Fusion of Skeletal and Silhouette-based Features for Human Action Recognition with RGB-D Devices

Alexandros Andre Chaaraoui, José Ramón Padilla-López
Department of Computer Technology
University of Alicante, P.O. Box 99, E-03080, Alicante, Spain
alexandros,jpadilla@dtic.ua.es

Francisco Flórez-Revuelta
Faculty of Science, Engineering and Computing
Penrhyn Road, KT1 2EE, Kingston upon Thames, United Kingdom
F.Florez@kingston.ac.uk

Abstract

Since the Microsoft Kinect has been released, the usage of marker-less body pose estimation has been enormously eased. Based on 3D skeletal pose information, complex human gestures and actions can be recognised in real time. However, due to errors in tracking or occlusions, the obtained information can be noisy. Since the RGB-D data is available, the 3D or 2D shape of the person can be used instead. However, depending on the viewpoint and the action to recognise, it might present a low discriminative value. In this paper, the combination of body pose estimation and 2D shape, in order to provide additional characteristic value, is considered so as to improve human action recognition. Using efficient feature extraction techniques, skeletal and silhouette-based features are obtained which are low dimensional and can be obtained in real time. These two features are then combined by means of feature fusion. The proposed approach is validated using a state-of-the-art learning method and the MSR Action3D dataset as benchmark. The obtained results show that the fused feature achieves to improve the recognition rates, outperforming state-of-the-art results in recognition rate and robustness.

1. Introduction

In recent years, interest has grown on affordable devices (e.g. *Microsoft Kinect* or *ASUS Xtion Pro*) that capture depth quite reliably. Such devices provide a depth image (D), along with an RGB image (thus RGB-D). A depth image can be further processed to obtain body pose estimation by means of a silhouette, a volume, or a skeleton model consisting of a series of joint data which, in turn, can be fed to a machine learning algorithm in order to learn and recognise poses, actions, or complex activities (see Figure 1).

Reliability and accuracy of RGB-D devices have been studied in several works [1, 14], which show that the extraction of a skeleton from depth information is not straightforward. Among several difficulties, lack of precision and occlusions, caused by body parts or other objects present in the scene, stand out [10, 17, 22]. Therefore, in order to provide robustness to human action recognition, possible errors in skeletal data should be considered, either improving the tracking and body pose estimation process, or relying on additional data as RGB colour, 3D (volume) or 2D (silhouette) depth-based information.

The silhouette of the human body has been proven to be

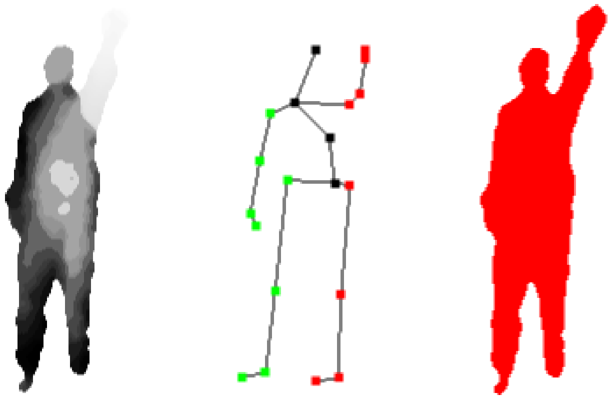


Figure 1: Depth image (left), extracted skeleton (centre) and silhouette (right).

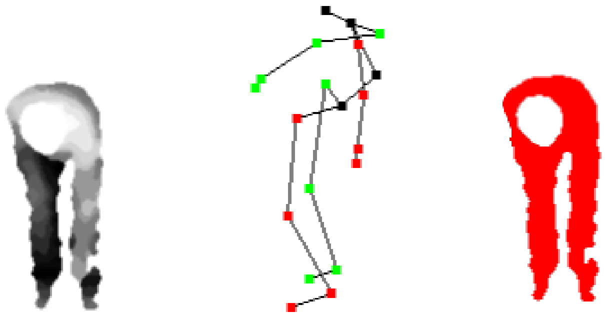


Figure 2: Example of a noisy skeleton (instance from action *Pick-up and throw* from the MSR Action3D dataset [11]

useful for action recognition [16]. Relying on background subtraction or human body detection techniques, the silhouette can be extracted out of RGB images. Different holistic features are then extracted based on the points of the silhouette or its boundary; which are afterwards made translation and scale invariant. Although good results can be obtained for body motion based action recognition, the silhouette fails when more fine-grained motion recognition is required (as for gestures), or when the only available viewpoint has been recorded from an unfavourable viewing angle. In other words, whereas a side view of *kicking* or *punching* can be easily recognised based on the shape of the silhouette, this may be undistinguishable from a front view, not only for a computer vision algorithm, but even for a human.

In conclusion, depending on scenario-related constraints, as viewing angles and the type of actions we want to recognise, either the 3D pose information contained in the skeleton or the shape of the silhouette may contain the most useful information. Furthermore, even if the 3D skeletal information seems to be in a clear advantage, though it may not always be reliable (see, for instance, Figure 2), the shape information provided by the silhouette could provide useful additional characteristic value in order to improve the classification. For this reason, in this proposal, the appropriate combination of skeletal- and silhouette based features is studied considering feature fusion techniques and state-of-the-art feature extraction and learning methods. Note that using the Microsoft Kinect device, it is possible to obtain in real time both the body pose estimation in form of skeletal data, and the silhouette of the *user* relying on depth-based segmentation.

2. Related work

2.1. Based on RGB-D data

The use of the different data provided by the RGB-D devices for human action recognition goes from employing only the depth data, or only the skeleton data extracted from the depth, to the fusion of both the depth and the skeleton data.

Li *et al*. [11] sample representative 3D points extracting the points on the contours of the projections of the 3D depth map onto the three orthogonal Cartesian planes. In order to reduce the size of the feature vector, the method selects a specified number of points at equal distance along the contours of the projections. Ballin *et al*. [3] estimate the 3D optical flow related to the tracked people from point cloud data, summarising it by means of a 3D grid-based descriptor. Wang *et al*. [22] fuse the skeleton information and a local occupancy pattern (LOP) based on the 3D point cloud around each joint. In a different approach, Wang *et al.* [21] treat an action sequence as a 4D shape and propose random occupancy pattern features, which are extracted from randomly sampled 4D sub-volumes with different sizes and at different locations. These features are robust to noise and less sensitive to occlusions. Oreifej and Liu [15] describe the depth sequence using a histogram that captures the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. Yang *et al*. [26] project the depth maps onto three orthogonal planes and accumulate the whole sequence generating a depth motion map (DMM), similar to the motion history images [4]. Histograms of oriented gradients (HOG) are obtained for each DMM. The concatenation of the three HOG represents an action.

Miranda *et al*. [13] describe each pose using a spherical angular representation of the skeleton joints. This way the feature vector is invariant to sensor orientation and global translation of the body, minimising issues with skeleton variations from different individuals. Yang and Tian [25] propose a new type of feature, the EigenJoints. They employ 3D position differences of joints to characterise action information including posture, motion and offset features. Xia *et al*. [24] transform each 3D joint location according to a a modified spherical coordinate system centred in the WAIST. This representation achieves view invariance by aligning the modified spherical coordinates with the person's specific reference direction. This work reduces the number of considered joints from 20 to 12, removing some that are close between them. Azary and Savakis [2] use sparse representations of spatio-temporal kinematic joint features and raw depth features, which are invariant to scale and position.

2.2. Based on human silhouettes

Silhouettes have been used extensively for human action recognition. Using a low resolution, this bidimensional data is sometimes employed as it is, for instance, relying on a volume of silhouettes as spatio-temporal feature. These binary masks can then be reduced in dimensionality as in [19], where principal component analysis (PCA) is employed. Lv and Nevatia [12] propose a log-polar histogram which is computed choosing the different radii of the bins based on logarithmic scale. A bag of rectangles has been proposed in [9], where a histogram of oriented rectangular patches is extracted over the whole silhouette. Similarly, in [23], a 3D histogram of oriented gradients (3DHOG) is extracted on densely distributed regions. A very popular feature is the one from Tran and Sorokin [20]. It combines silhouette shape and optical flow in the same feature vector. By means of radial histograms the silhouette shape and the X and Y axis optical flow are encoded and combined with the context of 15 surrounding frames. In [7], the authors present a low-dimensional radial feature that is based on the contour points of the silhouette, and shows suitability for real-time processing. Since the silhouette can be easily obtained as a binary mask using the depth data provided by the Microsoft Kinect device, all these silhouette-based features can also be applied on RGB-D images.

3. Skeletal and silhouette-based features

As has been previously introduced, our proposal consists in combining both skeletal and silhouette-based features in order to improve human action recognition, increasing the robustness to possible body pose estimation errors and taking advantage of the additional discriminative data the silhouette can provide.

In this section, the chosen characteristic features are detailed. These have been selected due to their outstanding performance when used for recognition on their own, and having in mind that they need to be suitable for feature fusion.

3.1. Skeletal feature

Human posture is represented by a skeleton model composed of 20 joints that is provided by the Microsoft Kinect SDK. Each joint is represented by its 3D position in the real world. In order to achieve invariance to scale and rotation, each sequence is normalised following the method in [8]:

1. Determine the normalising length as the distance from the TORSO to the NECK in the first skeleton of the sequence.

2. Determine the y-axis rotation of the first skeleton of the sequence with respect to the Kinect.

3. Set the average location of the TORSO, the LEFTSHOULDER, the RIGHTSHOULDER, the LEFTHIP and the RIGHTHIP as origin coordinate of each skeleton in the sequence.

4. Normalise the size and rotation of all the skeletons of the sequence according to the normalising length and rotation obtained in steps 1 and 2.

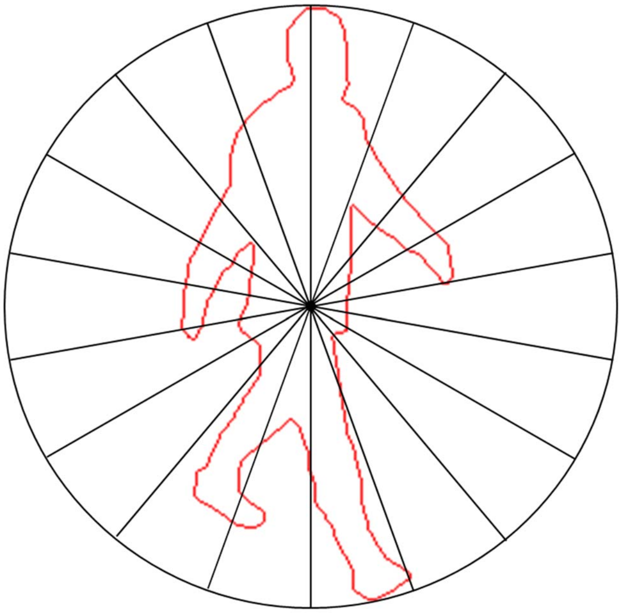


Figure 3: Sample silhouette on which a radial scheme is applied centred in the centroid of the contour points. A summary value is obtained for each radial bin.

3.2. Silhouette-based feature

The shape of the human silhouette is provided by its boundary. Therefore, an efficient feature extraction can be processed when relying only on the contour points of the silhouette. Since low dimensionality is also desired in order to speed up the classification stage and avoid extremely unbalanced feature sizes, we chose the radial silhouette-based feature from Chaaraoui et al. [7]. This feature showed to provide good results and support real-time recognition. It is obtained as follows:

1. First, the centroid of the contour points is obtained.

2. Second, a radial scheme is applied using the centroid as the origin. In this way, the silhouette can be divided in S radial bins of the same angular size.

3. Third, a summary value is obtained for each radial bin. This value is defined as the statistical range of the distances between the points of the bin and the centroid.

4. Fourth, the summary values are concatenated and normalised to unit-sum in order to obtain the feature vector of size S.

Figure 3 shows how the radial scheme is applied to obtain this feature. Please see [7] for greater detail.

3.3. Feature fusion

As has been mentioned before, we propose to combine the skeletal and the silhouette-based information at the feature level. By means of feature fusion, we can retain the different characteristic data enhancing the discriminative information so as to improve the classification.

In this sense, feature concatenation has been employed for feature fusion for two reasons: 1) We want to retain all the characteristic information provided by both feature types; and 2) since we are dealing with very low-dimensional features, the resulting increased size of the final feature is not critical.

4. Classification method based on bag of key poses

As has been seen in the last section, once feature fusion has been applied, for each frame, a pose representation feature is obtained which combines skeletal and silhouette-based features. This bimodal feature is used for classification of human actions.

As learning algorithm we employ a method based on bag of key poses [5, 6]. Similar to the bag-of-words-model, first, a codebook – called bag of key poses – is obtained using the K-means clustering algorithm. In this case, the words are made up of key poses. Then, instead of relying on histograms of word occurrences, sequences of key poses are built. For each training sequence, the pose representation feature of each frame is substituted with its nearest neighbour key pose out of the bag of key poses. In this way, sequences of key poses are obtained which model the possible transitions between key poses.

In the recognition stage, unknown video sequences are classified based on sequence matching. First, an equivalent sequence of key poses is obtained for the test sequence. Then, using the dynamic time warping (DTW) algorithm, the most similar training sequence is found considering a temporally alignment of the involved key poses.

As has been shown in [5, 6], this learning method handles multiple views successfully. In the present work, the skeleton and the silhouette could be considered as different views. Therefore, instead of applying a multi-view fusion of different viewing angles of the same data type, in this case, different data types of the same viewing angle are fused.

5. Experimentation

The proposed method has been evaluated with the MSR Action3D dataset [11]. This dataset contains 20 different actions, performed by ten different subjects and with up to three repetitions making a total of 567 sequences. However, ten sequences are not used because the skeletons were either missing or wrong, as explained by the authors[1]. The authors divided the dataset in three subsets of eight gestures each, as shown in Table 1, and most of the papers working with this dataset have also used them. This was due to the high computational cost of dealing with the complete dataset. The AS1 and AS2 subsets were intended to group actions with similar movement, while AS3 was intended to group complex actions together.

	Dataset					
	AS1		**AS2**		**AS3**	
Feature	K	Rate	K	Rate	K	Rate
Joints' 3D locations	9	88.57%	38	85.71%	6	94.59%
Silhouettes	34	71.43%	41	79.46%	22	85.59%
Fusion	9	92.38%	50	86.61%	25	96.40%

Table 2: Classification rate for each subset of the MSR Action3D dataset (cross-subject validation is used).

Similarly to [11], who first used this dataset, we perform a cross-subject validation, where actors 1, 3, 5, 7 and 9 are used for training, and actors 2, 4, 6, 8 and 10 are used for testing. We have performed an exhaustive study to select the number K of clusters that obtains the best result for each dataset. The parameter S, which determines the number of radial bins of the silhouette-based feature, has been established to a value of 24 bins, resulting in 15° sectors. With this setup, 25 tests have been executed.

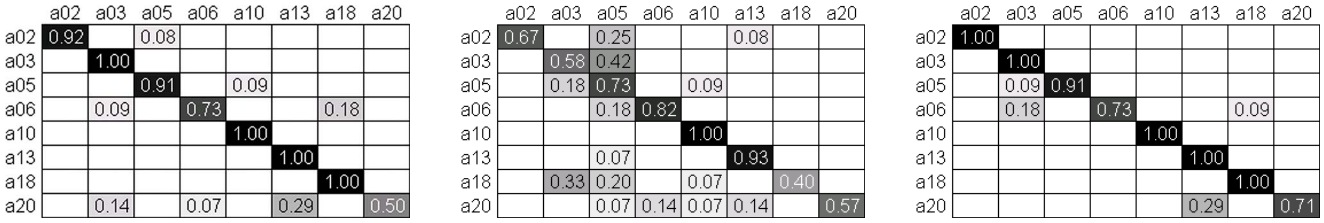
Table 2 shows the best recognition rate obtained using: 1) only the skeleton, 2) only the silhouette, and 3) fusing both of them. The worst results are always obtained using only the silhouettes, as most of the actions included in the dataset do not involve great changes in the 2D projection of the shape of the body. It can be observed that, while good results are obtained using only the joints' 3D locations, the fusion of both features steadily improves the recognition rate.

The success rates for each action in AS1 using the three different alternatives are shown in Figure 4. Despite the fact that the skeletal feature performs considerably better, for some specific actions, the silhouette-based feature obtains a slightly higher recognition rate. In the case of AS1, these actions are *high throw (a06)* and *pick-up and throw (a20)*. In both cases, a greater amount of body motion is involved than in the rest of actions, which explains why the silhouette performs well. At the same time, these actions involve a strongly actor-dependant motion as they can be performed in many ways, and their joint trajectory is certainly complex to track accurately. This could explain the poorer performance of the joints' 3D locations, which performs almost perfect at the other action classes.

[1] MSR Action Recognition Datasets and Codes, http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm (last access: 09/09/2013)

AS1		AS2		AS3	
Label	Action name	Label	Action name	Label	Action name
a02	Horizontal arm wave	a01	High arm wave	a06	High throw
a03	Hammer	a04	Hand catch	a14	Forward kick
a05	Forward punch	a07	Draw cross	a15	Side-kick
a06	High throw	a08	Draw tick	a16	Jogging
a10	Hand clap	a09	Draw circle	a17	Tennis swing
a13	Bend	a11	Two-hand wave	a18	Tennis serve
a18	Tennis serve	a14	Forward kick	a19	Golf swing
a20	Pick-up and throw	a12	Side-boxing	a20	Pick-up and throw

Table 1: Actions in each of the MSR Action3D subsets.



	a02	a03	a05	a06	a10	a13	a18	a20
a02	0.92		0.08					
a03		1.00						
a05			0.91		0.09			
a06		0.09		0.73			0.18	
a10					1.00			
a13						1.00		
a18							1.00	
a20		0.14		0.07		0.29		0.50

(a) Skeleton

	a02	a03	a05	a06	a10	a13	a18	a20
a02	0.67		0.25			0.08		
a03		0.58	0.42					
a05		0.18	0.73		0.09			
a06			0.18	0.82				
a10					1.00			
a13			0.07			0.93		
a18		0.33	0.20		0.07		0.40	
a20			0.07	0.14	0.07	0.14		0.57

(b) Silhouette

	a02	a03	a05	a06	a10	a13	a18	a20
a02	1.00							
a03		1.00						
a05		0.09	0.91					
a06		0.18		0.73			0.09	
a10					1.00			
a13						1.00		
a18							1.00	
a20						0.29		0.71

(c) Fusion

Figure 4: Confusion matrices obtained using only the skeletal feature (a), the silhouette-based feature (b), or the fusion of both (c).

Regarding the confusion matrix obtained for the proposed feature fusion (Figure 4 (c)), it can be observed that in the case of the *high throw (a06)* action, the better performing silhouette-based feature does not improve the combined results. But in the case of the *pick-up and throw (a20)* action, the combination achieves its goal of improving the recognition rate. Also for the *horizontal arm wave (a02)*, the fused feature performs better, showing that we do not require the silhouette-based feature to outperform the skeletal feature in order to provide better discrimination. Moreover, the fusion of features does not lead to a performance loss in any case. This is a desirable behaviour because we do not want to improve the classification of some classes at the cost of obtaining a worse recognition for others.

Table 3 shows a comparison with other methods. Our proposal improves state-of-the-art results for subsets AS2 and AS3, as well as for the overall average. Our results are quite stable in the sense that other methods seem to obtain good results only for specific subsets.

We have repeated similar tests using *leave-one-actor-out* cross validation (LOAO). In this cross validation test, actor-invariance is specifically tested by training with all but one actor, and testing the method with the unseen one. This is repeated for all actors, averaging the returned accuracy scores. Naturally, these rates are normally lower than train and test set validations, as the cross-subject validation, and they provide a more confident result. Furthermore, overfitting can be avoided, since multiple train and tests sets are evaluated. Results are presented in Figure 4.

	Dataset					
	AS1		AS2		AS3	
Feature	K	Rate	K	Rate	K	Rate
Joints' 3D locations	6	88.83%	7	85.01%	9	95.22%
Silhouettes	36	69.46%	35	76.49%	15	81.05%
Fusion	16	90.65%	14	85.15%	21	95.93%

Table 4: Classification rate for each subset of the MSR Action3D dataset (LOAO cross validation is used).

6. Conclusion

In this paper, a proposal for the combination of body pose estimation in form of a skeletal feature and 2D shape based on the silhouette has been presented for human action recognition. Relying on efficient skeletal and silhouette-based features, feature fusion is applied in order to obtain a visual feature with a higher discriminative value and improve human action recognition. A state-of-the-art learning method has been used to validate the proposal. During the experimentation, outstanding results have been obtained both for the cross-subject and the more challenging LOAO cross validation tests. In view of the fact that in every test the fused feature has achieved to improve the recognition

	Dataset			
Method	**AS1**	**AS2**	**AS3**	**Average**
DMM-HOG [26]	**96.2**%	84.1%	94.6%	91.63%
EigenJoints [25]	74.5%	76.1%	**96.4**%	82.33%
OESGP [18]	80.6%	74.9%	87.1%	80.87%
Sparse Repr. (L1-norm) [2]	77.66%	73.17%	91.58%	80.80%
Sparse Repr. (L2-norm) [2]	76.60%	75.61%	89.47%	80.56%
Key poses and Decision Forests [13]	93.5%	52%	95.4%	80.30%
Histograms of 3D Joints [24]	87.98%	85.48%	63.46%	78.97%
Bag of 3D Points [11]	72.9%	71.9%	79.2%	74.67%
Joints' 3D locations	88.57%	85.71%	94.59%	89.62%
Fusion	92.38%	**86.61**%	**96.40**%	**91.80**%

Table 3: Comparison with other state-of-the-art methods (bold indicates highest rate).

rate with respect to the unimodal features, we can confirm that the initial hypothesis has been validated. The highest results so far have been obtained for two of the three subsets of the MSR Action3D dataset, which leads to the best average recognition rate. In this sense, it can be concluded that the shape information contained in the silhouette can provide useful discriminative data, especially when the body pose estimation fails.

This work opens several future lines: On the one hand, in order to validate our proposal, we have employed state-of-the-art visual feature extraction and action learning methods. These could be improved in order to explicitly consider different feature types. More complex skeletal features can be employed based on joint distances, quaternions, etc. Similary, instead of the silhouette, 3D volume information could be considered. On the other hand, the proposed method might benefit from further feature types. Both the employed skeletal and silhouette-based feature contain frame-wise pose or shape information. However, recent works suggest that spatio-temporal features, as shape-motion cues at pixel-level [15], can be very valuable in human action recognition. The addition of further feature types will also lead to reconsider suitable feature fusion techniques.

Last but not least, as it has been seen in the experimentation, the obtained improvement depends on the action classes. Therefore, a selective approach, which decides for each action whether or not different feature types have to be considered, could provide further improvement.

Acknowledgements.

This work has been partially supported by the European Commission under project "caring4U - A study on people activity in private spaces: towards a multisensor network that meets privacy requirements" (PIEF-GA-2010-274649) and by the Spanish Ministry of Science and Innovation under project "Sistema de visión para la monitorización de la actividad de la vida diaria en el hogar" (TIN2010-20510-C04-02). Alexandros Andre Chaaraoui and José Ramón Padilla-López acknowledge financial support by the Conselleria d'Educació, Formació i Ocupació of the Generalitat Valenciana (fellowships ACIF/2011/160 and ACIF/2012/064 respectively). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] M. Alnowami, B. Alnwaimi, F. Tahavori, M. Copland, and K. Wells. A quantitative assessment of using the Kinect for Xbox360 for respiratory surface motion tracking. pages 83161T–83161T–10, 2012. 1

[2] S. Azary and A. Savakis. 3D Action Classification Using Sparse Spatio-temporal Feature Representations. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, and M. Papka, editors, *Advances in Visual Computing*, volume 7432 of *Lecture Notes in Computer Science*, pages 166–175. Springer Berlin / Heidelberg, 2012. 2, 6

[3] G. Ballin, M. Munaro, and E. Menegatti. Human action recognition from rgb-d frames based on real-time 3d optical flow estimation. In A. Chella, R. Pirrone, R. Sorbello, and K. R. Jhannsdttir, editors, *Biologically Inspired Cognitive Architectures 2012*, volume 196 of *Advances in Intelligent Systems and Computing*, pages 65–74. Springer Berlin Heidelberg, 2013. 2

[4] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 2

[5] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. An efficient approach for multi-view human action recognition based on bag-of-key-poses. In A. A. Salah, J. Ruiz-del Solar, C. Meriçli, and P.-Y. Oudeyer, editors, *Human Behavior Understanding*, volume 7559 of *Lecture Notes in*

Computer Science, pages 29–40. Springer Berlin Heidelberg, 2012. 4

[6] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799 – 1807, 2013. Smart Approaches for Human Action Recognition. 4

[7] A. A. Chaaraoui and F. Flórez-Revuelta. Human action recognition optimization based on evolutionary feature subset selection. In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference*, GECCO '13, pages 1229–1236, New York, NY, USA, 2013. ACM. 3

[8] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert Systems with Applications*, 2013. DOI 10.1016/j.eswa.2013.08.009. 3

[9] N. İkizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In A. Elgammal, B. Rosenhahn, and R. Klette, editors, *Human Motion Understanding, Modeling, Capture and Animation*, volume 4814 of *Lecture Notes in Computer Science*, pages 271–284. Springer Berlin / Heidelberg, 2007. 3

[10] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012. 1

[11] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9 –14, june 2010. 2, 4, 6

[12] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, june 2007. 3

[13] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. M. Campos. Real-time gesture recognition from depth data through key poses learning and decision forests. In *Sibgrapi 2012 (XXV Conference on Graphics, Patterns and Images)*, Ouro Preto, MG, august 2012. IEEE. 2, 6

[14] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel. Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 1188 –1193, 28 2012-sept. 1 2012. 1

[15] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition (CVPR*, Portland, USA, 2013. 2, 6

[16] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, June 2010. 2

[17] 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 *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297 –1304, june 2011. 1

[18] H. Soh and Y. Demiris. Iterative temporal learning and prediction with the sparse online echo state gaussian process. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, june 2012. 6

[19] C. Thurau and V. Hlaváč. *n*-grams of action primitives for recognizing human behavior. In W. Kropatsch, M. Kampel, and A. Hanbury, editors, *Computer Analysis of Images and Patterns*, volume 4673 of *Lecture Notes in Computer Science*, pages 93–100. Springer Berlin / Heidelberg, 2007. 2

[20] D. Tran and A. Sorokin. Human activity recognition with metric learning. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision ECCV 2008*, volume 5302 of *Lecture Notes in Computer Science*, pages 548–561. Springer Berlin / Heidelberg, 2008. 3

[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, Lecture Notes in Computer Science, pages 872–885. Springer Berlin Heidelberg, 2012. 2

[22] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, 2012. 1, 2

[23] D. Weinland, M. Özuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 635–648. Springer Berlin / Heidelberg, 2010. 3

[24] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20 –27, june 2012. 2, 6

[25] X. Yang and Y. Tian. EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. In *Second International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR, 2012*, pages 14–19, Providence, Rhode Island, 2012. 2, 6

[26] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In N. Babaguchi, K. Aizawa, J. R. Smith, S. Satoh, T. Plagemann, X.-S. Hua, and R. Yan, editors, *ACM Multimedia*, pages 1057–1060. ACM, 2012. 2, 6