Learning Skeleton Stream Patterns with Slow Feature Analysis for Action Recognition

Yanhu Shan, Zhang Zhang, and Kaiqi Huang(✉)

Institute of Automation, CAS, Beijing, China
{yanhu.shan,zzhang,kqhuang}@nlpr.ia.ac.cn

Abstract. Previous studies on MoCap (Motion Capturing (MoCap) System tracks the key points which are marked with conspicuous color or other materials (such as LED lights). The motion sequences are collected into MoCap action datasets, e.g., 1973 [3] and CMU [4] MoCap action datasets.) action data suggest that skeleton joint streams contain sufficient intrinsic information for understanding human body actions. With the advancement in depth sensors, e.g., Kinect, pose estimation with depth image provides more available realistic skeleton stream data. However, the locations of joints are always unstable due to noises. Moreover, as the estimated skeletons of different persons are not the same, the variance of intra-class is large. In this paper, we first expand the coordinate stream of each joint into multi-order streams by fusing hierarchical global information to improve the stability of joint streams. Then, Slow Feature Analysis is applied to learn the visual pattern of each joint, and the high-level information in the learnt general patterns is encoded into each skeleton to reduce the intra-variance of the skeletons. Temporal pyramid of posture word histograms is used to describe the global temporal information of action sequence. Our approach is verified with Support Vector Machine (SVM) classifier on MSR Action3D dataset, and the experimental results demonstrate that our approach achieves the state-of-the-art level.

Keywords: Action recognition · Skeleton · Joint stream · Multi-order streams · Slow feature analysis

1 Introduction

Recently, human action recognition has been an important domain of computer vision because of its great application prospects in intelligent visual surveillance, human-computer interaction, smart home, etc. Human action can be treated as a 3D space-time volume concatenated by images. Low-level and mid-level features [7][2][5][20] are extracted for action description, and the results on several realistic action datasets [15][6][12] demonstrate its promise. However, the lack of high-level semantic information makes this kind of methods not handle complex actions. Several previous studies [3][1] use skeleton of human body for gesture/action representation, and these work suggest that skeleton provides enough

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part III, LNCS 8927, pp. 111–121, 2015.
DOI: 10.1007/978-3-319-16199-0_8

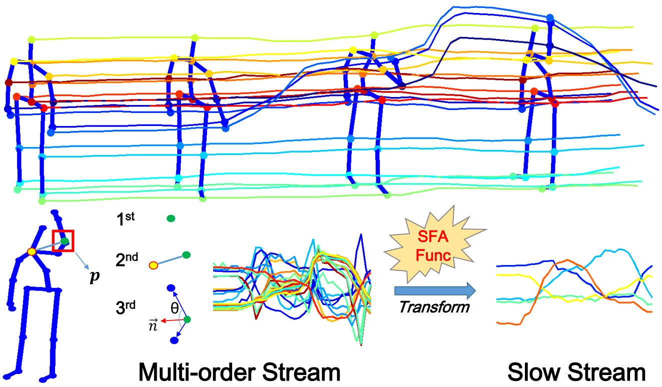


Fig. 1. (Best viewed in color) Kinetic stream pattern learning and transformation. The slow streams are the transformed 5 slowest streams.

information to describe human body actions. However, skeletons estimated from RGB image sequence are not accurate enough for action representation. Fortunately, motion capturing technique [14] can estimate the 3D skeleton joints easily with data from depth camera and help us to avoid the influence of the limits from preprocessing techniques.

Recent studies [21][8] employ both 3D skeleton joints and depth images to improve the capability of feature representation. Skeleton visualization intuitively demonstrates that although the skeleton sequences are unstable, skeletons contain sufficient information of human body actions. Thus, our work focuses on action representation with only skeleton joint streams, i.e., 3D skeleton joint trajectories as shown in Fig. 1.

Earlier work of Campbel and Bolick [1] represents action sequence by projecting pre-existing 3D joints trajectories to curves in subspaces of phase space. The poses in an action form a curve. Although the work can only recognize limited motions with simple descriptors, this work provides a new thinking of action recognition with joint streams. In order to obtain a better action representation, Lv et al. [11] model the dynamics of single joints in the skeleton with Hidden Markov Model (HMM), and the HMM models are combined to form a strong multi-class AdaBoost classifier. This approach can effectively improve the discrimination of action representation in data from Motion Capture (MoCap) system, however, it is still a challenging work to model joint streams with HMM due to lots of noises in the estimation of 3D skeleton sequences. Moreover, modeling a HMM for each class with single joint exists the risk of overfitting when the data volume is small. Zhao et al. [26] learn a vocabulary for each normalized distance stream of a pairwise joints, which reduces the noises in 3D skeleton sequences, and a gesture is represented by combining the corresponding words in different vocabularies. Nevertheless, high-level information is lacked in the feature description.

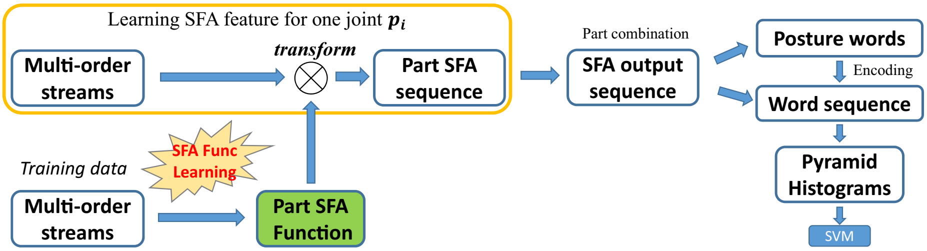


Fig. 2. The framework of our approach

Slow Feature Analysis (SFA) is a method for learning the invariant patterns from visual data. Researches [22] in neuroscience suggest that high-level visual perceptions vary more slowly over time in contrast to input signal. Thus, SFA has been employed in several previous work [25][16] to describe the dynamic of video on realistic datasets.

Inspired by this, we propose an approach to learning high-level patterns from skeleton joint streams with SFA and encoding the high-level information into skeleton postures for action representation. Fig. 2 shows the framework of our approach. Firstly, we construct a multi-order kinetic stream for each key joint by applying the original skeleton sequence. The new streams contain not only the local dynamic of joints but also the dynamic of the center joints relative to others. Secondly, we learn a pattern of each multi-order stream with SFA and encode the high-level dynamic pattern into the original skeleton stream as a part SFA sequence of the joint. The part SFA sequences of all joints are combined into a SFA posture sequence. Then, a dictionary of postures is learned, and the posture sequence is encoded with the posture words in the dictionary. To incorporate the global temporal order information, temporal pyramid is applied for action representation. The classifier we used is SVM.

2 Method

The main stages of our method is threefold: Section 2.1 presents how to generate the multi-order streams from original skeleton sequence data. Then, the SFA is introduced in Section 2.2. In Section 2.3, we propose the method of action representation and classification.

2.1 Multi-order Streams

With the motion capturing technique [14], 20 joints in each frame are estimated from depth video. The position of joint p at frame t has 3 coordinates $p(t) = [x(t), y(t), z(t)]$. The position changes of p over time form a point stream shown in Fig. 1. As the intra-class variance for one action performed by different subject (or even the same subject but different times) is always large and the joint streams contain lots of noises, it is challenging to learn a invariable action pattern. To disperse the intra-variance and avoid the risk of overfitting, we learn an action pattern of each joint in the human body skeleton individually. For one joint p, we extract the high-order streams between p and other joints as well as the first-order position stream of the joint to make sure that the combined stream contain both local and body structural information. The multi-order stream of one joint contains three parts as shown in Fig. 1:

1stOrder : The first-order stream is the split of the original joint shift in 3 coordinate over time, i.e.,

$$s^{1st} = [x(1{:}T), y(1{:}T), z(1{:}T)]^\top, s^{1st} \in R^{3\times T} \tag{1}$$

where

$$x(1{:}T) = [x(1), x(2), ..., x(T)] \tag{2}$$

is the stream in x coordinate, and T is the number of frames in the sequence. Similarly, $y(1{:}T)$ and $z(1{:}T)$ are the streams of the other two coordinates.

2ndOrder : The second-order stream describe the variations of the distances between joint p and the other joints, i.e.,

$$s^{2nd} = \{D_{pq}(1{:}T)\}, q \in \mathcal{J} \mid q \neq p, s^{2nd} \in R^{19\times T} \tag{3}$$

where $D_{pq}(1{:}T)$ is the distance sequence of p and q over time and $\mathcal{J}$ denotes the 20 joint set. The distance sequence of pairwise joints can reflect the dynamic information of joint p relative to others, and the body structural information of human action is naturally encoded into the distance streams. Euclidean distance is used in our work to measure the distance between two joints. For joint p, a 19-d stream is generated as the second-order stream.

3rdOrder : As shown in Fig. 1, the third-order streams are composed of two parts. One is the angle sequence $\theta(1{:}T)$ over time, where θ is the angle of two skeleton segments (segment is the link between two joints in the skeleton) centering on joint p. The other is a sequence formed by the normal vector $\boldsymbol{n_i}$ of the plane determined by the above two segments in frame i. The normal vector sequence $\boldsymbol{n}(1{:}T)$ can be decomposed into 3 streams by considering 3 coordinates of the vector respectively. As these 4 streams are determined by three points, they are combined as the third-order stream and denoted as s^{3rd} in space $R^{4\times T}$.

Streams of all the three orders are combined as the final multi-order stream

$$\mathbf{s} = [s^{1st}, s^{2nd}, s^{3rd}]^\top, s \in R^{26\times T} \tag{4}$$

of joint p. The sequence of each dimension in $\mathbf{s}$ is normalized to zero mean and unit variance to reduce the variances among skeletons of different subjects. Note that some terminal joints connected with only one joint (such as head, hands and feet) are not processed, because there is no angle on these points.

Moreover, the joints shared by more than two segments are employed repeatedly to form different multi-order streams containing different 3rd-order streams. To distinguish the multi-order streams sharing the same joint center, we call the joint p connecting two certain segments a joint unit.

2.2 SFA Function Learning

Slow Feature Analysis (SFA) have been used for learning the invariant patterns from visual data. It can extract high-level visual perceptions vary more slowly over time compared with multi-dimensional input signal and thus can be applied to describe the dynamic changes of human action with temporal sequence. The method is mathematically defined as follows:

Given a multi-dimensional input signal $\mathbf{s}(t)$ from training data, the SFA is to learn a function set $\mathbf{g}(\mathbf{s}) = [g_1(\mathbf{s}), ..., g_M(\mathbf{s})]^\top$ which makes the M-dimensional output $\mathbf{o}(t) = [o_1(t), ..., o_M(t)]^\top$ vary as slow as possible, where $o_j(t) = g_j(\mathbf{s}(t))$. The function learning process can be described as an optimization problem

$$\min_{g_j(t)} \left\langle \dot{o}_j^2 \right\rangle_t \tag{5}$$

subject to

$$\langle o_j \rangle_t = 0 \quad (\textit{zero mean}) \tag{6}$$

$$\langle o_j^2 \rangle_t = 1 \quad (\textit{unit variance}) \tag{7}$$

$$\forall j < j' : \; \langle o_j, o_{j'} \rangle_t = 0 \quad (\textit{decorrelation}), \tag{8}$$

where $\langle \; \rangle_t$ is a mean function over time, $\langle o_j \rangle_t$ and $\dot{o}_j$ are the temporal average and the first order derivative of the j-th dimension signal sequence, respectively. The objective of Eqn.(5) is to minimize the temporal variance measured by the average square of the first order derivative. Eqn.(6) is a normalization for convenience, and Eqn.(7) is to avoid the trivial solution $o_j = \mathit{const}$ which means that the output signal carries no information of changes. The constraint in Eqn.(8) has two roles: ensuring that different dimension output signals carry different types of information and sorting the order of different dimension signals from slowest to fastest.

The transformation function can be unified as

$$g_j(\mathbf{s}) = \mathbf{w}_j^\top \mathbf{h}(\mathbf{s}) = \sum_{k=1}^{K} w_{jk} h_k(\mathbf{s}). \tag{9}$$

When g_j is linear function, $\mathbf{h}(\mathbf{s}) = \mathbf{s}$, and in the case of nonlinear, $\mathbf{h}(\mathbf{s}) = [h_1(\mathbf{s}), ..., h_K(\mathbf{s})]^\top$ is a set of polynomial (usually quadratic) expansion functions for linearization. Note that $\mathbf{h}(\mathbf{s})$ is centralized by minus $\langle \mathbf{h}(\mathbf{s}) \rangle_t$, i.e., $\mathbf{h}(\mathbf{s}) = \mathbf{h}(\mathbf{s}) - \langle \mathbf{h}(\mathbf{s}) \rangle_t$. Thus, the objective function of Eqn.(5) can be rewritten as

$$\left\langle \dot{o}_j^2 \right\rangle_t = \mathbf{w}_j^\top \left\langle \dot{\mathbf{h}}(\mathbf{s}) \dot{\mathbf{h}}(\mathbf{s})^\top \right\rangle_t \mathbf{w}_j = \mathbf{w}_j^\top \mathbf{A} \mathbf{w}_j, \tag{10}$$

and

$$\langle o_j, o_{j'} \rangle_t = \mathbf{w}_j^\top \left\langle \mathbf{h}(\mathbf{s})\mathbf{h}(\mathbf{s})^\top \right\rangle_t \mathbf{w}_{j'} = \mathbf{w}_j^\top \mathbf{B} \mathbf{w}_{j'}. \tag{11}$$

Considering constraint in Eqn.(7), the objective function can be evolved into

$$\left\langle \dot{o}_j^2 \right\rangle_t = \frac{\left\langle \dot{o}_j^2 \right\rangle_t}{\langle o_j, o_j \rangle_t} = \frac{\mathbf{w}_j^\top \mathbf{A} \mathbf{w}_j}{\mathbf{w}_j^\top \mathbf{B} \mathbf{w}_j}. \tag{12}$$

The optimization problem can be solved by the generalized eigenvalue approach

$$\mathbf{AW} = \mathbf{BW\Lambda} \tag{13}$$

The eigenvectors $[\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M]$ corresponding to the M smallest eigenvalues sorted in ascending order are the weights of SFA function $\mathbf{g}(\mathbf{s}) = [g_1(\mathbf{s}), g_2(\mathbf{s}), ..., g_M(\mathbf{s})]$ in Eqn.(9).

2.3 Action Representation and Classification

For each joint unit, we can learn a set of SFA functions, and the SFA function sets of all joint units are combined as $\mathbf{G} = [\mathbf{g}^1, \mathbf{g}^2, ..., \mathbf{g}^N]^\top$, where $\mathbf{g}^i$ is the SFA function set of joint units i. With the learnt SFA function $\mathbf{G}$, the multi-order stream $\mathbf{s}^i$ of joint i can be transformed into a new slow stream feature $\hat{\mathbf{s}}^i = \mathbf{g}^i(\mathbf{s}^i)$ with the size of $M \times T$. Combination $\mathbf{S}(t) = [\hat{\mathbf{s}}^1(t), \hat{\mathbf{s}}^2(t), ..., \hat{\mathbf{s}}^N(t)]^\top$ is used as a stable expression of action sequence. The dimension of $\mathbf{S}(t)$ is $d' = M \times N$.

A posture can be described with the d' dimension vector at the corresponding time/frame, and each action is a posture sequence over time. Although the sequence contains lots of postures, some postures in a short time are very similar. Moreover, several actions share many postures. To describe the action sequence robustly with some key postures, we quantize the postures by clustering the observed posture vectors into a posture dictionary. K-means is employed here to cluster K centers as posture words, and 1-NN is used to label observational vectors with the posture words. Thus, each action can be transformed as a sequence of posture words corresponding to the observational postures.

As known that temporal information is very important for action representation. In order to encode the temporal information of one action into the final action descriptor, we apply a three-tier temporal pyramid with partitions 4×1, 2×1 and 1×1. For each subregions, we count the numbers of different posture words to obtain a histogram, then, the histograms generated from all 7 subregions are concatenated as the final action representation.

Multi-class Support Vector Machine (SVM) with RBF kernel is utilized for action classification. Parameter cost term and kernel bandwidth are optimized using a greedy search with a 5-fold cross-validation on the training data.

3 Experimental Results and Analysis

In this section, we show the verification of our approach on the public MSR Action3D dataset [8], and the experimental results demonstrate that the proposed method can achieve the state-of-the-art level.

MSR Action3D dataset contains 20 actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw*. Each action was performed by 10 subjects for two or three times. The dataset contains 557 action samples, and the frame rate of sequences is 15 f/s. The original action data of this dataset consists of depth image sequences. 20 3D skeleton joint positions are estimated from each depth image by applying the real time skeleton tracking technique [14].

Due to the large amount of computation for classifying all the actions, the dataset is divided into 3 subsets: AS1, AS2 and AS3, and each subset contains 8 actions. The partition follows the rule that AS1 and AS2 group actions with similar movement, while AS3 groups complex actions together. The actions in the three subsets are:

AS1: *Horizontal arm wave, Hammer, Forward punch, High throw, Hand clap, Bend, Tennis serve, Pickup & throw*

AS2: *High arm wave, Hand catch, Draw x, Draw tick, Draw circle, Two hand wave, Slide boxing, Forward kick*

AS3: *High throw, Forward kick, Side kick, Jogging, Tennis swing, Tennis serve, Golf swing, Pickup & throw*

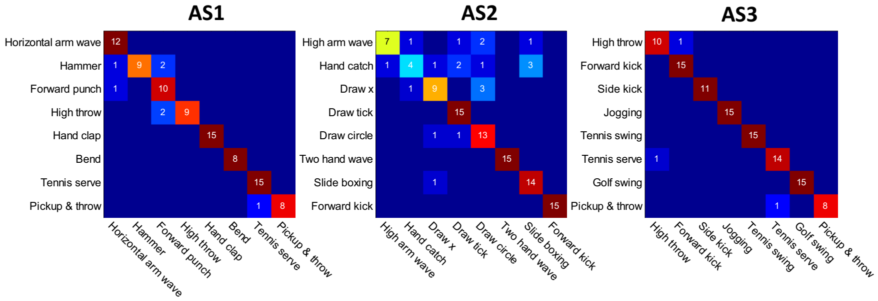
We evaluate our method on MSR Action3D dataset by using the 2-fold cross-subject test setting following the benchmark system [8], i.e., subjects 1,3,5,7,9 are used for training and 2,4,6,8,10 are used for testing. We do not compare with some methods [9,10] which just simply split data into two parts, because the performance of various 2-fold divisions vary widely. As mentioned in Section 2.1, some joint units share one joint center. The number of joint units can be confirmed by the existing angles between two connected segments in the skeleton. The joint of shoulder center connects with 4 segments, and we use the smallest 4 angles in the 3D space to form joint units. Removing the angles connected to hands and feet, the remaining 16 angles in the skeleton are used in our experiments. Thus, N mentioned in Section 2.3 is equal to 16. The number of SFA function M is empirically set to 15. We use the quadratic expansion function $h(d) = [d_1, d_2, ..., d_n, d_1 d_1, d_1 d_2, ..., d_n d_n]^\top$ to expand the d-dimension of stream feature $\mathbf{s}$ in Eqn.(9).

Table 1 compares the results between our approach and the state-of-the-art methods published in recent years. We can see from the results that our method achieves good performance on the three action sets of MSR Action3D dataset, and the average accuracy outperforms those of state-of-the-art methods.

Fig. 3 shows the confusion matrixes for the three action scenes of MSR Action3D dataset. The method works well on AS1 and AS3 action sets while the performance is relatively poor on AS2 set. Contrasting the actions in the three sets, some actions in AS2, e.g. {*High arm wave, Hand catch, Slide boxing*} and {*Draw X, Draw circle*}, are more similar than others, moreover, these actions are always with short durations where the high-level visual patterns are hard to learn by SFA. The highest performance on complex AS3 action set demonstrates that

Table 1. Result comparison with other published methods. 'D' and 'S' in the 'Data' column represent depth and skeleton information respectively.

			Accuracy (%)			
Method	Year	Data	AS1	AS2	AS3	Average
Li et al. [8]	2010	D	72.9	71.9	79.2	74.7
Xia et al. [23]	2012	D	87.98	**85.48**	63.46	78.97
Yang et al. [24]	2012	D	74.5	76.1	96.4	82.33
Vieira et al. [18]	2012	D	84.70	81.30	88.40	84.8
Wang et al. [21]	2012	D + S	-	-	-	88.20
Zhao et al. [26]	2013	S	-	-	-	81.70
Oreifej et al. [13]	2013	D	-	-	-	88.36
Wang et al. [19]	2013	S	-	-	-	90.22
Vemulapalli et al. [17]	2014	S	-	-	-	89.48
Our method		S	**92.47**	82.14	**97.17**	**90.59**



(a) (b) (c)

Fig. 3. The confusion matrixes for the three action sets of MSR Action3D dataset

the learnt high-level patterns by SFA contribute to recognizing complex human actions.

In order to have a deep insight of the proposed approach, comparison experiments are done to analyze the contributions of multi-order stream and SFA transformation, and the results are laid out in Table 2. It's obvious that high-order (2^{nd} and 3^{rd}) streams contain more action information, and multi-order streams can improve the performance of recognition accuracies on all subsets. Moreover, the description capability of each type of streams can be improved with SFA transformation. Combined with the slow streams shown in Fig. 1, we can know that SFA can be used to learn the intrinsic information from streams with noise, and the transformed streams are more stable. Note that the SFA function number M is empirically set by considering all actions in the three subsets, thus, some learned relative faster functions will influent the pow of SFA stream feature and make the performance reduce in some subsets.

Table 2. Result comparison of stream orders and SFA transformation

	Accuracy (%)			
Method	AS1	AS2	AS3	Average
1^{st} Order	59.14	61.61	76.42	65.72
2^{nd} Order	75.26	62.50	87.74	75.17
3^{rd} Order	74.19	64.29	87.74	75.41
Multi-Order	83.49	74.54	90.29	82.78
1^{st} Order + SFA	64.52	60.71	73.58	66.27
2^{nd} Order + SFA	69.89	72.32	88.68	76.96
3^{rd} Order + SFA	78.49	68.75	90.57	79.27
Multi-Order + SFA	**92.47**	**82.14**	**97.17**	**90.59**

4 Conclusion

This paper has proposed an approach to recognize human actions with skeleton joint streams. We generate multi-order streams from original data to improve the description capability of skeleton joint streams. Then, the SFA is employed to decrease the intra-variance of data. Temporal pyramid of posture word histograms is used to describe the global temporal information of action sequence. The experimental results demonstrate that both multi-order streams and the SFA contribute to the recognition accuracy. Compared to the state-of-the-art methods, the part of action representation in our system has big room of improvement by employing discriminative information of action sequence. Moreover, online action recognition with skeleton point information will be more widely used in applications such as human-computer interaction, entertainment or even robot control. Thus, we can optimize our method from these aspects.

Acknowledgments. This work is supported by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61105009 and Grant No. 61175007).

References

1. Campbell, L., Bobick, A.: Recognition of human body motion using phase space constraints. In: Proceedings of the Fifth International Conference on Computer Vision, 1995, pp. 624–630 (1995)
2. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72 (2005)

3. Gunnar, J.: Discriminative video pattern search for efficient action detection. Perception and Psychophysics **14**(2), 201–211 (1973)
4. Han, L., Wu, X., Liang, W., Hou, G., Jia, Y.: Discriminative human action recognition in the learned hierarchical manifold space. Image and Vision Computing **28**(5), 836–849 (2010)
5. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2046–2053 (2010)
6. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proc. Int. Conf. Comput. Vis. (November 2011)
7. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003, vol. 1, pp. 432–439 (2003)
8. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9–14 (2010)
9. Lu, C., Jia, J., Tang, C.K.: Range-sample depth feature for action recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
10. Luo, J., Wang, W., Qi, H.: Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1809–1816. IEEE (2013)
11. Lv, F., Nevatia, R.: Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 359–372. Springer, Heidelberg (2006)
12. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: Proc. Conf. Comput. Vis. Pattern Recognit. (June 2009)
13. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 716–723 (2013)
14. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1297–1304 (2011)
15. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. CoRR abs/1212.0402 (2012)
16. Theriault, C., Thome, N., Cord, M.: Dynamic scene classification: Learning motion descriptors with slow features analysis. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2603–2610 (2013)
17. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
18. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.M.: STOP: space-time occupancy patterns for 3D action recognition from depth map sequences. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 252–259. Springer, Heidelberg (2012)
19. Wang, C., Wang, Y., Yuille, A.: An approach to pose-based action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 915–922 (2013)

20. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, March 2013
21. 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 (CVPR), pp. 1290–1297 (2012)
22. Wiskott, L., Sejnowski, T.: Slow feature analysis: Unsupervised learning of invariances. Neural Computation **14**(4), 715–770 (2002)
23. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27 (2012)
24. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 14–19 (2012)
25. Zhang, Z., Tao, D.: Slow feature analysis for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3), 436–450 (2012)
26. Zhao, X., Li, X., Pang, C., Zhu, X., Sheng, Q.Z.: Online human gesture recognition from motion data streams. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, pp. 23–32. ACM, New York (2013)