Joint Orientations from Skeleton Data for Human Activity Recognition

Annalisa Franco(✉), Antonio Magnani, and Dario Maio

C.d.L. Ingegneria e Scienze Informatiche,
University of Bologna, Via Sacchi, 3, Cesena, FC, Italy
{annalisa.franco,antonio.magnani,dario.maio}@unibo.it

Abstract. The recognition of activities performed by humans, in a non-intrusive and non-cooperative way, is a very relevant task in the development of Ambient Intelligence applications aimed at improving the quality of life by realizing digital environments that are adaptive, sensitive and reactive to the presence (or absence) of the users and to their behavior. In this paper, we present an activity recognition approach where angle information is used to encode the human body posture, i.e. the relative position of its different parts; such information is extracted from skeleton data (joint orientations), acquired by a well known cost-effective depth sensor (Kinect). The system is evaluated on a well-known dataset (CAD-60 (Cornell Activity Dataset) for comparison with the state of the art; moreover, due to the lack of datasets including skeleton orientations, a new benchmark named OAD (Office Activity Dataset) has been internally acquired and will be released to the scientific community. The tests confirm the efficacy of the proposed model and its feasibility for scenarios of varying complexity.

1 Introduction

Automated high-level human activity analysis and recognition play a fundamental role in many relevant and heterogeneous application fields such as video-surveillance, ambient assisted living, automatic video annotation or human-computer interfaces. Of course different applications need specific approaches to be designed and implemented; general-purpose solutions, though highly desirable, are very difficult to implement due to the differences in the source of information, the requirements in terms of efficiency, the environmental factors which have a significant impact on performance, etc. This work focuses on human activity recognition in indoor environments which has typical applications in fall-detection of elderly people, abnormal human behavior detection or human computer interfaces. In our opinion unobtrusiveness is one of the most important and interesting features of ambient intelligence applications; to meet this requirement, the proposal of this paper is a vision-based technique where simple cameras are used as input devices and the users are not require to wear neither to actively interact with sensors of different nature.

S. Battiato et al. (Eds.): ICIAP 2017, Part I, LNCS 10484, pp. 152–162, 2017.
https://doi.org/10.1007/978-3-319-68560-1_14

With respect to other application scenarios such as video-surveillance, indoor environments offer several advantages: the input data are somehow more "controlled" and easier to process (e.g. to segment the subjects in the scene), the number of possible users is generally limited and input devices, such as RGB-D cameras, can be successfully adopted for data acquisition. The problem of activity recognition is however still complex if we consider that the users are not cooperative and a real-time processing is needed to produce timely and useful information. This paper proposes an activity recognition technique based on the use of RGB-D cameras, and in particular the Kinect sensor, for data acquisition. To the best of our knowledge all the existing techniques based on skeleton data only exploit 3D joint position, while joint orientation is typically neglected. Aim of this work is to evaluate the reliability of the joint orientation estimates provided by Kinect and to verify their effectiveness for action recognition.

The paper is organized as follows: an overview of the state-of-the-art is provided in Sect. 2, Sect. 3 presents the proposed approach, the results of the experimental evaluation are given in Sect. 4 and finally Sect. 5 draws some conclusions an presents possible future research directions.

2 State of the Art

Vision-based activity recognition techniques do not require the use of special devices and the only source of information is represented by cameras placed in the environment which continuously acquire video sequences. Many works adopt common RGB cameras to acquire information from the environment, but undoubtedly the widespread diffusion of low-cost RGB-D sensors, such as the well-known Microsoft Kinect, greatly boosted the research on this topic. Even though a few hybrid approaches combining gray-scale and depth information have been proposed (e.g. [1]), RGB-D sensors alone have been widely used for activity analysis [2] and several benchmarks have been released to facilitate the comparative evaluation of recognition algorithms [3,4]. The most attractive feature of the Kinect sensor is the ability to capture depth images, coupled with the possibility of tracking rather accurately skeletons of individuals in the scene. The skeleton representation provided by Kinect which consists of a set of joints, each described in terms of position and orientation in the 3D space. Such information is extremely useful for human activity analysis as confirmed by many approaches in the literature. A few works exploit only the depth information (and not the skeleton), and typically perform an image segmentation to identify some relevant posture features from the human body [5]. Most of the approaches perform a skeleton analysis, adopting different representations of the set of joints such as the simple joint coordinates, normalized according to some body reference measure [6,7] or joint distances [8], EigenJoints in [9] where PCA is applied to static and dynamic posture features to create a motion model, histograms of 3D joints [10], kinematic features, obtained observing the angles between couples of joints [11], Gaussian Mixture Models representing the 3D positions of skeleton joints [12], Dynamic Bayesian Mixture Model of 3D skeleton features [13]

or spatio-temporal interest points and descriptors derived from the depth image [14]. Another common approach is to adopt a hierarchical representation where an activity is composed of a set of sub-activities, also called *actionlets* [15–18]. Finally a few works also analyze the interaction of humans with objects to obtain a better scene understanding. The authors of [18] adopt a Markov random field where the nodes represent objects and sub-activities, and the edges represent the relationships between object affordances, their relations with sub-activities, and their evolution over time is proposed; in [19] the authors propose a graph-based representation.

3 Proposed Approach

The idea behind the proposed approach is to encode each frame of a video sequence as a set of angles derived from the human skeleton, which summarize the relative positions of the different body parts. This proposal presents some advantages: the use of skeleton data ensures a higher level of privacy for the user with respect to RGB sequences, and the angle information derived from skeletons is intrinsically normalized and independent from the user's physical build. The skeleton information extracted by the Kinect [20] consists of a set of n joints $J = \{j_1, j_2, ..., j_n\}$ where the number n of joints depends on the software used for the skeleton tracking (i.e. typical configurations include 15, 20 or 25 joints). Each joint $j_i = \left(\mathbf{p_i}, \overrightarrow{\mathbf{o_i}}\right)$ is described by its 3D position $\mathbf{p_i}$ and its orientation $\overrightarrow{\mathbf{o_i}}$ with respect to "the world". Our approach exploits the information given by joint orientations to compute relevant angles whose spatio-temporal evolution characterizes an activity. We consider three different families of angles (see Fig. 1a and b):

- θ_{ab}: angle between the orientations $\overrightarrow{\mathbf{o}_a}$ and $\overrightarrow{\mathbf{o}_b}$ of joints j_a and j_b. Angles θ_{ab} are computed for the following set of couples of joints:

$$A_\theta = \{(j_1, j_3), (j_1, j_5), (j_3, j_4), (j_5, j_6), (j_0, j_{11}), (j_0, j_{12}), (j_7, j_8), (j_9, j_{10})\}$$

- φ_{ab}: angle between the orientation $\overrightarrow{\mathbf{o}_a}$ of j_a and the segment $\overrightarrow{j_a j_b}$ connecting j_a to j_b (we can consider the segment as the bone that interconnects the two joints). Angles φ_{ab} are computed for the following set of couples of joints:

$$A_\varphi = \{(j_3, j_1), (j_3, j_4), (j_4, j_3), (j_4, j_{11}), (j_{11}, j_4), (j_5, j_1), (j_5, j_6), (j_6, j_5),$$
$$(j_6, j_{12}), (j_{12}, j_6), (j_2, j_7), (j_7, j_2), (j_7, j_8), (j_2, j_9), (j_9, j_2), (j_9, j_{10})\}$$

- α_{bac}: angle between the segment $\overrightarrow{j_a j_b}$ connecting j_a to j_b and $\overrightarrow{j_a j_c}$ that connects j_a to j_c. Angles α_{abc} are computed for the following triplets of joints:

$$A_\alpha = \{(j_2, j_7, j_8), (j_7, j_8, j_{13}), (j_2, j_9, j_{10}), (j_9, j_{10}, j_{14})\}$$

We consider only subset of the possible angles, mainly obtained from the joints of the upper part of the body, because not all the angles are really informative: for example the angles between head and neck are almost constant over

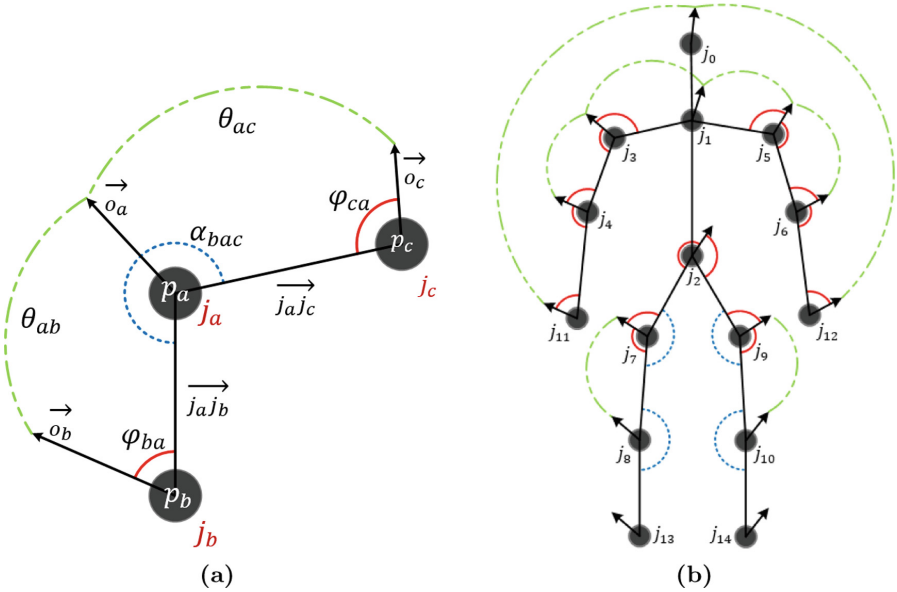


Fig. 1. (a) Representation of a subset of joints $j_a = (p_a, \vec{o_a})$, $j_b = (p_b, \vec{o_b})$ and $j_c = (p_c, \vec{o_c})$ and related angles θ, φ and α. (b) The 28 angles used in our experiments computed from a skeleton configuration with 15 joints.

time and does not provide useful information for activity discrimination. Different configurations of angles have been evaluated and compared in (see Sect. 5). Therefore, each frame f_i of the video sequence $S_i, i = 1, .., l$ is represented by a vector obtained as the ordered concatenation of the values of $\theta_i \mid i \in A_\theta$, $\varphi_j \mid j \in A_\varphi$, $\alpha_k \mid k \in A_\alpha$

$$\mathbf{v}_i = (\theta_1, ..., \theta_m, \varphi_1, ...\varphi_n, \alpha_1, ..., \alpha_s)$$

of size $(m + n + s)$.

It is worth noting that the number of frames for each video sequence can be extremely high and certainly not all the resulting feature vectors are significant: the variation of the angles between two subsequent frames is minimal and usually unnoticeable. We decided therefore to adopt a Bag of Word model [21] with a two-fold objective: minimizing the representation of each sequence keeping only the relevant information and producing fixed-length descriptor which can be used to train an activity classifier. The idea is to represent each activity as an histogram of occurrences of some reference postures (see Fig. 2 for a visual representation), derived from the analysis of the training set. A reference dictionary is first built by applying the K-means clustering algorithm [22] to the set of posture features extracted from the training sequences. Since some subjects could be left-handed, all the angle features are mirrored with respect to the x-axis. We denote with k the number of clusters determined (i.e. the

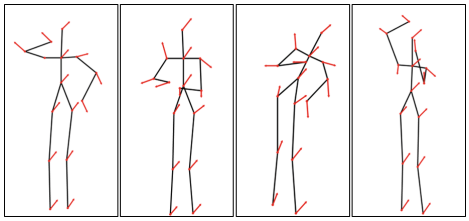


Fig. 2. Visual representation of a subset of key poses corresponding to some cluster centroids of the dictionary W.

dictionary size). The dictionary should encode the basic postures assumed during the different actions in the training set and will be used to represent each sequence as an histogram of occurrences of such basic elements. Given a set of training sequences $TS = \{S_i, i = 1, .., d\}$, representative of the different actions, the k-means clustering algorithm is applied to the associated set of feature vectors $FV = \{\mathbf{v_i}, i = 1, .., d\}$ to obtain a set of k clusters: the cluster centroids are used as words of the reference dictionary $W = \{w_i, i = 1, .., k\}$. The number of clusters k determines the size of the dictionary and is one of the most relevant parameters of the proposed approach. Each sequence is then encoded as a normalized histogram of occurrences of the words in W. Of course the angle features are continuous values and a precise correspondence between the words in the dictionary and the descriptors is very unlikely; therefore when computing the histogram each feature vector f_i is associated to the closest word w_j^* in the dictionary: $j^* = \text{argmin}_j \, ||f_i - w_j||$.

A Random Forest Classifier [23] is trained to discriminate the different activities represented in the training set; the classifier consists of an ensemble of decision trees, each trained on a subset of the patterns and a subset of the features and the final classification is obtained combining the decisions of the single sub-trees.

4 Experiments

Several experiments have been conducted to evaluate the sensitivity of the proposed approach to its main parameters (i.e. the set of angles selected and the dictionary size). Despite of the large number of existing benchmarks for activity recognition from skeleton information, joint orientations are generally not available. We used for testing the well-known CAD-60 [15,24], released by the Cornell University, and a newly acquired dataset. CAD-60 contains 60 RGB-D videos where 4 different subjects (two male and two female, one left-handed) perform 12 daily activities in 5 environments (office, kitchen, bedroom, bathroom and living room). The authors of the benchmark propose two settings named *new person*, where a leave-one-out cross-validation is adopted, and *have seen* where the training set includes data from all the subjects. We adopted the *new person* testing protocol, in accordance with all the related works in the literature, to

allow for a comparison of the results. Moreover, analogously to other works, the recognition accuracy is measured separately for the different rooms.

4.1 Office Activity Dataset (OAD)

Due to the lack of datasets including information on joint orientations, we decided to acquire a new database of human activities to perform further tests. Data acquisition was carried out in a single environment (office) from several perspectives based on the action being performed. From this point of view the benchmark is more complex than CAD-60 because all the activities need to be compared for activity recognition and the higher number of subjects increases the variability of each action. It contains 14 different activities: *drinking, getting up, grabbing an object from the ground, pour a drink, scrolling book pages, sitting, stacking items, take objects from a shelf, talking on the phone, throwing something in the bin, waving hand, wearing coat, working on computer, writing on paper*. Data was collected from 10 different subjects (five males and five females) aged between 20 and 35, one subject left-handed. The volunteers received only basic information (e.g. *"pour yourself a drink"*) in order to be as natural as possible while performing actions. Each subject performs each activity twice, therefore we have collected overall 280 sequences.

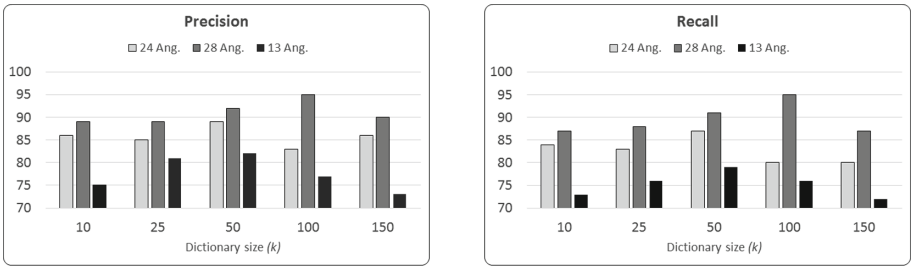
The device used for data acquisition is the Microsoft Kinect V2 whose SDK allows to track 25 different joints (19 of which have their own orientation). For testing, we adopted the same *"new person"* setting of the CAD-60 dataset: a leave-one-out cross-validation with rotation of the test subject. The set of angles used for testing the proposed approach is however the same used for CAD-60. The dataset will be made available online in the Smart City Lab web site (http://smartcity.csr.unibo.it).

4.2 Results

Performance evaluation starts from the analysis of the confusion matrix M where a generic element $M(i, j)$ represents the percentage of patterns of class i classified by the system as belonging to class j. Further synthetic indicators can be derived from the confusion matrix; in particular, we computed precision P and recall R as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}$$

where TP, FP and FN represent respectively the True Positives, False Positives and False Negatives which can be easily derived from the extra-diagonal elements of the confusion matrix. In analogy to the proposal in [8], each video sequence is partitioned into three subsequences which are used independently in the tests. The results obtained are summarized in Fig. 3 where the Precision (P) and Recall (R) values are reported for different experimental settings, i.e. variable dictionary size (k) and three subsets of angles considered for skeleton representation. In particular, the efficacy of the joint orientations is assessed by comparing the



(a) **(b)**

Fig. 3. Precision (a) and recall (b) values on CAD-60 with different configurations of angles, as a function of the dictionary size (k).

results of two different settings - 24 angles, (α angles omitted) and 28 angles - with those obtained using only A_α angles, computed between all the existing pairs of neighboring segments (13 angles, no joint orientation is used in this case). The results show that, overall, the accuracy of the proposed technique is good. As expected the dictionary size has a significant impact on the performance; it is worth noting that different actions have often very similar postures (e.g. drinking and talking on the phone) and a value of k excessively low probably determines the reference posture of such activities to collapse in a single word, thus making difficult to correctly distinguish them. On the other hand, a high value of k produces very sparse feature vectors, more sensitive to the presence of noise. The best results have been reached with a value of $k = 100$ which also allows to efficiently perform the classification task. Also the angle configuration is important; the use of 28 angles produces better results both in terms of precision and recall with respect to the version with 24 angles. The limited accuracy of the configuration with 13 angles, where the orientation is not exploited, confirm the effectiveness of joint orientation for accurate posture representation. These results also show that the significance of the angles varies greatly and a few strategical angles can greatly improve the recognition performance. As to the computational complexity, the proposed approach is very efficient, and all the angle configuration are suitable for a real time processing.

The confusion matrix, reported in Table 1, allows to analyze the main causes of errors. The mismatch occurred are all rather comprehensible since they are related to very similar activities (e.g. cooking-chopping, cooking-stirring). In these cases the skeleton information is probably too synthetic to discriminate the two actions which are very similar in terms of posture. A comparison with the state of the art is provided in Table 2 which summarizes the results published in the benchmark website. Despite of the very good accuracy reached by different approaches in recent years, the proposed approach outperforms existing methods, both in terms of precision and recall.

The results on the Office Activity Dataset are reported in Tables 3 and 4 for the standard configuration with 28 angles and $k = 100$. The overall results confirm that this benchmark is more difficult for several reasons: (i) the activities are

Table 1. Confusion matrix using $k = 100$ words and a configuration of 28 angles on CAD-60.

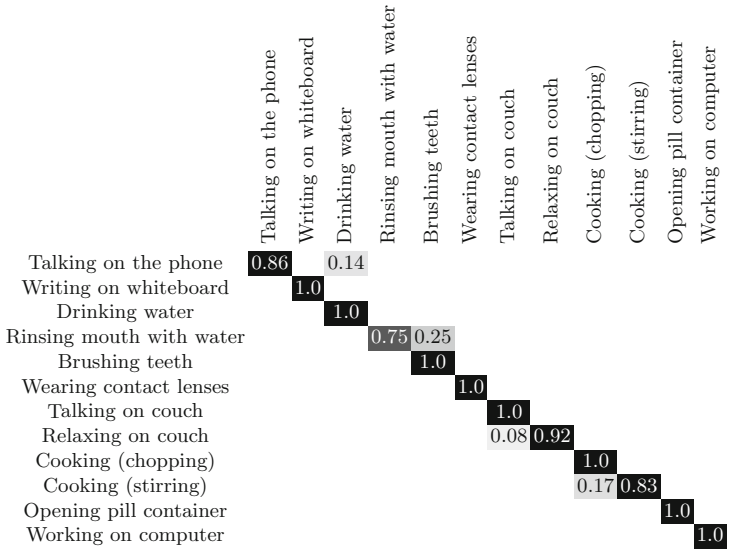


Table 2. Precision (P) and recall (R) of the proposed approach on CAD-60, compared to the results published in the benchmark website. *"*"* indicates that a different protocol was used.

Algorithm	P	R
Proposed approach	**95.0**	**95.0**
Sung et al. [15,24] - 2012	67.9	55.5
Koppula et al. [18] - 2012	80.8	71.4
Zhang and Tian [11] - 2012	86	84
Ni et al. [25] - 2012	Accur: 65.32	-
Yang and Tian [9] - 2013	71.9	66.6
Piyathilaka and Kodagoda [12] - 2013	70*	78*
Ni et al. [1] - 2013	75.9	69.5
Gupta et al. [5] - 2013	78.1	75.4
Wang et al. [17] - 2013	Accur: 74.70	-
Zhu et al. [14] - 2014	93.2	84.6
Faria et al. [13] - 2014	91.1	91.9
Shan and Akella [7] - 2014	93.8	94.5
Gaglio and Lo Re [6] Morana - 2014	77.3	76.7
Parisi et al. [26] - 2015	91.9	90.2
Cippitelli et al. [8] - 2016	93.9	93.5

Table 3. Precision (P) and Recall (R) values of the proposed approach for each activity on OAD.

Action	P	R
Drinking	60.87	77.78
Getting up	81.25	72.22
Grabbing object from ground	83.33	83.33
Pouring a drink	75.00	83.33
Scrolling book pages	80.95	94.44
Sitting	59.09	72.22
Stacking items	90.00	100.00
Taking objects from shelf	100.00	94.44
Talking on phone	86.67	72.22
Throwing something in bin	75.00	33.33
Waving	66.67	66.67
Wearing coat	100.00	100.00
Working on computer	94.12	88.89
Writing on paper	78.95	83.33
Overall	**80.85**	**80.16**

Table 4. Confusion matrix using $k = 100$ words and a configuration of 28 angles on OAD.

	Drinking	Getting up	Grabbing obj.	Pouring a drink	Scrolling book	Sitting	Stacking items	Taking objects	Talking on phone	Throwing something	Waving	Wearing coat	Working on computer	Writing on paper
Drinking	0.78			0.06					0.11		0.06			
Getting up		0.72				0.28								
Grabbing obj.			0.83	0.06		0.06					0.06			
Pour a drink				0.83	0.17									
Scrolling book					0.94						0.06			
Sitting		0.17	0.06			0.72				0.06				
Stacking items							1.0							
Taking objects								0.94		0.06				
Talking on phone	0.17			0.06					0.72		0.06			
Throwing something	0.11		0.11		0.06	0.17	0.11			0.33	0.06			0.06
Waving	0.17			0.11							0.67			0.06
Wearing coat												1.0		
Working on computer													0.89	0.11
Writing on paper	0.06										0.06		0.06	0.83

not partitioned according to the room where they are performed and the probability of misclassification increases; (ii) the number of subjects is higher and the variability in executing the actions increases proportionally. For instance the

worst results have been measured for the activity "throwing something in bin" that the different subjects executed very differently. Other mismatches occur between the activities "sitting" and "getting up"; in principle the reference postures of the two actions are similar, but their temporal ordering in the execution is different and probably the BoW representation adopted is not able to capture this aspect. However in general the good performance of the proposed approach is confirmed on this dataset as well.

5 Conclusions

A human activity recognition technique based on skeleton information has been proposed in this work. In particular, the effectiveness of joint orientations, typically neglected by the works in the literature, has been evaluated on different benchmarks. The efficacy of the proposal have been confirmed; the results obtained overcome the state-of-the-art in the well-known CAD-60 benchmark and good accuracy levels can be reached also on the newly acquired OAD dataset. Future researches will be devoted to the study of techniques able to couple the human posture information (encoded according to the model proposed here) to the information from the surrounding environment (e.g. about interactions with objects or facial expressions) which would certainly increase the performance and enable a fine-grained classification of activities.

References

1. Ni, B., Pei, Y., Moulin, P., Yan, S.: Multilevel depth and image fusion for human activity detection. IEEE Trans. Cybern. **43**(5), 1383–1394 (2013)
2. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans. Cybern. **43**(5), 1318–1334 (2013)
3. Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C.: RGB-D-based action recognition datasets: a survey. Pattern Recogn. **60**, 86–105 (2016)
4. Ni, B., Wang, G., Moulin, P.: RGBD-HuDaAct: a color-depth video database for human daily activity recognition. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision. ACVPR, pp. 193–208. Springer, London (2013). doi:10.1007/978-1-4471-4640-7_10
5. Gupta, R., Chia, A.Y.S., Rajan, D.: Human activities recognition using depth images. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, pp. 283–292. ACM, New York (2013)
6. Gaglio, S., Re, G.L., Morana, M.: Human activity recognition process using 3-D posture data. IEEE Trans. Hum.-Mach. Syst. **45**(5), 586–597 (2015)
7. Shan, J., Akella, S.: 3D human action segmentation and recognition using pose kinetic energy. In: 2014 IEEE International Workshop on Advanced Robotics and its Social Impacts, pp. 69–75 (2014)
8. Cippitelli, E., Gasparrini, S., Gambi, E., Spinsante, S.: A human activity recognition system using skeleton data from RGBD sensors. Intell. Neurosci. **2016**, 21 (2016)
9. Yang, X., Tian, Y.: Effective 3D action recognition using eigenjoints. J. Vis. Commun. Image Represent. **25**(1), 2–11 (2014). Visual Understanding and Applications with RGB-D Cameras

10. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: CVPR Workshops, pp. 20–27. IEEE Computer Society (2012)
11. Zhang, C., Tian, Y.: RGB-D camera-based daily living activity recognition. J. Comput. Vis. Image Process. **2**(4), 12 (2012)
12. Piyathilaka, L., Kodagoda, S.: Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In: 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA), pp. 567–572 (2013)
13. Faria, D.R., Premebida, C., Nunes, U.: A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pp. 732–737 (2014)
14. Zhu, Y., Chen, W., Guo, G.: Evaluating spatio-temporal interest point features for depth-based action recognition. Image Vis. Comput. **32**(8), 453–464 (2014)
15. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from RGBD images. In: 2012 IEEE International Conference on Robotics and Automation, pp. 842–849 (2012)
16. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297 (2012)
17. Wang, J., Liu, Z., Wu, Y.: Learning Actionlet Ensemble for 3D Human Action Recognition. In: Wang, J., Liu, Z., Wu, Y. (eds.) Human Action Recognition with Depth Cameras. SCS, pp. 11–40. Springer, Cham (2014). doi:10.1007/978-3-319-04561-0_2
18. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. CoRR abs/1210.1207 (2012)
19. Koppula, H., Saxena, A.: Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), JMLR Workshop and Conference Proceedings, vol. 28, pp. 792–800 (2013)
20. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
21. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009 - British Machine Vision Conference, London, United Kingdom, pp. 124.1–124.11. BMVA Press (2009)
22. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press Professional Inc., San Diego (1990)
23. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
24. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from RGBD images. In: Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition. AAAIWS 11-16, pp. 47–55. AAAI Press (2011)
25. Ni, B., Moulin, P., Yan, S.: Order-preserving sparse coding for sequence classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 173–187. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33709-3_13
26. Parisi, G.I., Weber, C., Wermter, S.: Self-organizing neural integration of pose-motion features for human action recognition. Front. Neurorobot. **9**(3), 1–14 (2015)