Academia Română

Institutul de Matematică Simion Stoilow

Modele Perceptuale pentru Analiza Vizuală Computațională

Autor,
Elisabeta Marinoiu

Conducător,
C.S. I Dr. Cristian
Sminchișescu

Rezumat Teză de Doctorat

București
2018

Capitolul 1

Introducere

Când privim o imagine sau când urmărim un video, înțelegem imediat spațiul 3d implicit, putem identifica și descrie care sunt principalii actori precum și principalele acțiuni și obiecte. Analiza proceselor implicate în percepția vizuală umană poate oferi intuiții în direcția construirii unor sisteme vizuale automate mai robuste și mai adaptabile.

În această teză ne propunem să construim sisteme vizuale automate care integrează, atunci când acest lucru este posibil, elemente de percepție vizuală umană. Mai întâi adresăm o problemă mai generală, descrierea automată a unui video în limbaj natural. Apoi ne concentrăm pe analiza perceptuală și reconstruirea automată a unei categorii vizuale mai particulară, dar foarte dificilă – oamenii, precum și a posturii lor în spațiul 2d și 3d. În această teză tratăm doar cazul monocular atât pentru imagini cât și pentru videouri. Rezolvarea acestor probleme are implicații în domenii foarte diverse, cum ar fi: indexarea automată a videourilor, modelarea comportamentală, terapie asistată sau mașini autonome. Propunem un model automat de descriere al videourilor construit pe baza unui mecanism de atenție spațio-temporal și a unei rețele neuronale recurente bazată pe memorie de scurtă durată, lungă (en., *long short-term memory*; LSTM). Modelul integrează și informație provenită din propuneri spațio-temporale în video precum și răspunsuri ale clasificatorilor pentru diverse categorii semantice. Problema este dificilă din cauza variabilității propozițiilor adnotate și a categoriilor semantice întâlnite. Metoda propusă produce rezultate de top,

și poate localiza conceptele semantice (subiecte, verbe, obiecte) în spațiu și timp fără a avea nevoie de adnotări adiționale.

În continuare, prezentăm un aparat experimental care are drept scop să creeze o legătură între percepția umană și obținerea unor măsurători cantitative. Pentru acest scop, am cerut mai multor oameni să reproducă posturi 3d pe care le arătăm pentru scurt timp într-o serie de imagini cu o singură persoană. În acest timp, le înregistrăm mișcările corpului și fixația ochilor folosind aparatură specializată. Oferim o analiză extensivă a consistenței mișcărilor ochilor între participanți precum și rezultate cantitative și calitative privind reproducerile posturilor. Datele colectate și intuițiile rezultate în urma analizei acestora sunt folosite mai departe pentru a învăța metrici perceptuale care, atunci când sunt integrate în modele automate de estimare a posturii 3d într-o imagine, produc rezultate mai stabile și mai plauzibile.

În cele din urmă, propunem o extensie a metodelor de estimare a posturii unei singure persoane, către o metodologie care permite estimarea posturii în spațiul 2d, 3d precum și a formei corpului și a distanței față de cameră pentru mai mulți oameni din imagini naturale. Principalele dificultăți provin din varietatea de forme pe care o poate lua corpul uman, din ocluzii și vederi doar parțiale ale acestuia, din interacțiuni foarte apropiate între oameni precum și din ambiguitatea rezultată în urma proiecției în spațiul 2d al imaginii. Ne folosim de o rețea profundă de înțelegere vizuală cu sarcini multiple peste care adăugăm un pas de optimizare ghidat de elemente vizuale semantice. Impunem o serie de constrângeri care țin de scenă, cum ar fi, faptul că oamenii stau pe un același plan sau că nu se pot întrepătrunde. Acestea ne ajută să obținem rezultate de top precum și reconstrucții calitative promițătoare în imagini naturale.

Capitolul 2

Modele de atenție spațio-temporală pentru descrierea videourilor

Descrierea automată a videourilor este dificilă din cauza interacțiunilor complexe din scenele reale. Un sistem inteligent ar trebui să poată localiza și urmări obiectele, acțiunile și interacțiunile prezente în video și să genereze o descriere pe baza acestora. Însă cele mai multe dintre metodele existente de descriere automată a videourilor aleg să construiască o propoziție folosindu-se direct de datele video evitând astfel partea de localizare și recunoaștere. Acest tip de procesare omite informație foarte utilă pentru capacitatea de generalizare și localizare a modelului. În acest capitol prezentăm un model automat de descriere a videourilor care combină un mecanism de atenție spațio-temporală și clasificatori de imagine prin intermediul unei rețele neuronale recurente (LSTM). Sistemul propus obține rezultate de top pe baza de date YouTube [4] având și avantajul de a oferi localizarea spațio-temporală a unor concepte semantice (subiecte, verbe, obiecte) fără a folosi adnotări specifice problemei de localizare.

2.1 Metodologie

Abordarea noastră pentru problema de descriere automată a videourilor are două componente: prima constă în găsirea suportului spațio-temporal al cuvintelor în cadrul videoului, iar a doua constă în ghidarea procesului de generare de propoziții prin

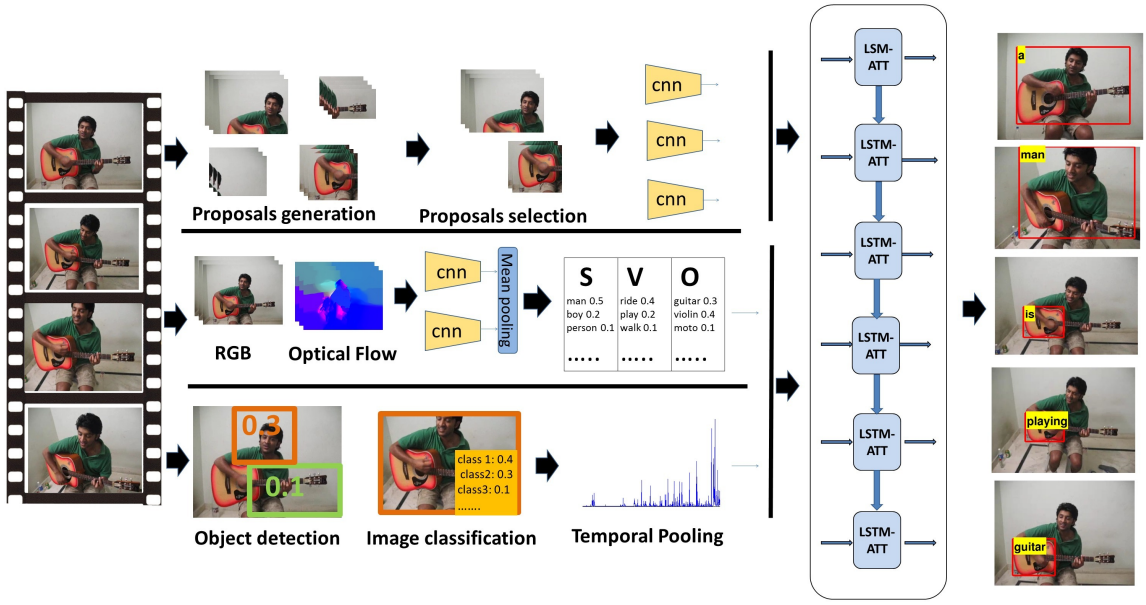


Figura 2-1: Vedere de ansamblu a metodologiei propuse pentru descrierea automata a videourilor cu o propozitie și localizarea spațio-temporală a conceptelor semantice.

includerea de informație semantică în procesul de învățare. Mecanismul de atenție funcționează peste o mulțime de propuneri spațio-temporale în video și este integrat cu o rețea neuronală recurentă. Informația semantică este obținută în două feluri: (a) învățând să estimăm care sunt subiectele, verbele și obiectele și (b) folosind modele preantrenate pentru clasificarea de imagini și detecția de obiecte. O vedere de ansamblu a modelării și a pașilor computaționali este prezentată în figura 2-1.

2.1.1 Propuneri spațio-temporale de obiecte

Folosim metodologia din [14] pentru a genera un set inițial (aproximativ 1000) de propuneri spațio-temporale de obiecte pentru fiecare video. Acestea sunt apoi filtrate și ordonate folosind o serie de euristici care iau în seamă dimensiunea și probabilitatea de a conține un obiect. Un video va fi reprezentat prin primele 20 de astfel de propuneri spațio-temporale. Pentru fiecare propunere, extragem în fiecare cadru caracteristici din stratul fc_7 al rețelei VGG-19. O propunere va fi reprezentată prin media acestor caracteristici de-a lungul cadrelor care intră în componența sa.

2.1.2 LSTM bazat pe atenție

Încorporăm un mecanism de atenție *soft* în LSTM pentru a permite modelului să se concentreze selectiv asupra diferitelor parți de video reprezentate prin propunerile spațio-temporale de obiecte, la fiecare moment de timp când este emis un cuvânt. Modelul învață să atribuie fiecărei propuneri o pondere, atunci când emite un cuvânt. Valoarea acesteia este influențată de importanța pe care a avut-o acea zonă spațio-temporală în producerea cuvântului respectiv. Astfel, modelul propus poate indica localizarea în video a cuvintelor folosite în descrierea videoului.

2.1.3 Informație semantică

Pentru a îmbunătăți modelul, propunem să folosim și informație semantică, sub mai multe forme: răspunsuri ale clasificatoarelor de imagine și detectoarelor de obiecte preantrenate precum și răspunsuri ale clasificatoarelor învățate de noi pentru de subiecte, verbe și obiecte. Generarea unei descrieri textuale a unui video presupune identificarea actorilor și a interacțiunilor acestora și apoi construirea unei propoziții corecte gramatical. Pentru a putea produce în mod automat descrieri cât mai aproape de cele propuse de oameni, mai întâi ne propunem să obținem pentru un video o reprezentare de forma Subiect (S), Verb (V) și Obiect (O) (în mod similar cu lucrări anterioare [7]). Apoi integrăm această reprezentare cu rețeaua neuronală recurentă bazată pe mecanismul de atenție. Pentru a învăța această reprezentare de nivel înalt a fiecărui video, mai întâi procesăm adnotările textuale din datele de antrenare și din fiecare propoziție extragem tuplul (S, V, O) – de exemplu, propoziția *A cat plays with a toy* este reprezentată ca (cat, play, toy). Din acestea construim trei vocabulare separate pentru subiecte, verbe și obiecte și folosim modelul Least Squares Support Vector Machine (LS-SVM) pentru a învăța pentru fiecare video o reprezentare în acest spațiu.

2.2 Detalii experimentale

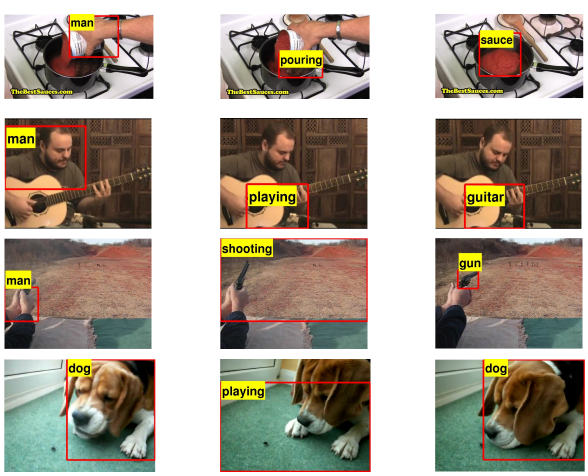
Descrierea bazei de date. Prezentăm experimente pe baza de date YouTube [4] care conține 1,967 videouri scurte (având durata între 10 și 25 secunde). Acestea care au fost obținute de pe YouTube și de obicei conțin o singură acțiune principală. Fiecare video are asociate aproximativ 40 de descrieri în limba engleză, colectate prin intermediul aplicației Amazon Mechanical Turk.

Metrici de evaluare. Pentru raportarea rezultatelor folosim două metrici, BLEU [17] și METEOR [11], care au fost inițial propuse pentru evaluarea traducerii automate dintr-o limbă în alta și care au fost adoptate de lucrările anterioare în evaluarea descrierilor automate ale videourilor și imaginilor.

2.2.1 Rezultate experimentale

Rezultate cantitative. Rezultatele obținute cu modelele propuse sunt arătate în tabelul 2.1. Modelul nostru care combină un mecanism de atenție cu o rețea LSTM (LSTM-ATT) obține rezultate comparabile cu ale celorlalte metode din literatură. Dacă adaugăm caracteristici semantice peste acest model, obținem cele mai bune rezultate conform metricii BLEU și rezultate promițătoare sub metrica METEOR. Impactul pe care îl au caracteristicile SVO este prezentat atât în cazul în care sunt folosite doar acestea cât și în cazul în care sunt folosite împreună cu caracteristicile de detecție (DET) și de clasificare (CLS). În cazul folosirii doar a caracteristicilor SVO, cel mai bun rezultat este obținut cu metoda LSTM2-ATT(SVO), sub ambele metrici de evaluare (BLEU@4 52.0%, METEOR 32.3%). Atunci când folosim toate caracteristicile semantice, metoda care are cele mai bune rezultate sub metrica BLEU este LSTM-ATT(SVO,DET,CLS) (50.6%) iar conform metricii METEOR LSTM2-ATT(SVO,DET,CLS) (32.4%).

Rezultate calitative. Mecanismul de atenție propus peste propunerile spațio-temporale de obiecte ne permite obținerea unei explicații vizuale a ceea ce a considerat modelul că este cel mai relevant suport vizual pentru emiterea unui anumit cuvânt.



Ours: A **man** is **pouring** some **sauce**.
Ref: A man is pouring some sauce into a pan.

Ours: A **man** is **playing guitar**.
Ref: A man is playing a guitar.

Ours: A **man** is **shooting** with a **gun**.
Ref: A guy is shooting a gun.

Ours: A **dog** is **playing** with a **dog**.
Ref: A dog is playing with a fly.

Figura 2-2: Propunerile cu cea mai mare pondere pentru emiterea fiecărui cuvânt din propoziție. Arătăm doar suportul spațio-temporal al cuvintelor principale, ignorând cuvintele de legătură. Propoziția completă este arătată în dreapta împreună cu cea mai apropiată propoziție de referință (adnotată de oameni). Pentru fiecare propunere arătăm un singur cadru ales aleator.

Acest lucru poate fi făcut prin inspectarea ponderilor învățate de model precum și a propunerilor asociate cu acestea. În figura 2-2 arătăm propunerea spațio-temporală care a avut cea mai mare pondere în generarea unui anume cuvânt.

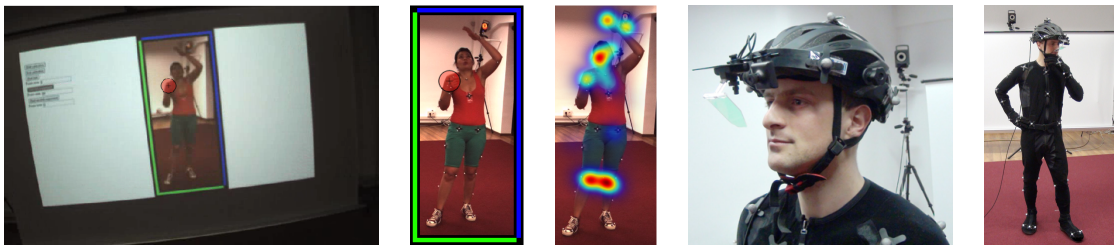
Metoda	**BLEU@1**	**BLEU@2**	**BLEU@3**	**BLEU@4**	**METEOR**
FGM [19]	-	-	-	13.68	23.9
S2VT[20]	-	-	-	-	29.8
MM-VDN[22]	-	-	-	37.64	29.00
LSTM-YT-coco [21]	-	-	-	33.29	29.07
LSTM-YT-coco+flicker [21]	-	-	-	33.29	28.88
Temporal attention [23]	-	-	-	41.92	29.60
LSTM-E (VGG+C3D) [16]	78.8	66.0	55.4	45.3	31.0
h-RNN[24]	81.5	70.4	60.4	49.9	32.6
HRNE with attention[15]	79.2	66.3	55.1	43.8	**33.1**
GRU-RCNN [1]	-	-	-	49.63	31.70
LSTM	78.0	66.4	56.7	45.4	31.2
LSTM(SVO)	80.1	68.1	57.5	45.8	31.2
LSTM(DET,CLS)	81.2	68.9	57.9	46.2	31.1
LSTM(SVO,DET,CLS)	80.8	69.3	59.3	48.3	30.7
LSTM-ATT	80.1	68.9	59.4	48.7	31.9
LSTM-ATT(SVO)	81.0	70.5	61.2	50.5	32.3
LSTM-ATT(DET,CLS)	81.9	70.9	60.9	50.5	31.6
LSTM-ATT(SVO,DET,CLS)	82.0	71.6	62.4	51.5	32.0
LSTM2-ATT(SVO)	**82.4**	**71.8**	**62.5**	**52.0**	32.3
LSTM2-ATT(DET,CLS)	80.6	68.1	57.4	46.0	31.8
LSTM2-ATT(SVO,DET,CLS)	81.5	70.8	61.5	50.6	32.4

Tabela 2.1: Comparație cu metodele precedente folosind metricile BLEU@1–BLEU@4 și METEOR. Valorile sunt raportate ca procente (%).

Capitolul 3

Pictorial Human Spaces: Un studiu computațional privind percepția oamenilor asupra posturilor articulate 3-dimensionale

Analiza mișcării oamenilor în imagini și videouri precum și a componentelor 2d și 3d care reies din aceasta este una din problemele centrale ale problemei vederii automate. Cu toate acestea nu există studii care să dezvăluie cât de bine percep oamenii alți oameni în imagini și cât de precișsi sunt în acest lucru. În acest capitol ne propunem să dezvăluim o parte din procesele – precum și nivelul de precizie – implicate în percepția 3d a oamenilor în imagini, prin evaluarea performanței umane. În plus, arătăm care sunt diferențele cantitative și calitative între oameni și calculator atunci când li se prezintă același stimul vizual și demonstrăm că metrici care incorporează elemente de percepție umană pot produce rezultate mai plauzibile atunci când sunt integrate în algoritmi de estimare automată a posturilor umane.



(a) (b) (c) (d) (e)

Figura 3-1: Ilustrare a aparatului nostru pentru percepția posturilor umane. (a) Ecranul pe care imaginile sunt proiectate, așa cum este capturat de către camera externa a sistemului de captură a mișcărilor ochilor. (b) Rezultatul mapării distribuției fixațiilor pe imaginea de rezoluție înaltă, după ce s-a realizat detecția marginilor, urmărirea și alinierea acestora. (c) Hartă de activări cu distribuția tuturor fixațiilor aparținând unuia dintre participanții la studiu pentru această postură. (d) sistemul nostru de captură a mișcării ochilor care se poartă pe cap, văzut în detaliu. (e) Sistemul de captura 3d a mișcării.

3.1 Aparatul pentru percepția posturilor umane

Propunem un aparat experimental care ne permite să stabilim o legătură între un fenomen parțial subiectiv cum este percepția 3d a posturilor umane și o serie de măsurători cantitative. Abordarea noastră constă în a îmbrăca oamenii într-un costum de captură a mișcării corpului în 3d și a-i echipa cu un sistem de captură a mișcării ochilor în timp ce le arătăm imagini cu oameni în diferite posturi. Aceste imagini au fost obținute la rândul lor folosind același sistem de captură a mișcării corpului în 3d (fig. 3-1). Cerându-le oamenilor să reproducă posturile pe care le arătăm, putem crea o legătură între percepție și măsurători. Folosim un sistem de top pentru captura mișcării (VIicon) împreună cu un sistem mobil de rezoluție înaltă pentru captarea mișcării ochilor.

3.1.1 Proiectarea experimentului și colectarea de date

Participanți și organizare generală. Mai întâi analizăm performanța a 10 participanți, 5 bărbați și 5 femei, care nu au probleme vizuale și nu întâmpină nicio dificultate în mobilitate. Profesia lor nu necesită niște aptitudini neuro-motoare superioare (cum

este de exemplu cazul dansatorilor profesioniști, actorilor sau sportivilor). Acest grup va fi referit ca grupul participanților obișnuiți *regular*. De asemenea, analizăm și performanța a încă 4 participanți, 2 bărbați și 2 femei, care studiază coregrafie pentru balet clasic și modern și sunt în an terminal. Acest grup va fi referit ca grupul participanților calificați *skilled*. Imaginile au fost proiectate pe un ecran de 1.2 metri, la o distanță de 2.5–3 metri față de participanți.

Fiecărui participant i s-a cerut să stea nemișcat și să privească câte o imagine pe rând până când aceasta dispărea, iar apoi să reproducă postura, luându-și cât timp are nevoie. Pentru fiecare postură proiectată pe ecran, înregistrăm atât fixația ochilor cât și mișcările în 3d ale participanților, pe durata reproducerii posturilor, după ce imaginea a dispărut de pe ecran. După ce au trecut cele 5 secunde în care imaginea era arătată, participanții nu mai au posibilitatea să vadă imaginea și trebuie să refacă postura bazându-se pe ce au memorat. Arătăm participanților un număr de 120 de imagini care reprezintă dreptunghiul de delimitare al unei persoane. Majoritatea imaginilor sunt frontale, 100 dintre ele conțin posturi în picioare, ușor de reprodus, iar 20 sunt mai greu de reprodus deoarece presupun așezarea pe podea și rezultă des în acoperirea unor părți din postură. Imaginile arătate sunt selectate din baza de date Human3.6M [9], dintr-o varietate de activități zilnice.

3.2 Analiza datelor

3.2.1 Înregistrările mișcărilor ochilor

Concordanța statică și dinamică. În această secțiune analizăm cât de concordanți au fost participanții privind modul în care au fixat în imagine. În primul rând suntem interesați în a evalua concordanța statică, ce consideră doar locațiile fixate, iar apoi cea dinamică, unde luăm în considerare și ordinea locațiilor fixate. Pentru a evalua cât de mult sunt participanții în acord privind locațiile fixate, prezicem fixațiile fiecărui participant, pe rând, folosindu-ne de fixațiile celorlalți participanți [6, 13]. Acest lucru în facem luând în considerare aceiași postură, precum și posturi diferite. Fig. 3-

2 indică o concordanță bună a participanților.

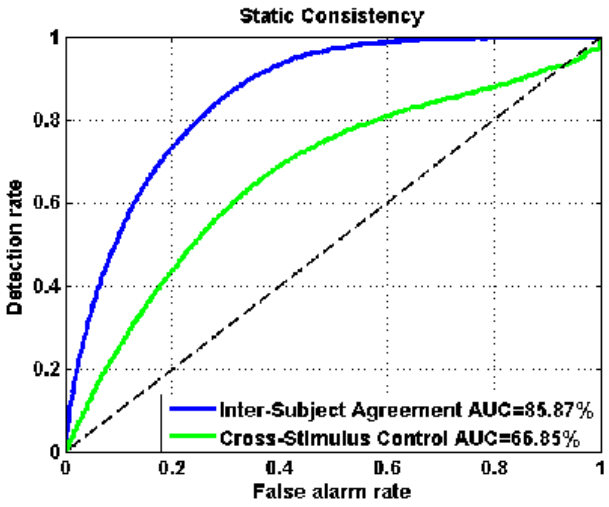


Figura 3-2: Concordanța statică între participanți. Fixațiile unui participant sunt prezise folosind datele celorlalți atât în cadrul aceleiași imagini (cu albastru) cât și selectând aleator o altă imagine din cele 120 (cu verde).

Pentru a evalua concordanța participanților privind ordinea în care fixează regiunile de interes, ne folosim de modelul Markov cu stări ascunse propus de [13]. Stările corespund regiunilor de interes fixate de participanți, iar tranzițiile corespund sacadelor. Pentru fiecare postură arătată, învățăm un model dinamic din traiectoriile a 9 participanți și calculăm verosimilitatea traiectoriei celui de-al zecelea participant sub modelul antrenat. Acest proces este repetat pe rând pentru fiecare dintre participanți, și se calculează media peste verosimilități. Media verosimilităților (normalizată prin dimensiunea traiectoriei) este -9.38. Rezultatele sunt comparate cu verosimilitatea unei traiectorii generate aleator precum și cu verosimilitatea unei traiectorii obținute privind o altă postură. Verosimilitatea medie este mult mai mică în cazul traiectoriilor generate aleator (-42.03) decât în cazul celor care aparțin participanților. De asemenea, verosimilitatea traiectoriilor obținute privind alte posturi -17.12 este mult mai mică decât verosimilitatea traiectoriilor obținute pe baza aceleiași imagini, ceea ce indică o concordanță bună a participanților privind ordinea de fixare a regiunilor de interes.

Care sunt articulațiile cele mai fixate? Vrem să studiem dacă anumite articulații

ale posturii umane sunt fixate mai mult decât altele și dacă acest lucru se întâmplă indiferent de postura arătată sau dacă diferă în funcție de postură. În acest scop, ne uităm la numărul fixațiilor care se suprapun peste o anumită articulație. Figura Fig. 3-3 arată distribuție fixațiilor peste articulațiile corpului, făcând media peste toate posturile arătate. Putem observa că încheieturile mâinilor și zona capului sunt cele mai privite, iar tendința este de a fixa mai mult părțile superioare ale corpului decât părțile inferioare.

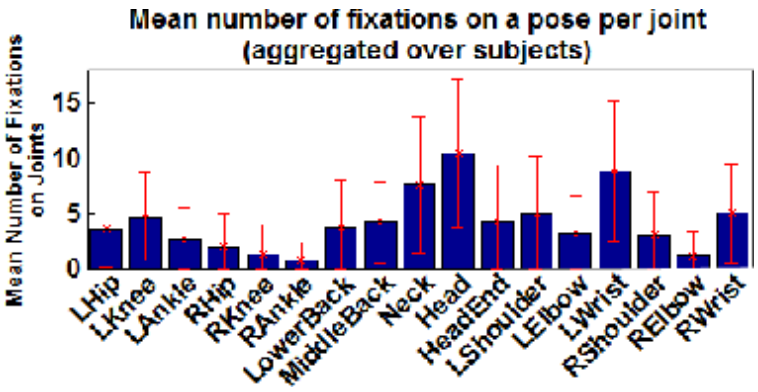


Figura 3-3: Numărul fixațiilor peste fiecare articulație. Media și deviația standard sunt calculate peste cele 120 de imagini, agregând informația de la cei 10 participanți obișnuiți pentru fiecare postură.

3.2.2 Reproducerea posturilor 3d

În această secțiune completăm studiul mișcării ochilor cu o analiză privind precizia cu care oamenii pot reproduce posturi ale altor oameni arătate în imagini.

Cât de precis reproduc oamenii posturi 3d? Am comparat precizia reproducerii posturilor de către participanții obișnuiți și de către cei calificați atunci când li se arată aceleași imagini. Ne dorim să înțelegem dacă o educație formală pentru profesii ce necesită aptitudini neuro-motoare ridicate, și o postură corectă a corpului influențează percepția și capacitatea de a reproduce posturi, măsurată sub metrica MPJPE. Tabelul 3.1 arată erorile de reproducere sub metrica MPJPE atât pentru participanții obișnuiți cât și pentru cei calificați. Se poate observa că eroarea participanților calificație este cu doar 7.1 mm mai mică decât eroarea participanților

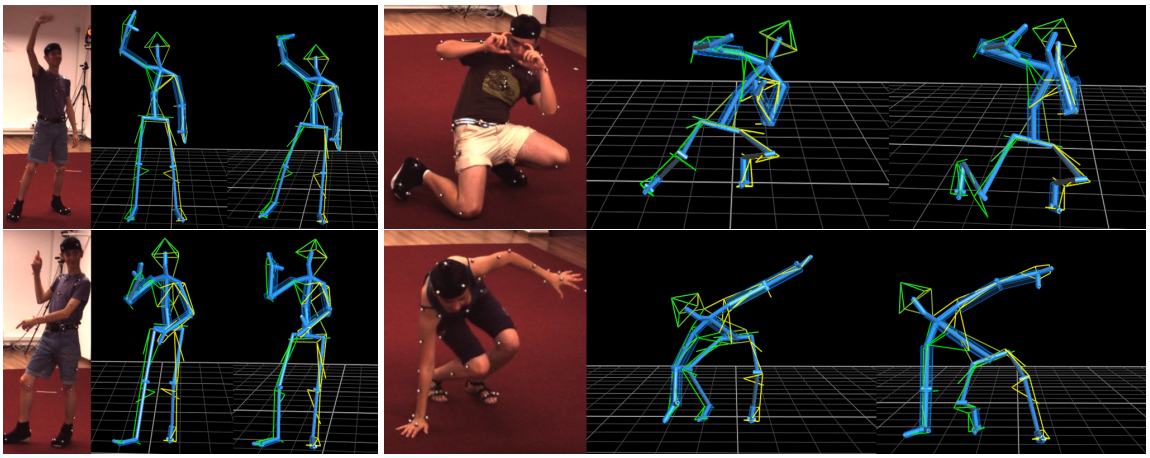


Figura 3-4: Exemple de reproduceri pentru două posturi ușoare (stânga) și două grele (dreapta). Pentru fiecare postură prima reproducere aparține unui participant obișnuit, iar cea de-a doua unui participant calificat.

obișnuiți. Această diferență mică între cele doua grupuri poate sugera următoarele: (1) metrica de eroare folosită nu este suficient de sensibilă pentru a surprinde criteriile optimizate de oameni atunci când percep și reproduc o postură 3d; (2) capacitatea ridicată de coordonare a corpului și mobilitatea superioară nu aduc un beneficiu în contextul studiului nostru. Exemple de reproduceri atât ale subiecților obișnuiți cât și ale celor calificați pot fi văzute în figura 3-4.

	Eroarea MPJPE (mm)		
Metoda	Ușoare	Dificile	Ambele
Participanți obișnuiți	91.7 ± 35.9	156.7 ± 58.1	102.5 ± 58.1
Participanți calificați	83.9 ± 31.6	153.1 ± 65.6	95.4 ± 47.3
KDE	100.33 ± 39.54	267.42 ± 133.22	128.18 ± 89.45

Tabela 3.1: Rezultate reproducerii posturilor umane și estimarea algoritmului KDE atât în cazul posturilor ușoare cât și în cazul celor dificile sub metrica MPJPE.

3.3 Metrici perceptuale pentru estimarea automată a posturilor umane 3d

În această secțiune ne concentrăm pe două aspecte: (1) să înțelegem care sunt tipurile de erori pe care le fac oamenii în procesul de reproducere a posturilor și să le comparăm cu cele generate automat; (2) să învățăm metrici perceptuale care reflectă mai bine semantica unei posturi umane.

3.3.1 Diferențe între performanța oamenilor și cea a unui sistem automat

Ne propunem să dezvăluim diferențele cantitative și calitative între posturile reproduse de oameni și cele estimate în mod automat, pornind de la aceleași imagini. Folosim un model de predicție structurat, Kernel Dependency Estimation (KDE) [5], pentru a învăța o mapare a caracteristicilor extrase dintr-o imagine a unei persoane într-o reprezentare a posturii ei în spațiul 3d sub forma unui lanț cinematic al articulațiilor principale. Pentru a antrena modelul ne folosim de baza de date Human80K. Aceasta este un subset al bazei Human3.6m, din care am ales cele 120 de imagini pe care le-am arătat participanților.

Care este diferența dintre eroarea obținută de oameni și un model de învățare automată (KDE)? În tabelul 3.1 arătăm media erorii de reproducere pentru participanții obișnuiți și pentru cei calificați precum și media erorii de predicție a algoritmului KDE sub metrica MPJPE. În cazul posturilor ușoare diferența dintre participanții calificați și predicțiile KDE este de 16mm, iar diferența dintre reproducerile participanților obișnuiți și KDE este de 9mm. În cazul posturilor grele, diferența dintre participanți și KDE este mult mai mare: 110mm în cazul participanților obișnuiți și 113mm în cazul celor calificați. Aceasta indică faptul că deși atât performanța oamenilor cât și a algoritmului KDE este afectată de dificultatea posturilor, algoritmul întâmpină dificultăți mai mare în estimarea corectă a posturilor decât oamenii, atunci când imaginile arătate reprezintă posturi parțial acoperite sau întinse pe jos.

3.3.2 Învățarea metricilor perceptuale

În această secțiune prezentăm o abordare pentru învățarea unor metrici care să surprindă diferențele perceptuale între posturi. În acest mod țintim să reducem diferența dintre modul cum percep oamenii o postură și ceea ce metrica MPJPE evaluează. Ne folosim de reproducerile de posturi obținute atât de la participanții obișnuiți cât și de la cei calificați pentru a învăța o metrică perceptuală peste posturi. Pentru aceasta ne folosim de metoda Relevant Component Analysis (RCA) [2] care modifică spațiul caracteristicilor printr-o transformare globală liniară, atribuind ponderi ridicate "caracteristicilor relevante" și ponderi scăzute celor "irelevante".

3.3.3 Metrici perceptuale în estimarea automată a posturilor

În această secțiune integrăm metrica perceptuală nou învățată în modelul KDE. Antrenăm modelul pe baza de date Human80K [8] folosind atât un kernel Gausian $K_Y(x,y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$, cât și un kernel perceptual $K_Y(x,y) = e^{-\frac{d_P^2(x,y)}{2\sigma^2}}$ peste variabilele țintă, unde $d_P(x,y)$ este metrica perceptuală între posturile x și y învățată în §3.3.2.

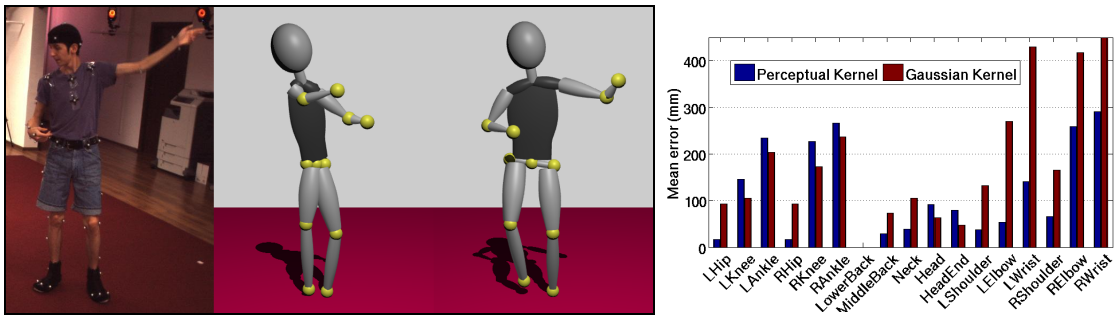
În tabelul 3.2 arătăm media MPJPE și media erorii perceptuale în cazul posturilor obținute cu kernelul Gausian și respectiv cu cel perceptual. Se poate observa că acelea obținute cu kernelul perceptual au erori mai mici sub ambele metrici.

Metoda	MPJPE mediu	Eroarea perceptuală medie
KDE - Gaussian	113.07 ± 72	18.44 ± 10.25
KDE - Perceptual	109.16 ± 66	14.92 ± 9.43

Tabela 3.2: Eroarea posturilor estimate automat folosind un kernel Gausian și respectiv unul perceptual.

Posturi care au erori MPJPE similare pot arăta perceptual foarte diferit dacă un subset de articulații au erori foarte mari. În figura 3-5 arătăm un exemplu în care estimarea obținută folosind un kernel perceptual, deși nu este perfectă, arată calitativ mai bine decât cea obținută folosind kernelul Gausian. Arătăm de asemena

și distribuția erorilor articulațiilor pentru cele două estimări. Se poate vedea că deși multe dintre articulații au erori similare, 4 dintre acestea (încheieturile mâinilor, și coatele) au erori foarte mari în cazul estimării cu kernelul Gausian ceea ce face postura foarte diferită de cea de referință.



(a) (b) (c) (d)

Figura 3-5: a) Imagine de test, b) Estimarea KDE folosind kernel Gausian, c) Estimarea KDE folosind kernel perceptual, d) Eroarea pe fiecare articulație.

Capitolul 4

Estimarea monoculară a posturii și a formei corpului pentru mai mulți oameni în imagini naturale *Importanța multiplelor constrângeri de scenă*

Modelele automate pentru înțelegerea vizuală a oamenilor au beneficiat de progresul obținut în rețelele de învățare profundă, în modelarea parametrică a oamenilor și în construirea de baze de date 2d și 3d de dimensiuni mari. Cu toate acestea, modelele 3d existente au presupuneri rigide în ceea ce privește scena, considerând fie că este o singură persoană în imagine, fie că aceasta este vizibilă în totalitate, fie că există un fundal simplu sau mai multe camere. În acest capitol integrăm o rețea neuronală profundă cu sarcini multiple cu un model uman parametric și elemente de modelare a scenei. Arătăm experimente atât pe imagini cu o sigură persoană cât și cu mai multe persoane și evaluăm în mod sistematic fiecare componentă a modelului, demonstrând o performanță sporită. De asemenea, aplicăm modelul nostru și în cazul imaginilor cu mai mulți oameni, acoperiți parțial, și cu fundaluri diverse, obținând rezultate de calitate din punct de vedere perceptual.

4.1 Modelarea mai multor persoane în scenă

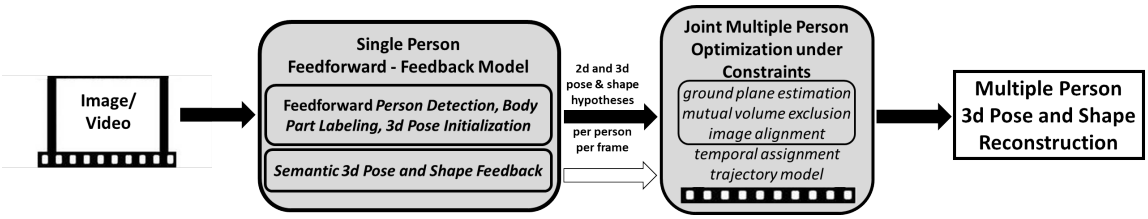


Figura 4-1: Schema modelului nostru monocular de estimare a posturii 3d și a formei corpului pentru mai multe persoane. Sistemul nostru combină un model pentru o singură persoană care încorporează o inițializare *feedforward* și un *feedback* semantic cu constrângeri de scenă cum ar fi estimarea unui plan comun, constrângerea ca două persoane să nu se întrepătrundă și inferența comună pentru mai multe persoane. Pentru un video monocular, urmărirea oamenilor de-a lungul videoului este rezolvată folosind algoritmul Ungar, iar optimizarea traiectoriei este realizată împreună, pentru toți oamenii, ținând cont de toate constrângerile și de concordanța cu imaginea.

Formularea problemei. Fără a restrânge generalitatea, considerăm N_p persoane unic detectate într-un video cu N_f cadre. Obiectivul nostru este de a infera cele mai bune variabile care controlează postura $\mathbf{\Theta} = [\boldsymbol{\theta}_p^f] \in \mathbb{R}^{N_p \times N_f \times 72}$, parametrii care controlează dimensiunea corpului $\mathbf{B} = [\boldsymbol{\beta}_p^f] \in \mathbb{R}^{N_p \times N_f \times 10}$ și distanța față de cameră a fiecărei persoane $\mathbf{T} = [\mathbf{t}_p^f] \in \mathbb{R}^{N_p \times N_f \times 3}$, cu $p \in N_p$ și $f \in N_f$. Ca prima pas, definim un cost per cadru, pentru o singură persoană $L_I^{p,f}(\mathbf{B}, \mathbf{\Theta}, \mathbf{T})$:

$$L_I^{p,f} = L_S^{p,f} + L_G^{p,f} + L_R^{p,f} + \sum_{\substack{p'=1 \\ p' \neq p}}^{N_p} L_C^f(p, p'), \tag{4.1}$$

unde costul L_S ia în calcul evidența vizuală calculată în fiecare cadru sub forma etichetării semantice a corpului, L_C penalizează întrepătrunderea a două persoane, și L_G încorporează constrângerea ca unii oameni stau pe același plan suport. Termenul $L_R^{p,f} = L_R^{p,f}(\boldsymbol{\theta})$ este o mixtură de Gausiene similară cu [3]. Costul pentru mai mulți oameni aflați într-o imagine, ținând cont de toate constrângerile este

$$L_I^f = \sum_{p=1}^{N_p} L_I^{p,f}. \tag{4.2}$$

Dacă avem la dispoziție un video monocular, costul static L^f este augmentat cu un model de traiectorie aplicat fiecărei persoane, după ce am rezolvat urmărirea fiecărei persoane de-a lungul videoului. Costul complet pentru un video este

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_T = \sum_{p=1}^{N_p} \sum_{f=1}^{N_f} \left(L_I^{p,f} + L_T^{p,f} \right), \tag{4.3}$$

unde L_T poate încorpora cunoștințe anterioare despre mișcarea umană, cum ar fi faptul că aceasta este netedă în timp, presupuneri legate de viteză sau accelerație constantă sau modele mai sofisticate de mișcare învățate din date.

Pentru a infera postura și poziția 3d a mai multor persoane ne bazăm pe un model uman parametric, SMPL [12], și pe o rețea profundă cu sarcini multiple pentru înțelegerea vizuală a oamenilor, DMHS [18]. În practică nu putem presupune un număr constant de persoane de-a lungul unui video și mai întâi trebuie să inferăm parametrii $\mathbf{B}$,$\mathbf{\Theta}$, $\mathbf{T}$ independent pentru fiecare cadru, minimizând suma primelor două funcții de cost L_S și L_C. Apoi urmărim temporal persoanele obținute în fiecare cadru rezolvând în mod optim o problemă de atribuire și apoi reoptimizăm funcția obiectiv adăugând constrângerile de a avea un suport de plan comun L_G, și constrângerea temporală, L_T. O vedere de ansamblu a metodei este arătată în figura 4-1.

4.2 Detalii experimentale

Am validat experimental metoda noastră pe două baze de date, CMU Panoptic [10] și Human3.6M [9], și am evaluat calitativ pe videouri naturale (vezi figura 4-2). Fiind dat un video cu mai multe persoane, mai întâi detectăm persoanele în fiecare cadru și apoi, pe baza metodei DMHS, obținem predicții pentru fiecare articulație 2d, segmentarea semantică și postura 3d.

Human3.6M este o bază de date de dimensiuni mari, care conține imagini ale unei singure persoane înregistrate în condiții de laborator folosind un sistem de captură al mișcării. Am selectat trei din cele mai dificile trei acțiuni (*sitting*, *sitting down* și *walking dog*) pentru a evalua modelul pentru o singură persoană. Pentru evaluare

Metoda	Walking Dog	Sitting	Sitting Down
DMHS [18]	78	119	106
Cost semantic	75	109	101
Multi View	51	71	65
Netezire	**48**	**68**	**64**

Tabela 4.1: Eroarea medie la nivel de articulație 3d (în mm) pe baza de date Human3.6M, evaluată pe partea de test a trei acțiuni dificile. Observăm importanță diferitelor constrângeri în îmbunătățirea erorii de estimare.

	Haggling		Mafia		Ultimatum		Pizza		Medie	
Metoda	Postură	Translație	Postură	Translație	Postură	Translație	Postură	Translație	Postură	Translație
DMHS [18]	217.9	-	187.3	-	193.6	-	221.3	-	203.4	-
Cost 2d	**135.1**	282.3	174.5	502.2	**143.6**	357.6	177.8	419.3	157.7	390.3
Cost semantic	144.3	260.5	179.0	459.8	160.7	376.6	178.6	413.6	165.6	377.6
Netezire	141.4	260.3	173.6	454.9	155.2	368.0	173.1	403.0	160.8	371.7
Netezire Estimare plan	140.0	**257.8**	**165.9**	**409.5**	150.7	**301.1**	**156.0**	**294.0**	**153.4**	**315.5**

Tabela 4.2: Erorile (în mm) pentru estimarea posturii 3d și a translației pe baza de date Panoptic (9,600 de cadre și 21,404 de detecții de persoane). Aceste rezultate demonstrează beneficiile fiecărei componente și importanța constrângerii la nivel de plan, în special, pentru estimarea corectă a translației.

folosim partea oficială de test corespunzătoare celor trei acțiuni; aceasta conține aproximativ 160,000 de cadre. Rezultatele sunt afișate în tabelul 4.1. Metoda propusă obține îmbunătățiri față de metoda de bază, DMHS, folosind constrângeri la nivel semantic și de formă.

Baza de date CMU Panoptic. Am utilizat patru activități (*Haggling*, *Mafia*,*Ultimatum* și *Pizza*) care conțin mai multe persoane interacționând între el. În total, am folosit 9,600 de cadre, respectiv 21,404 de detecții de persoane. Menționăm că nu am antrenat nicio componentă a metodei noastre pe această bază de date.

Metodologia de evaluare. Pentru evaluarea fiecărei posturi prezise mai întâi o aliniem cu postura de referință pe baza articulației pelvisului, iar apoi calculăm folosind erorii peste fiecare articulație (en., *mean per joint position error*; MPJPE). Translația estimată este evaluată folosind distanța Euclidiană. Aceste două metrici sunt calculate independent pentru fiecare cadru al unei secvențe date și apoi mediate peste toate persoanele și cadrele.

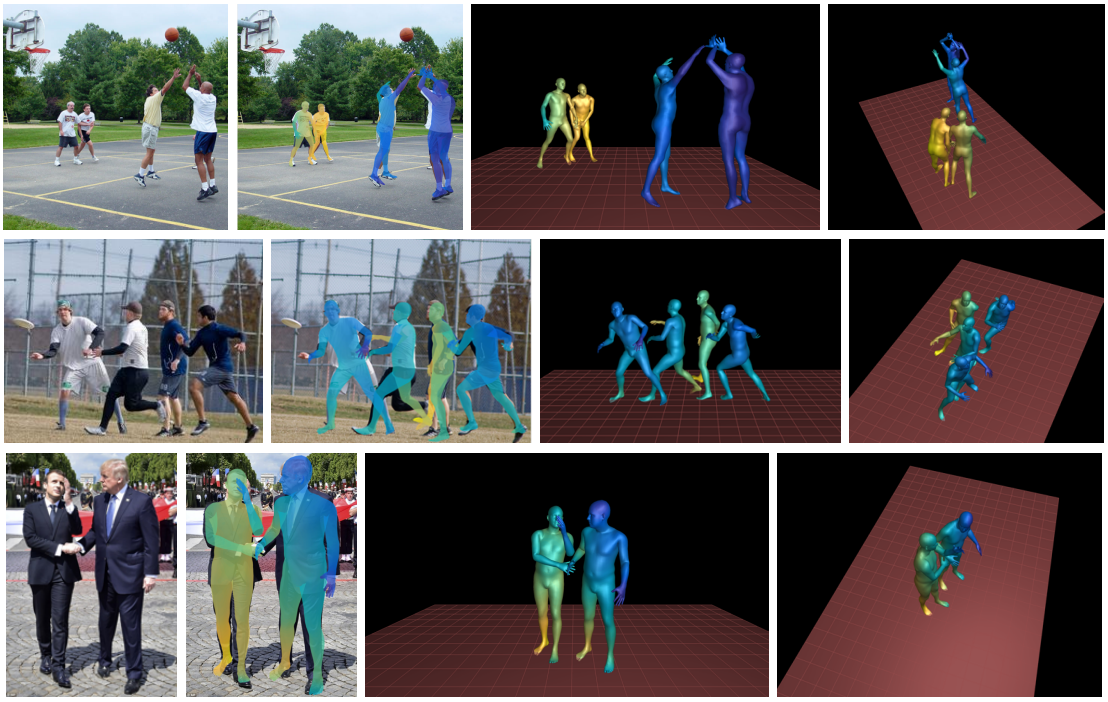


Figura 4-2: Exemple de reconstrucții automate 3d a mai multor persoane din imagini monoculare și naturale. Pe fiecare rând, de la stânga la dreapta avem: imaginea originală, modelul prezis suprapus peste imaginea originală, două vederi din unghiuri diferite ale reconstrucțiilor 3d (acestea includ și planul de bază estimat). Remarcăm că posturile dificile, ocluziile, interacțiunile dintre subiecți sunt toate rezolvate corect în reconstrucția finală.

Studiu sistematic. Am testat în detaliu fiecare componentă din sistemul propus și am colectat rezultatele în tabela 4.2. Comparativ cu metoda DMHS, metoda noastră reduce semnificativ eroarea MPJPE, de la 203.4 mm la 153.4 mm (o scădere relativă de aproximativ 25%), și, în plus, calculează și translația fiecărei persoane în scenă. Eroarea de translație este în medie de 315.5 mm. Constrângerile bazate pe proiecția semantică elimină ambiguitățile pozițiilor 3d în cazul a mai multor persoane și reduce astfel eroarea de translație comparativ cu metoda care utilizează doar constrângeri bazate pe proiecțiile 2d. Metoda de netezire temporală îmbunătățește în continuare eroarea de translație. Cea mai importantă componentă este, în schimb, constrângerea bazată pe planul de bază – această aduce o îmbunătățire relativă de 15% a erorii de translație, de la 371 mm la 315 mm. Din punct de vedere calitativ, metoda noastră produce reconstrucții 3d care sunt perceptual plauzibile și care se aliniază corect cu imaginile originale, deși sarcina este una dificilă (prezența a mai multor persoane, vizibilitate parțială a unor persoane, unghiuri de captură neconvenționale).

În concluzie, am prezentat un model monocular care estimează într-un mod integrat posturile 2d și 3d și forma corpului pentru mai multe persoane, fiind ghidat de constrângeri de scenă multiple.

Bibliografie

[1] N. Ballas, L. Yao, C. Pal, and A. C. Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

[2] A. Bar-hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *International Conference on Machine Learning*, 2003.

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[4] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, June 2011.

[5] C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *International Conference on Machine Learning*, pages 153–160, 2005.

[6] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 2009.

[7] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[8] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1661–1668, 2014.

[9] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. Submitted.

[10] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[11] A. Lavie and A. Agarwal. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248:1–16, 2015.

[13] S. Mathe and C. Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in Neural Information Processing Systems*, 2013.

[14] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014.

[15] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *arXiv preprint arXiv:1511.03476*, 2015.

[16] Y. Pan, T. M. , T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *arXiv preprint arXiv:1505.01861*, 2015.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[18] A. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[19] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, August 2014.

[20] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *ICCV*, 2015.

[21] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL HLT*, 2015.

[22] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, and K. Saenko. A multi-scale multiple instance video description network. In *arXiv preprint arXiv:1505.05914*, 2015.

[23] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.

[24] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, June 2016.