedan konačni kanal.

Rad [103], za razliku od većine drugih koji svoje učenje temelje na statičkim podacima, radi učenje na eng. *streaming* podacima (koji stalno dolaze i nisu ručno označeni). To im omogućava dinamički izbor najboljih mogućih značajki (svaki puta kada dođe nova

koja je bolja ona se uzima u razmatranje). Kombiniranjem neuralnih mreža i aktivnog učenja sustav inkrementalno poboljšava naučeni model. Sustav je prošao testiranje na četiri velike baze pohranjenih videa o ljudskim aktivnostima.

2.5.5 Trenutno otvoreni problemi u prepoznavanju aktivnosti

Unatoč velikom broju sustava prepoznavanja vrlo malo ih je u stanju nastaviti sa radom uslijed bitne promjene uvjeta scene.

2.6 ZAKLJUČAK

U pregledu je dan kratak sašetak većine relevantnih radova iz polja prepoznavanja ljudskih aktivnosti unutar zadnjik 15-ak godina.

Prepoznavanje ljudskih aktivnosti iznimno je multidisciplinarno područje u kojem se neprastano predlažu novi načini rješavanja dosta kompleksnih i teško rješivih problema poput praćenja uslijed otežanih uvjeta rada te praćenja i prepoznavanja složenih aktivnosti itd.

Nadalje, pregled sugerira da je gotovo nemoguće napraviti učinkovit i pouzdan sustav bez korištenja višeslojne obrade podataka osim u slučajevima ukoliko se cilja samo na specifičnu domenu sa čvrsto ispoštovanim postavljenim hipotezama.

Zaključno, ideja autora je napraviti sustav koji bi bio u stanju pratiti i prepoznavati ljudske pokrete na otvorenom prostoru (ulice grada, more, rijeke, jezera itd.), bez ikakve ljudske intervencije u rad sustav. Dakle, potpuno autonoman sustav sa jednom kamerom (monokularni video) bez dodatnih pomagala koji je u stanju razlučiti opasne od uobičajenih radnji. Sekundarni cilj je poboljšati postojeći algoritam koji radi praćenje i prepoznavanje u normalnim okolnostima tako da je on primjenjiv na domene gdje prije nije bio moguć (npr. na vodenu površinu).

3 LITERATURA

[1] Aggarwal, J. K.; Ryoo, M.S.: *Human activity analysis: A review* . Journal of the Association for Computing Machinery (ACM), Volume 43, Issue 3, Article Nr.16, April 2011

[2] Moeslund, T. B.; Hilton; A.;Krüger, *V.: A survey of advances in vision-based human motion capture and analysis.* Computer Vision and Image Understanding, 104(2), 90–126. February 2006

[3] Poppe, R. : *A survey on vision-based human action recognition*. Image and Vision Computing, 28(6), 976–990, June 2010

[4] Moeslund, T. B.; Granum, E: *A Survey of Computer Vision-Based Human Motion Capture.* Computer Vision and Image Understanding, 81, 231–268. 2001

[5] Bagheri, M. A; Gao, Q.; Guerrero, S. E.; Clapés, A.; Nasrollahi, K; Holte, M. B.; Moeslund, T. B.; *Keep it Accurate and Diverse Enhancing Action Recognition Performance by Ensemble Learning.* IEEE Conference on Computer Vision and Pattern Recognition Workshop, 7 Jun 2015

[6] Afsar, P.; Cortez, P.; Santos, H. : *Automatic visual detection of human behavior: A review from 2000 to 2014*. Expert Systems with Applications, An internationa Journal, 42 6935–6956, 2015

[7] Barrón, C.; Kakadiaris, I. A. : *Estimating anthropometry and pose from a single uncalibrated image.* Computer Vision and Image Understanding, 81(3), 269–284, 2001.

[8] Barron, C.; Kakadiaris, I. A. : *On the improvement of anthropometry and pose estimation from a single uncalibrated image*. Machine Vision and Applications, 14(4), 229–236, 2003.

[9] Parameswaran, V.; Chellappa, R. ; *View independent human body pose estimation from a single perspective image*. Proceedings of the 2004 IEEE computer society confeXuce on computer vision and pattern recognition. CVPR 2004 (Vol. 2, pp. II–16). IEEE.

[10] Taylor, C. J. : *Reconstruction of articulated objects from point correspondences in a single uncalibrated image*. Proceedings. IEEE conference on computer vision and pattern recognition, 2000 (Vol. 1, pp. 677–684). IEEE.

[11] Krahnstoever, N., Sharma, R. (2004). Articulated models from video. In Computer vision and pattern recognition, 2004. Proceedings of the 2004 IEEE computer society conference on CVPR 2004 (Vol. 1, pp. I–894). IEEE.

[12] Krahnstöver, N., Yeasin, M., Sharma, R. (2003). Automatic acquisition and initialization of articulated models. Machine Vision and Applications, 14(4), 218–228.

[13] Song, Y., Goncalves, L., Perona, P. (2003). Unsupervised learning of human motion models. Advances in Neural Information Processing Systems, 14

[14] Sidenbladh, H., Black, M. J. (2001). Learning image statistics for bayesian tracking. Proceedings. Eighth IEEE international conference on computer vision, 2001. ICCV 2001 (Vol. 2, pp. 709–716). IEEE.

[15] Sidenbladh, H., & Black, M. J. (2003). Learning the statistics of people in images and video. International Journal of Computer Vision, 54(1–3), 183–209.

[16] Ramanan, D., Forsyth, D. A., Zisserman, A. Strike a pose: Tracking people by finding stylized poses. IEEE Computer society conference on computer vision and pattern recognition, 2005. CVPR 2005 (Vol. 1, pp. 271–278). IEEE.

[17] Micilotta, A. S., Ong, E.-J., & Bowden, R. (2005). Detection and tracking of humans by probabilistic body part assembly. In BMVC

[18] Ramanan, D., Forsyth, D. A., Zisserman, A. " Tracking People by Learning their Appearance" IEEE Pattern Analysis and Machine Intelligence (PAMI). Jan 2007

[19] Roberts, T. J., McKenna, S. J., Ricketts, I. W. Human pose estimation using learnt probabilistic region similarities and partial configurations. In Computer vision-ECCV 2004 (pp. 291–303). Springer.

[20] Ronfard, R., Schmid, C., Triggs, B. Learning to parse pictures of people. In Computer vision-ECCV 2002 (pp. 700–714). Springer.

[21] Viola, P., Jones, M. Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, 2001. CVPR 2001 (Vol. 1, pp. I–511). IEEE.

[22] C. Stauffer and W.E.L. Grimson. Adaptive Background Mixture Models for RealTime Tracking. In Computer Vision and Pattern Recognition, Santa Barbara, CA, USA, June 1998.

[23] McKenna, S. J., Jabri, S., Duric, Z., Wechsler, H. Tracking interacting people. In Fourth IEEE international conference on automatic face and gesture recognition, 2000. Proceedings (pp. 348–353). IEEE.

[24] Cucchiara, R., Grana, C., Piccardi, M., Prati, A. Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(10), 1337–1342.

[25] Zhao, T., & Nevatia, R. (2004). Tracking multiple humans in crowded environment. Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004 (Vol. 2, pp. II–406). IEEE.

[26] Kim, K., Chalidabhongse, T. H., Harwood, D., & Davis, L. Real-time foreground–background segmentation using codebook model. Real-Time Imaging, 11 (3), 172–185.

[27] Fihl, P., Corlin, R., Park, S., Moeslund, T. B., Trivedi, M. M. Tracking of individuals in very long video sequences. In Advances in visual computing (pp. 60–69). Springer.

[28] Heikkila, M., & Pietikainen, M. A texture-based method for modeling the background and detecting moving objects. IEEE transactions on pattern analysis and machine intelligence, 28(4), 657–662.

[29] Oliver, N. M., Rosario, B., & Pentland, A. P. A bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 831–843.

[30] Monnet, A., Mittal, A., Paragios, N., & Ramesh, V. (2003). Background modeling and subtraction of dynamic scenes. In Proceedings. Ninth IEEE international conference on computer vision, 2003 (pp. 1305–1312). IEEE.

[31] Prati, A., Mikic, I., Trivedi, M. M., & Cucchiara, R. Detecting moving shadows: Algorithms and evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(7), 918–923.

[32] Cucchiara, R., Grana, C., Piccardi, M., & Prati, A. Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(10), 1337–1342, 2003

[33] On hue and saturation of natural and non-natural materials Gaši, Ana; Dujmić, Hrvoje; Papić, Vladan; Turić, Hrvoje.Proceedings of 2008 International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2008). Split, 2008.

[34] Elgammal, A., Harwood, D., & Davis, L. (2000). Non-parametric model for background subtraction. In Computer vision-ECCV 2000 (pp. 751–767). Springer.

[35] AHMED ELGAMMAL, RAMANI DURAISWAMI, DAVID HARWOOD, LARRY S. DAVIS: Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance, PROCEEDINGS OF THE IEEE · AUGUST 2002

[36] Guha, P., Mukerjee, A., & Venkatesh, K. (2005). Efficient occlusion handling for multiple agent tracking by reasoning with surveillance event primitives. In 2nd Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, 2005 (pp. 49–56). IEEE

[37] McKenna, S. J., Jabri, S., Duric, Z., & Wechsler, H. (2000). Tracking interacting people. In Fourth IEEE international conference on automatic face and gesture recognition, 2000. Proceedings (pp. 348–353). IEEE.

[38] Yang, T., Pan, Q., Li, J., & Li, S. (2005b). Real-time multiple objects tracking with occlusion handling in dynamic scenes. IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005 (Vol. 1, pp. 970–975).IEEE.

[39] Zhao, T., & Nevatia, R. (2004). Tracking multiple humans in crowded environment. Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004 (Vol. 2, pp. II–406). IEEE.

[40] Schindler, K., & Wang, H. (2006). Smooth foreground-background segmentation for video processing. In Computer vision-ACCV 2006 (pp. 581–590). Springer.

[41] Sheikh, Y., & Shah, M. (2005). Bayesian modeling of dynamic scenes for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11), 1778–1792

[42] Chen, M., Ma, G., & Kee, S. (2005). Pixels classification for moving object extraction. In Application of computer vision 2005. Seventh IEEE workshops on WACV/MOTIONS'05 Volume 1 (Vol. 2, pp. 44–49). IEEE.

[43] Eng, H., Toh, K.-A., Kam, A. H., Wang, J., & Yau, W.-Y. (2003). An automatic drowning detection surveillance system for challenging outdoor pool environments. In Ninth IEEE international conference on computer vision, 2003. Proceedings (pp. 532–539). IEEE.

[44] Haritaoglu, I., Harwood, D., & Davis, L. S. (2000). W 4: Real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 809–830.

[45] Gutchess, D., Trajkovics, M., Cohen-Solal, E., Lyons, D., & Jain, A. K. (2001). A background model initialization algorithm for video surveillance. In Computer vision, 2001 . Eighth IEEE international conference on ICCV 2001. Proceedings (Vol. 1, pp. 733–740). IEEE.

[46] Wang, H., & Suter, D. (2005). Background initialization with a new robust statistical approach. In 2nd Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, 2005 (pp. 153–159). IEEE.

[47] Feng Wang, Shuguang Dai (2009): Adaptive background update based on mixture models of Gaussian, Proceedings of the 2009 IEEE International Conference on Information and Automation June 22 -25, 2009, Zhuhai/Macau, China

[48] Sidenbladh, H. (2004). Detecting human motion with support vector machines. Proceedings of the 17th international conference on pattern recognition. ICPR 2004 (Vol. 2, pp. 188–191). IEEE

[49] Gonzalez, J. J., Lim, I. S., Fua, P., & Thalmann, D. (2003). Robust tracking and segmentation of human motion in an image sequence. In International conference on acoustics, speech, and signal processing, Hong Kong (Vol. 1, pp.29–32).

[50] Sangi, P., Heikkila, J., & Silvén, O. (2001). Extracting motion components from image sequences using particle filters. In Proceedings of the scandinavian conference on image analysis (pp. 508–514).

[51] Bradski, G. R., & Davis, J. W. (2002). Motion segmentation and pose recognition with motion history gradients. Machine Vision and Applications, 13(3), 174–184.

[52] Hu M (1962) Visual pattern recognition by moment invariants. IRE Trans Inf Theory IT-8(2)

[53] Starner T, Pentland A (1995) Visual recognition of American sign language using hidden Markov models. In: Proceedings of International Workshop on Automatic Face and Gesture Recognition

[54] Viola, P., Jones, M. J., & Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. In Ninth IEEE international conference on computer vision, 2003. Proceedings (pp. 734–741). IEEE.

[55] Mohan, A., Papageorgiou, C., & Poggio, T. (2001). Example-based object detection in images by components. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(4), 349–361 .

[56] Okuma, K., Taleghani, A., De Freitas, N., Little, J. J., & Lowe, D. G. (2004). A boosted particle filter: Multitarget detection and tracking. In Computer vision-ECCV 2004 (pp. 28–39). Springer.

[57] Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(5), 564–577.

[58] Yang, C., Duraiswami, R., & Davis, L. (2005a). Fast multiple object tracking via a hierarchical particle filter. Tenth IEEE international conference on computer vision, 2005. ICCV 2005 (Vol. 1, pp. 212–219). IEEE

[59] Capellades, M. B., Doermann, D., DeMenthon, D., & Chellappa, R. (2003). An appearance based approach for human and object tracking. In Image processing, 2003. 2003 International conference on ICIP 2003. Proceedings (Vol. 2, pp. II–85). IEEE

[60] Mittal, A., & Davis, L. S. (2003). M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. International Journal of Computer Vision, 51 (3), 189–203.

[61] Andrew Senior, Arun Hampapur, Ying-Li Tian, Lisa Brown, Sharath Pankanti, Ruud Bolle (2005) Appearance models for occlusion handling. Elsevir Image and Vision Computing 24 (2006) 1233–1243

[62] Khan, S., & Shah, M. (2000). Tracking people in presence of occlusion. Asian conference on computer vision (Vol. 5). Citeseer.

[63] Park, S., & Aggarwal, J. K. (2006). Simultaneous tracking of multiple body parts of interacting persons. Computer Vision and Image Understanding, 102(1), 1–21 .

[64] Leibe, B., Seemann, E., & Schiele, B. (2005). Pedestrian detection in crowded scenes. In Computer vision and pattern recognition, 2005. IEEE computer society conference on CVPR 2005 (Vol. 1, pp. 878–885). IEEE.

[65] Wu, Y., & Yu, T. (2006). A field model for human detection and tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5), 753–765.

[66] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In Computer vision and pattern recognition 2005. IEEE computer society conference on CVPR 2005 (Vol. 1, pp. 886–893). IEEE.

[67] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. The 8th ICCV, Vancouver, Canada, pages 454–461, 2001.

[68] Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In Computer vision-ECCV 2006 (pp. 428–441). Springer.

[69] Wu, B., & Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. Tenth IEEE international conference on computer vision, 2005. ICCV 2005 (Vol. 1, pp. 90–97). IEEE.

[70] Antonini, G., Martinez, S. V., Bierlaire, M., & Thiran, J. P. (2006). Behavioral priors for detection and tracking of pedestrians in video sequences. International Journal of Computer Vision, 69(2), 159–180.

[71] Roth, D., Doubek, P., & Van Gool, L. J. (2005). Bayesian pixel classification for human tracking. In WACV/MOTION (pp. 78–83)

[72] Xu, L.-Q., & Puig, P. (2005). A hybrid blob-and appearance-based framework for multi-object tracking through complex occlusions. In 2nd Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, 2005 (pp. 73–80). IEEE

[73] Felzenszwalb, P. F., & Huttenlocher, D. P. (2000). Efficient matching of pictorial structures. IEEE conference on computer vision and pattern recognition, 2000. Proceedings (Vol. 2, pp. 66–73). IEEE.

[74] Agarwal, A., & Triggs, B. (2004). 3d human pose from silhouettes by relevance vector regression. In Computer vision and pattern recognition 2004. Proceedings of the 2004 IEEE computer society conference on CVPR 2004 (Vol. 2, pp. II–882). IEEE.

[75] Howe, N. R. (2004). Silhouette lookup for automatic pose tracking. In CVPRW'04. Conference on computer vision and pattern recognition workshop, 2004 (pp. 15–22). IEEE.

[76] Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. Proceedings. Ninth IEEE international conference on computer vision, 2003 (pp. 750–757). IEEE.

[77] Sminchisescu, C., Kanaujia, A., Li, Z., & Metaxas, D. (2005). Discriminative density propagation for 3d human motion estimation. IEEE Computer society conference on computer vision and pattern recognition, 2005. CVPR 2005 (Vol. 1, pp. 390–397). IEEE.

[78] Hua, G., Yang, M.-H., & Wu, Y. (2005). Learning to estimate human pose with dana driven belief propagation. In Computer vision and pattern recognition, 2005. IEEE computer society conference on CVPR 2005 (Vol. 2, pp. 747–754). IEEE

[79] Bregler, C., Malik, J., & Pullen, K. (2004). Twist based acquisition and tracking of animal and human kinematics. International Journal of Computer Vision, 56(3), 179–194.

[80] Interactive, C., (2005). Megamocap: Specialist in motion capture and 3D character animation. <http://www.charactermotion.com/about/index.html>.

[81] Jang, B. (2002). CMU graphics lab motion capture database. <http://mocap.cs.cmu.edu/faqs.php>.

[82] Grochow, K., Martin, S. L., Hertzmann, A., & Popović, Z. (2004). Style-based inverse kinematics. ACM transactions on graphics (TOG) (Vol. 23, pp. 522–531).ACM

[83] LAWRENCE, N. D. 2004. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. Proc. NIPS 16.

[84] Ong, E.-J., & Hilton, A. (2006). Learnt inverse kinematics for animation synthesis. Graphical Models, 68(5), 472–483.

[85] Stauffer, C., & Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 747–757.

[86] Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. International Journal of Computer Vision, 74(1), 17–31 .

[87] Junejo, I. N., Javed, O., & Shah, M. (2004). Multi feature path modeling for video surveillance. Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004 (Vol. 2, pp. 716–719). IEEE

[88] Xiang, T., & Gong, S. (2006). Beyond tracking: Modelling activity and understanding behaviour. International Journal of Computer Vision, 67(1), 21–51 .

[89] Wang, L., Ning, H., Tan, T., & Hu, W. (2004). Fusion of static and dynamic body biometrics for gait recognition. IEEE Transactions on Circuits and Systems for Video Technology, 14(2), 149–158.

[90] BenAbdelkader, C., Cutler, R., & Davis, L. (2002). Motion-based recognition of people in eigengait space. In Fifth IEEE international conference on automatic face and gesture recognition, 2002. Proceedings (pp. 267–272). IEEE.

[91] Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In Ninth IEEE international conference on computer vision, 2003. Proceedings (pp. 726–733). IEEE.

[92] Ricquebourg, Y., & Bouthemy, P. (2000). Real-time tracking of moving persons by exploiting spatio-temporal image slices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 797–808.

[93] Rittscher, J., Blake, A., & Roberts, S. J. (2002). Towards the automatic analysis of complex human body motions. Image and Vision Computing, 20(12), 905–916

[94] Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3), 257–267.

[95] Yilmaz, A., & Shah, M. (2005). Actions sketch: A novel action representation. IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005 (Vol. 1, pp. 984–989). IEEE.

[96] Shao, L., Zhen, X., Tao, D., & Li, X. (2014). Spatio-temporal laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics, 44(6), 817–827.

[97] Xiantong Zhen (2013); Feature Extraction and Representation for Human Action Recognition, Ph.D thesis partial fulfillment

[98] Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience, 4(3), 179–192.

[99] Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience, 2(9), 661–670.

[100] Billard, A., Epars, Y., Calinon, S., Schaal, S., & Cheng, G. (2004). Discovering optimal imitation strategies. Robotics and Autonomous Systems, 47(2), 69–77.

[101] Del Vecchio, D., Murray, R. M., & Perona, P. (2003). Decomposition of human motion into dynamics-based primitives with application to drawing tasks. Automatica, 39(12), 2085–2098.

[102] Xu, W., Yang, M, Yu, K: 3D Convolutional Neural Networks for Human Action Recognition. Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010.

[103] Mahmudul, H., Roy-Chowdhury, A.K.: Continuous Learning of Human Activity Models using Deep Nets. Computer Vision–ECCV 2014, 2014 - Springer