ACADEMIA ROMÂNĂ

INSTITUTUL DE MATEMATICĂ SIMION STOILOW

Modele pentru Detecția și Recunoașterea Automată a Activităților Umane în Video

Autor,
Mihai ZANFIR

Conducător,
C.S. I Dr. Cristian SMINCHIȘESCU

REZUMAT TEZĂ DE DOCTORAT

București
2018

Capitol 1 - Introducere.

Această teza se concentrează pe problema recunoașterii de activități umane din reprezentări tridimensionale ale corpului uman. Aceasta este o problema importantă, cu numeroase aplicații în domenii precum sisteme de supraveghere de interior, divertisment, conducere autonomă și interacțiune om-robot.

Dificultatea problemei de înțelegere a omului din imagini provine din provocările date de detecția părțile corpului uman și din cuplarea sarcinilor de recunoaștere și reconstrucție. În cazul primei provocări, trebuie să se țină seama de variabilitatea posturii și formei corpului uman, diversitatea scenelor și a dezordinii de fundal. Pentru ultima provocare, trebuie să se țină cont de inter- și intra- variabilitațile diferitelor clase de acțiuni, și de asemenea de erorile de prezicere a modelelor de reconstrucție ale posturii 3d.

Capitol 2.

În Capitolul 2 propunem un cadru de lucru rapid și non-parametric pentru recunoașterea cu latență redusă a acțiunilor și activităților umane. Demonstrăm performanța modelului folosind reconstrucții 3d automate din senzori de adâncime și arătăm cum acest cadru de lucru se pliază natural pe recunoaștere cu latență mică, învățare dintr-un singur exemplu și detecție de acțiuni din date video nesegmentate cu acuratețe mare.

Central metodologiei noastre se află descriptor-ul "Moving Pose", o reprezentare dinamică inedită a framelor video, ce capturează nu numai postura 3d a corpului, ci și viteza și accelerația incheieturilor corpului uman într-o fereastră mică de timp în jurul framei curente. Considerăm că datorită constrângerilor fizice precum inerție sau latență în actuarea musculară, mișcările corpului asociate unei acțiuni pot fi bine aproximate cel mai adesea de o funcție pătrătica, ce ține cont de prima și de a două derivată a mișcării corpului uman relativ la timp.

Inspirați de aceasta, pentru fiecare framă dintr-o secvență video calculăm descriptorul "Moving Pose" (MP), ca o concatenare a posturii umane 3d normalizate $\mathbf{P} = [\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}]$ împreună cu derivatele sale de ordin 1 și 2, $\delta\mathbf{P}(t_0)$ și $\delta^2\mathbf{P}(t_0)$. Derivatele sunt estimate numeric folosind o fereastră temporală de 5 frame centrată în frama pe care o procesăm curent: $\delta\mathbf{P}(t_0) \approx \mathbf{P}(t_1) - \mathbf{P}(t_{-1})$ and $\delta^2\mathbf{P}(t_0) \approx \mathbf{P}(t_2) + \mathbf{P}(t_{-2}) - 2\mathbf{P}(t_0)$. Pentru o aproximare numerică mai bună, mai întâi aplicăm fiecărei coordonate ale vectorului

de postură uman normalizat, de-a lungul dimensiunii de timp, un filtru gaussian 5 pe 1 ($\sigma = 1$).

Descriptorul MP propus codifică informația cinematică și de postura pentru a descrie segmentele de acțiuni. Pentru a scoate în evidență puterea lui discriminativa și pentru a oferi flexibilitate acestuia în faza de antrenament (incluzând învățarea dintr-un singur exemplu), folosim o schemă non-parametrică de clasificare de acțiuni bazată pe metoda de k-cei mai apropiați vecini (kNN).

Metoda noastră cu latență redusă pentru clasificarea unei secvențe de testare se aplică după cum urmează: la momentul de timp t, după ce am observat t frame de test, lăsăm fiecare dintre descriptorii vecini kNN din setul de antrenament să voteze pentru clasa din care face parte, pentru un număr total de kt voturi. Pentru a lua o decizie, aplicăm o schemă simplă de respingere. Dacă voturile acumulate pentru cea mai reprezentată clasa c_j sunt suficiente comparativ cu toate celelalte clase și destul de multe frame au fost observate, atunci o raportăm drept câștigătoare. De asemenea, învățăm un scor pentru votul fiecărei frame din setul de antrenament, astfel încât framelor cele mai reprezentative pentru o anumită acțiune le este alocată o putere discriminativa mai mare. Abordarea noastră completă este să incorporam și informație temporală globală pentru a limita căutarea de vecini apropiați doar către acele mostre ce sunt localizate într-o poziție similară în secvență de antrenare, relativ la prima framă.

Combinând descriptorii MP locali și discriminativi, cu o schemă de clasificare conștientă de temporalitate, ne permite să ținem cont de două aspecte importante în clasificarea de acțiuni: puterea discriminativa a framelor cheie și dinamica lor, precum și de scurgerea temporală a unei acțiuni.

Sistemul nostru (vezi Table 1) îmbunătățește metoda de top actuală cu 3.5% pe setul de date MSR Action3D. Acest set de date constă din secvențe de acțiuni temporal segmentate, capturate cu o camera RGB-D. Scheletul 3d, reprezentat de un set de poziții 3d ale incheieturilor corpului, este disponibil pentru fiecare framă și este obținut prin metoda lui [3].

Capitol 3.

În capitolul 3 dezvoltăm în continuare reprezentări ce nu sunt doar bazate pe postura

Tabela 1: Studiu comparativ de clasificare pe setul de date MSR Action3D.

Metoda	Acuratețe(%)
Recurrent Neural Network [10]	42.5
Dynamic Temporal Warping [12]	54
Hidden Markov Model [9]	63
Latent-Dynamic CRF [11]	64.8
Canonical Poses [2]	65.7
Action Graph on Bag of 3D Points [7]	74.7
Latent-Dynamic CRF [11] + MP	74.9
EigenJoints [18]	81.4
Actionlet Ensemble [16]	88.2
MP (Ours)	**91.7**

corpului uman și generalizăm formularea "Moving Pose", considerând relații topologice între perechi de caracteristici simple (*e.g.* poziții 2d sau 3d). De asemenea, exploatăm relațiile cinematice ale acestor relații pentru a forma sub-grafuri de decriptori de frame potrivite pentru recunoașterea diferitelor tipuri de acțiuni. Spre deosebire de "Moving Pose", aceste relații se află la un nivel mai înalt de abstractizare. În loc de a măsura în mod exact geometria, noi construim modele bazate pe clasificatori "soft" ce răspund la relații dintre diferitele încheieturi ale corpului uman și obiecte. Considerăm trei tipuri de relații topologice: *sus-jos*, *stânga-dreapta* și *față-spate*. Analizând cazul 2d, fără a pierde din generalitate, presupunem ca ne sunt data o pereche de puncte de caracteristici (i, j), la locațiile spațiale $p_i = (x_i, y_i)$ și $p_j = (x_j, y_j)$. Atunci există două tipuri de relații topologice ce pot fi modelate cu un predictor categoric folosind funcțiile logistice, $R_x(x_i, x_j)$ și $R_y(y_i, y_j)$. R_x și R_y sunt clasificatori binari "soft" ce răspund la relațiile topologice *stânga-dreapta* și *sus-jos*:

$$R_x(x_i, x_j) = \frac{1}{1 + \exp(-w_x(x_i - x_j))} \quad R_y(y_i, y_j) = \frac{1}{1 + \exp(-w_y(y_i - y_j))}. \qquad (1)$$

Dându-se locațiile incheieturilor corpului uman și a obiectelor din scenă, există un număr exponențial de subseturi de relații ce se poate forma. În shimb, pentru o clasificare precisă e nevoie doar de un grup mic de relații. Găsirea acestui set mic necesită o procedură de căutare eficientă în spațiul mare de seturi posibile. De exemplu, pentru un număr de 20 de

încheieturi, numărul posibil de relații ce se poate forma este de 570 și prin urmare avem un număr de 2^{570} descriptori de framă "Moving Relations" (MR). Formalizăm sarcina de căutare a descriptorului optim d_a^* pentru o clasa de acțiuni dată a, ca maximizarea diferenței empirice așteptate dintre răspunsurile clasificatorului "soft" pe secvențele pozitive și pe secvențele negative:

$$d_a^* = \underset{d \in D}{\operatorname{argmax}} \left(\frac{1}{N_a} \sum_{s \in S(a)} C(s,d) - \frac{1}{N_{\neg a}} \sum_{s \in S(\neg a)} C(s,d) \right). \tag{2}$$

Abordarea propusă de noi este să începem de la o metodă stocastică, extinsă din [6], pentru a estima relevanța fiecărei relații și a fiecărei încheieturi în mod separat, și de a forma mai apoi, într-o manieră lacomă, un descriptor inițial MR cu performanță bună pentru toate acțiunile. Începând de la acest descriptor comun MR, folosim o căutare bazată pe algoritmi genetici [5], pentru a învața diferiții descriptori d_a^* ce sunt optimizați pentru fiecare clasă de acțiuni în parte.

Tabela 2: Studiu comparativ a acurateței de clasificare pe setul de date CAD-120.

Metoda	Acuratețe(%)
Moving Pose [20]	67.5
Koppula et. al. [4]	83.1
Moving Relations descoperite fara GA (Noi)	93.3
Moving Relations descoperite cu GA (Noi)	**99.2**

Experimentele noastre demonstrează puterea MR, care în combinație cu o schemă modificată de clasificare kNN, depășește semnificativ în performanță metode curente mai sofisticate. Spre deosebire de multe alte metode, a noastră este robustă la caracteristici lipsă și este aplicabilă în astfel de scenarii fără a necesita modificări.

Capitol 4.

În acest capitol al tezei, propunem o arhitectură profundă, multi-sarcină, pentru înțelegerea automată a corpului uman în 2d și 3d (DMHS), incluzând recunoașterea și reconstrucția, în imagini monoculare. Sistemul prezice segmentarea fundal-față, etichetarea părților corpului la nivel de pixel și estimează postura 2d și 3d a omului în scenă. Conceptual, fiecare din

etapele noastre de procesare produce estimate pentru recunoaștere și reconstrucție și este constrânsa de funcții de cost specifice. Fiecare sarcină constă dintr-un număr de șase etape recurente ce primesc la intrare imaginea, rezultatele etapei anterioare, precum și intrări de la stagiile celorlalte sarcini. Intrările fiecărei etape sunt procesate și combinate individual cu rețele convoluționale pentru a produce ieșirile corespunzătoare. Sarcina de estimare a posturii 2d este bazată pe o arhitectură convoluționala recurentă similară cu cea din [17]. Dându-se o imagine RGB, $I \in \mathbb{R}^{w \times h \times 3}$, ne dorim să prezicem în mod corect locațiile celor N_J încheieturi definite anatomic, $p_k \in \mathcal{Z} \subset \mathbb{R}^2$, with $k \in \{1 \dots N_J\}$.

Pentru segmentarea semantică a părților corpului, atribuim fiecărei locații din imagine $(u, v) \in \mathcal{Z} \subset \mathbb{R}^2$ una dintre cele N_B etichete de părți anatomice ale corpului (incluzând o etichetă suplimentară pentru fundal), b_l cu $l \in \{1 \dots N_B\}$. La fiecare etapă t, rețeaua prezice, pentru fiecare locație de pixel, probabilitatea prezenței fiecărei părți de corp, $B^t \in \mathbb{R}^{w \times h \times N_B}$. La prima etapă de procesare folosim reprezentări convoluționale bazate pe imagine și pe hărțile de activare 2d, J^1, pentru a prezice etichetele curente pentru părți de corp B^1. Pentru fiecare din etapele următoare folosim în plus informația prezentă în etichetele de părți de corp produse în etapa anterioară, B^{t-1}, și ne bazăm pe o serie de patru straturi convoluționale c_B^t ce învață să combine caracteristicile de imagine $\mathbf{x}$ cu B^{t-1}.

Modulul de reconstrucție 3d folosește informație provenită din hărțile de încheieturi 2d și etichetare de părți de corp, J^t și B^t. Adițional, introducem o funcție antrenabilă c_D^t, definită, în mod similar ca c_B^t, peste caracteristicile de imagini, pentru a obține hărțile de activare ale reconstrucțiilor corpului D^t. Acest modul urmează un flux similar celorlalte: refolosește estimate din etape precedente, R^{t-1}, împreună cu ieșirile celorlalte sarcini, pentru a produce hărțile de reconstrucție finale R^t. În figura 0-1 se pot vizualiza etapele de procesare și dependințele acestui modul.

Modul nostru de proiectare a arhitecturii ne permite să legăm un protocol complet de antrenament, folosind multiple seturi de date ce altfel ar acoperi doar o parte din componentele modelului nostru: imagini complexe cu date 2d, dar fără etichete de părți de corp și 3d țintă asociate, sau date complexe 3d cu variabilitate de fundal 2d limitată. În experimente detaliate bazate pe seturi 2d și 3d dificile, ne evaluăm substructurile modelului nostru și efectul diferitelor tipuri de date de antrenare asupra funcției de cost pe sarcini multiple și

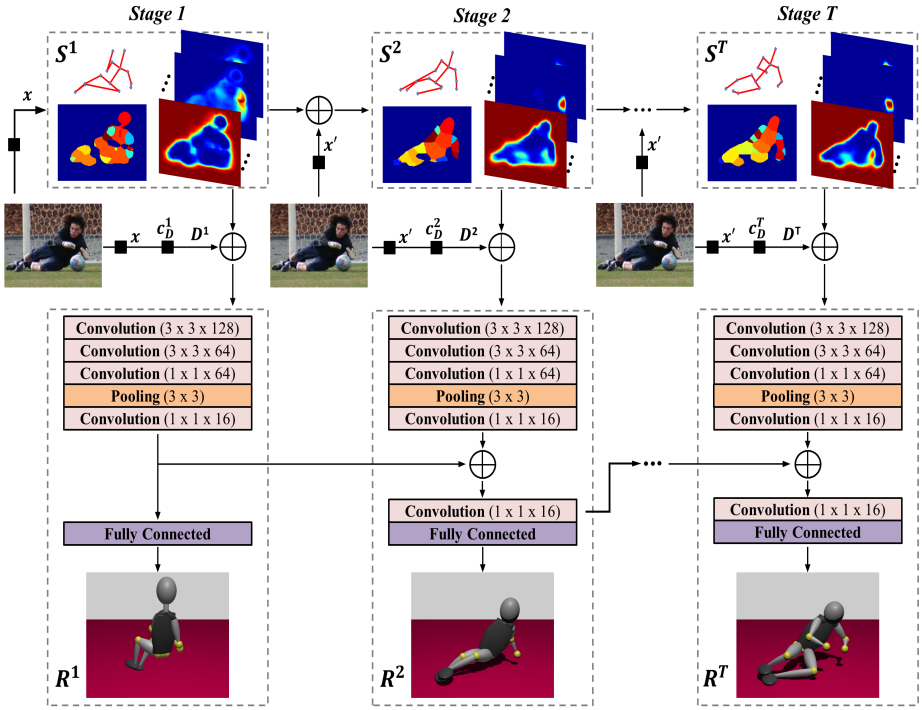

Figura 0-1: Modulul de reconstrucție 3d.

arătăm că obținem rezultate de top.

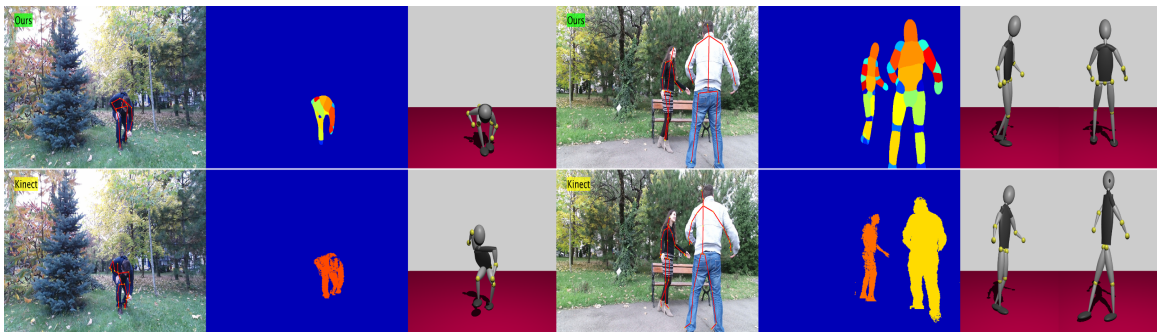

Figura 0-2: Comparație calitativă pe segmentare și reconstrucție între sistemul nostru monocular (sus) și un sistem comercial RGB-D Kinect (jos).

Capitol 5.

În acest ultim capitol de teză, introducem sarcini de recunoaștere de acțiuni și emoții definite pe video-uri reale, înregistrate în timpul unor sesiuni de terapie cu copii autiști din cadrul setului de date multi-modal DE-ENIGMĂ [14]. Sesiunile sunt fie doar cu terapeut sau și asistate de un robot; primele sunt capturate doar de control, în timp ce ultimele sunt

cele ce de interes pentru noi în acest capitol. În terapiile asistate de un robot, un copil și un terapeut sunt așezați la o masă pe care se află un robot. Terapeutul controlează cu o telecomandă un robot ce folosește pentru a atrage atenția copilului în terapie. Sesiunile constau din o parte de joacă liberă (unde copilul se joacă cu ce jucării alege) și o parte efectiv de terapie. Terapia este bazată pe scenarii în care terapeutul arată cărți cu diferite emoții (fericit, trist, supărat, etc.) ce sunt reproduse și de către robot, și copilul trebuie să le imite sau să le potrivească.

Scenariile de terapie acoperă o varietate mare de gesturi corporale și acțiuni executate de către copii. Noi am adnotat un total de 3757 de secvențe video, cu o durată medie de 2.1 secunde. Adnotarea video-urilor de terapie se bazează pe o unealtă web dezvoltată de către noi ce poate să (i) selecteze marginile temporale și să (ii) atribuie o eticheta de clasă de acțiuni. Am inclus și caracteristici ce imbunătățesc experiența de adnotare, precum scurtături pentru ajustări temporale precise, selecție de reluări, vizualizare și filtrare de adnotări precedente, management de sesiune.

Experimentele prezentate în acest capitol folosesc un subset de 2031 de secvențe video adnotate cu etichete din 24 de clase comune de acțiuni pentru toți copiii. Chiar dacă clasele selectate se referă la comportamentul copilului, câteva dintre ele se leagă de terapeut e.g., *Pointing to therapist*, *Turning towards therapist*. O să ne referim la acestea ca secvențe de interacțiune. Printre secvențele adnotate, în jur de o treime sunt secvențe de interacțiune.

Scopul nostru pe termen lung este să interpretăm și să reacționăm în mod automat la acțiunile copilului în cadrul unui mediu dificil precum cel al unei sesiuni de terapie. Pentru a înțelege copilul, ne bazăm pe caracteristici de nivel înalt asociate cu postura și forma corpului său în 3d. Folosim metoda precedent discutată, DMHS, și o îmbunătățim pe sarcina de estimare a posturii 3d a oamenilor parțial vizibili (*i.e.* DMHSPV).

Ne bazăm de asemenea pe modelul prezentat în lucrarea [19] pentru a combina detecția de oameni și prezicerea posturii 2d și 3d din DMHSPV, cu o rafinare volumetrică a formei corpului uman bazat pe reprezentarea SMPL [8] – o numim DMHS-SMPL-F, iar varianta corectată temporal DMHS-SMPL-T.

Am experimentat cu diferite modele de recunoaștere de acțiuni din schelete umane și am efectuat studii comparative pentru diferite tipuri de reconstrucții 2d și 3d ale corpului uman.

Caracteristică de postură	MP - Copil	MP - Copil + Terapeut
Kinect [15]	46.96%	**47.49%**
DMHSPV	32.92%	34.95%
2D [1]	40.83%	44.14
DMHS-SMPL-F	43.53%	45.07%
DMHS-SMPL-T	44.20%	45.68%

Tabela 3: Rezultate comparative pentru diferite metode de estimare a posturii umane pentru clasificare de acțiuni în cadrul de lucru "Moving Pose". De asemenea investigăm impactul modelării și a terapeutului, asupra acurateței clasificării.

O selecție de video-uri din [14], incluzând cele folosite în experimentele de clasificare de acțiuni, a fost adnotată continuu cu emoții în spațiul de valență-excitare, de către 5 terapeuți specializați. Am pre-procesat datele că în [13] pentru a obține valorile pe framă pentru fiecare adnotator și pentru a le potrivi într-un singur semnal țintă de valență-excitare.

Axa emotională	Caracteristica de postură	RMSE ↓	PCC ↑	SAGR ↑
Valența	Kinect	0.116	**0.184**	0.787
	DMHS-SMPL-T	**0.099**	0.169	**0.844**
Excitare	Kinect	0.111	0.345	0.973
	DMHS-SMPL-T	**0.107**	**0.388**	**0.977**

Tabela 4: Prezicere continuă de emoții. Folosind estimatele pentru scheleții 3d de la DMHS-SMPL-T, obținem rezultate similare sau mai bune comparativ cu scheletele 3d produse de Kinect.

Bibliografie

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[2] C. Ellis, S.Z. Masood, M.F. Tappen, J.J. LaViola Jr., and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV*, August 2012.

[3] J. Shotton et al. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

[4] H. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML*, 2013.

[5] R. Leardi, R. Boggia, and M. Terrile. Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, 6, 1992.

[6] S. Li, E.J. Harner, and D.A. Adjeroh. Random knn feature selection-a fast and stable alternative to random forests. *BMC bioinformatics*, 12(1), 2011.

[7] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *WCBA-CVPR*, 2010.

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248:1–16, 2015.

[9] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV*, 2006.

[10] James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. In *ICML*, 2011.

[11] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*. IEEE Computer Society, 2007.

[12] Meinard Müller and Tido Röder. Motion templates for automatic classification and retrieval of motion capture data. In *SCA*, 2006.

[13] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. German Research Center for AI (DFKI), 2010.

[14] J. Shen, E. Ainger, A. M. Alcorn, S. Babović Dimitrijevic, A. Baird, P. Chevalier, N. Cummins, J. J. Li, E. Marchi, E. Marinoiu, V. Olaru, M. Pantic, E. Pellicano, S. Petrovic, V. Petrovic, B. R. Schadenberg, B. Schuller, S. Skendžić, C. Sminchisescu, T. T. Tavassoli, L. Tran, B. Vlasenko, M. Zanfir, V. Evers, and Consortium De-Enigma. Autism data goes big: A publicly-accessible multi-modal database of child interactions for behavioural and machine learning research. *International Society for Autism Research Annual Meeting*, 2018.

[15] J Shotton, A Fitzgibbon, M Cook, T Sharp, M Finocchio, R Moore, A Kipman, and A Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304. IEEE Computer Society, 2011.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[17] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, June 2016.

[18] X. Yang and Y. Tian. Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor. In *CVPR Workshops*, 2012.

[19] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints. In *CVPR*, 2018.

[20] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The "Moving Pose": An Efficient 3D Kinematics Descriptor for Low-Latency Action Detection and Recognition. In *International Conference on Computer Vision*, December 2013.