Better Performance in Human Action Recognition from Spatiotemporal Depth Information Features Classification

Naresh Kumar

Abstract The recent revolution of sensor-based depth information opens attracting scope to work for human activity recognition. The activities due to human being can have a great interest in every domain of real life where human is always a major factor. Activity recognition is having a key importance due to its advantages in several domains like surveillance systems at the airport, patient monitoring system, and care of elderly people. The variation in spatial and temporal parameters can present any activity efficiently. In the natural color vision, it is not efficient to give complete information because it represents flatness for every portion of the images. The author proposes the objective of this work to recognize daily life human activities by spatiotemporal depth information. This work is carried out by three phases which comprise preprocessing, feature extraction, and action classification. Actions may be performed by a single person or more than one person at a time. For this purpose, the Kinect sensor is used in the data collection phase. The spatiotemporal depth features are computed for recognition by support vector machine classifier. The research work of this problem is experimented on Intel i5 processor with clock speed 3.1 GHz under the windows 8 environment and processing work is performed by commercial software MATLAB 2015b. There are nine classes of human actions in the database described by RGB-D human activity recognition and video database, Cornell activity datasets, and Berkeley multimodal human action database. The accuracy of nine actions is 90.38%. The research work carried out here proves that using the proposed work, the research community and organizations can get better performance that is tough to achieve through the normal video frames of human activities.

Keywords Human action recognition (HAR) · Principal component analysis (PCA) · Spatiotemporal descriptors · Histogram of gradient (HOG) · Support vector machine (SVM)

N. Kumar (✉)
Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, India
e-mail: atrindma@iitr.ac.in

M. Pant et al. (eds.), *Computational Network Application Tools for Performance Management*, Asset Analytics,
https://doi.org/10.1007/978-981-32-9585-8_5

1 Introduction

Human is centered objective of all the happenings everywhere. The change in every domain creates keen interest for the researchers and common people to interpret the things. The still images are less promising than image sequences to ensure any prediction in the field of security, surveillance, and patient monitoring. In the context Ye et al.[1] presented a broad review of several metrologies from the depth information for human action recognition.

The sounding features of image sequences are to vary with space and time which becomes a key interest of researches. These features are termed as spatiotemporal features which are described by Harris 3D [2] Cuboid, Hessian, HOF [3], etc. Moreover, we introduced depth features [2, 4–6] also as the major part of this work.

1.1 Depth Map Information

Depth maps are the image frames which have per pixel depth information extracted from natural images; i.e., the pixel value of the object is varying due to the lack of flatness. The varied pixel information in the different plane of the part of any object is classified by its depth information. The depth maps of the images can be created by taking images by different views and angles which become the source of depth image information [6].

Several depth sensors are available in the market to capture depth images, e.g., Kinect SDK launched by Microsoft in the year 2011. Human activities may be categorized in various sub-categories which include pose estimation, gesture classification, single person, and group activities with any object or communication in the context of social media analytics [7].

1.2 Research Motivation and Challenges

Depth data provides the information with minimum noise which enhances the efficiency of the human action recognition system. Depth map-based recognition [1] of human activities opens a new door for researchers as it is being a novel technique of human activity recognition. Working with space–time interest point (STIP) for feature extraction, it creates a huge amount of multidimensional data which is very hard to process without rich processing devices.

The scope of research work is still open due to the lack of literature on depth map-based processing. Comparatively, high accuracy and cost-effectiveness ensure that depth information-based research is highly ranked. Human activity may involve as a single-person activity like gesture or action performed or two-person interaction, or it may also be a group activity as per the context. Our goal is to recognize all these

Table 1 Depth datasets available on human actions

Dataset	Modality	Metadata	Action class/ Video Seq.	Description
MSR action 3D	Depth map, skeletal joint	10 subjects, 2–3 repetitions	20/567	Game context action, and full body
Chaleara multimodal	RGBD, skeletal	20 subjects, 8–30 repetitions	20/13558	Multiple Indian gestures and ground truth
UTKinect action	RGBD, skeletal	10 subjects, 2 repetitions	10/200	Indoor actions and variant view point
MSR action pair	RGBD, skeletal	10 subjects, 2–3 repetitions	12/360	Action sequences of cues
UCLA multiview actions	RGBD, skeletal	10 subjects, ao 3 view repetitions	10/1475	Indoor home activity
NTU action dataset	RGBD, Skeletal, Infrared	40 subjects, 80 views	60/56880	Challenging data
CAD-60	RGBD Videos	4 subjects, 5 views	12/16S0	Indoor actions

ongoing activities with the help of depth image frames taken from a depth camera or depth images created by some other means.

The real-life action recognition becomes an emergent challenge in multimodal group communication when the communication is randomly started and ended by the participant in parallel fashion. How to train the machine to determine the various modes of activities by many persons together. This always requires a multimodal feature descriptor to deal with inter-class ambiguity, intra-class variation, and occlusion. In Table 1, a review of available depth datasets is presented.

Reaming text of the work is organized as follows. The related work and system architecture of the proposed work are explained in Sects. 2 and 3, respectively. Experimental evaluations and result discussion are presented in Sects. 4 and 5. Finally, in Sect. 6, the work is concluded with future directions of natural vision limitations to get resolve by deep learning networks.

2 Related Work

Human action recognition attracts researchers till today because of its multimodal complexity due to the several environmental issues. The novel steps in video analytics is feature extraction spatiotemporal interest point (STIP) detector [8] which built a basic building block of interest points by Harris and Forester operators to resolve the occlusion and dynamic cluttered issues of environmental complex background, but this descriptor is not independent of motion, contract, and rotation of image. The spatiotemporal descriptor which is based on 3D gradients [2] optimized the depth parameters of activity analysis.

Key frame-based history of image motion (MHI), energy of image (MEI) [9, 10], for temporal segmentation is introduced which invariant of linear motion for achieving a benchmark for activities in natural vision. Many of the feature extraction techniques for HAR based on depth sequences are new which still require to establish a state of art. Depth map-based activities are resolved by the bag of 3D points which modeled view variant action graph of the human body [2, 11].

Local binary pattern features based on three-view fusion of depth motion map [12] is used recognition the human action. In [2] three views of motion characteristics of actions were used by DMMs and then *l1*-norm based regularisation method is used for classification which gives 90.5% accuracy with efficient computation. Similarly, [13] presents the state of art to project DMMs onto three orthogonal planes from HOG normal is computed for action recognition.

A feature descriptors random occupancy pattern (ROP) [14] which is robust to noise and less sensitive to occlusion and treats depth data as 4D volume [14] which is sampled by sparse coding further to improve the robustness. In [15], depth video-based STIP filter DSTIP is developed after that, spatiotemporal depth cuboid similarity feature (DCSF) is established a benchmark to describe for actions class that adaptable size of 3D cuboids. Histogram of normal orientation [16] in 4D space, time, and spatial geometry of joints gave a new benchmark for changing shape and motion, for activity recognition.

The methods applicable in the joint trajectory of depth sequence are developed [17] which accumulates several cues from sparse coding, max polling, etc. Further, it was introduced the pyramid of adaptive spatial–temporal cue that pertains spatial and temporal orders. Space–time features based on depth map treat the depth images as 2D video features. Several key points-based scale and view-invariant feature extraction techniques have been listed which characterized shape, motion parameters [3, 18] for SIFT, HOG, HOF, STIP and [19] introduced a hierarchical kernel descriptor for extracting layer-by-layer feature extraction.

However, the silhouette of depth data encodes 3D shapes and their geometric estimates simultaneously. In [16], it is resolved the issues of motion estimation at the joint by computing histogram-oriented normal (HON4D) treat the depth map as 4D data composed of shape and spatiotemporal information. The deformable changing structure can be captured easily for both space and time parameters.

After the motivation for 3D information on how computer generates spatial information [4] proposed a descriptor robust to occlusions and view and scale invariant based on τ test. Template matching transforms the 3D problem to 2D video sequences [10] which makes the projection in 2D plane. The entire video sequence is projected into orthogonal DMMs planes [13] from which histogram of gradient (HOG) is computed from each of the planes. This approach is view independent, but due to noise and occlusions it is hardly reliable. To address these issues, [11, 14] proposed space–time occupancy patterns (STOP) for action representation in depth maps by partitioning 4D video into 4D cells of space and time information.

3 Proposed Methodology

3.1 Histogram of Depth Map Information

This work is carried out by three phases which comprise preprocessing, feature extraction, and action classification. In the literature, several datasets are available that are presented in Table 1 and the sample the used dataset in this work is presented in Fig. 2a, b. The dataset used in this work is extracted from the standard dataset. In this phase, the features of the image sequence are computed to get classified for the specified objective (Fig. 1).

Feature extraction is the main phase of this work. The region of interest (RoI) is computed for every frame, which is further divided into 8 × 5 grids which is presented by grid image in Fig. 1. The architecture of the proposed human action recognition system is given in Fig. 1. Each pixel p in depth map in represented by a triplet as $p = (x, y, d(x, y))$ where $d(x, y)$ is distance between pixel x and y which is set by Kinect sensor. Let $p = (x, y)$ be a point on the surface. We can compute normal vector N at point p by cross-product of tangent vectors to the plane. Tangent vectors S_x and S_y are given by (3.1) and (3.2) (Fig. 2).

$$S_x = \frac{\partial}{\partial x}\begin{bmatrix} x \\ y \\ d(x, y) \end{bmatrix}; \quad S_y = \frac{\partial}{\partial y}\begin{bmatrix} x \\ y \\ d(x, y) \end{bmatrix} \tag{3.1}$$

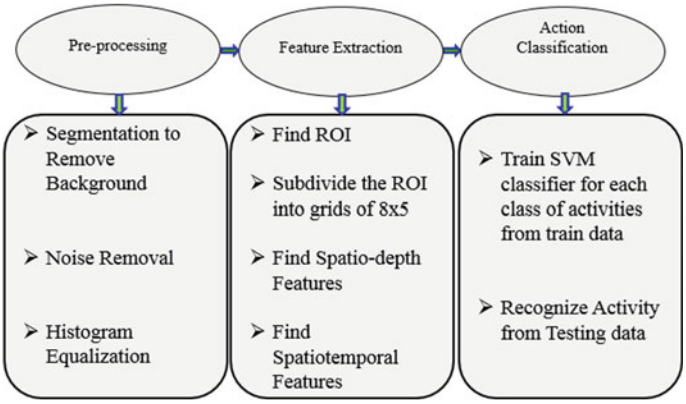
Pre-processing
Feature Extraction
Action Classification
➢ Segmentation to Remove Background
➢ Noise Removal
➢ Histogram Equalization
➢ Find ROI
➢ Subdivide the ROI into grids of 8x5
➢ Find Spatio-depth Features
➢ Find Spatiotemporal Features
➢ Train SVM classifier for each class of activities from train data
➢ Recognize Activity from Testing data

Fig. 1 Human action recognition system architecture flow diagram

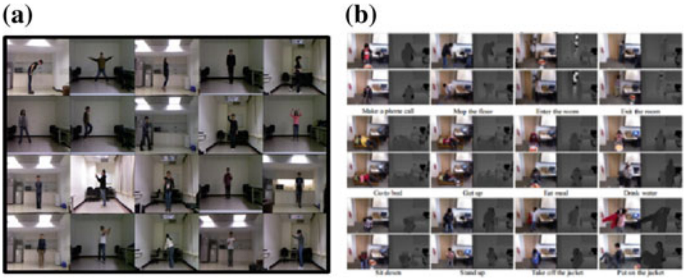

Fig. 2 **a** Multimedia computing laboratory depth HAR dataset (left). **b** Dataset for human activities samples (right)

$$N = S_x \times S_y \quad \text{i.e. } N = \begin{bmatrix} -\frac{\partial(x,y,d(x,y)}{\partial x} \\ -\frac{\partial(x,y,d(x,y)}{\partial y} \\ 1 \end{bmatrix} \tag{3.2}$$

3.1.1 Spatiotemporal Features

The data modality of videos and images demands high-computational resources and storage architectures. One of the feasible efforts [9, 17, 20] in this direction is to use dictionary learning and sparse data which contain salient information. Direct features learning from video data gives the idea of feature descriptors that can extract the features changing with time and space in video analytics. Such features ensure to resolve the specified issues presented in [5, 20–22]. We used dense sampling which can extract the video block of the form (x, y, t, σ, τ), which contains the spatial and temporal information at regular position and scales. In this block, σ and τ denote spatial and temporal parameters. The combination of dense sample features with histogram of oriented optical flow (HOOF) gives a promising result to support this work.

3.2 Feature Classification

To classify the action features, we used nonlinear support vector machine (SVM) for which it is selected Gaussian kernel. The mathematical system of Gaussian kernel for support vector is supported by (3.4) and (3.5).

$$V(H_N, H_d) = \exp\left(-\sum_{c \epsilon C} \frac{1}{M_c} D_c(H_N, H_d)\right) \tag{3.4}$$

$$D_c(H_N, H_d) = \frac{1}{2}\sum_{i=1}^{V}\frac{(h_{\mathrm{Ni}} - h_{\mathrm{di}})^2}{h_{\mathrm{Ni}} + h_{\mathrm{di}}} \tag{3.5}$$

In (3.4) and (3.5), $H_N = \{h_{\mathrm{Ni}}\}$ and $H_d = \{h_{\mathrm{di}}\}$ are histogram of depth feature and dense sampling of optical flow features, respectively. Ac is mean value of the distances between all training samples with vocabulary size V over channel C. Dc in (3.5) is the $\chi 2$ distance between the training feature vectors.

4 Experiments and Evaluations

The research work of this problem is experimented on Intel i5 processor with clock speed 3.1 GHz under the windows 8 environment and processing work is performed by commercial software MATLAB 2015b.

4.1 Data Collection and Preprocessing Phases

The dataset of 20 videos which comprise nine daily life activities is created by Microsoft Kinect SDK XBOX 360 sensor which is easily available at very economical rate. This dataset is created on four to six persons with their different surrounding conditions.

It is indicated by Fig. 3 the environmental background is encoded red for one person in blue, for another person in green. Segmentation is performed to remove the background component from handshaking Fig. 3a, and then we get Fig. 3b which represents the noise-free human body only after a background removal. Morphological operation is followed by background removal to remove the noise with the opening and closing morphological operations. The effect of noise removal can be observed in Figs. 4b, 5, 6 and 7.

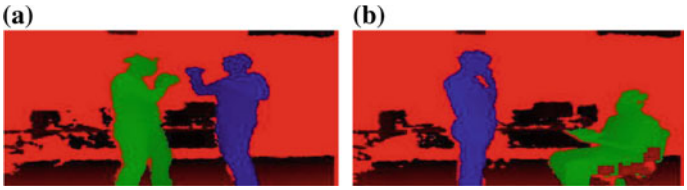

Fig. 3 **a** Both fighting. **b** Sitting and standing

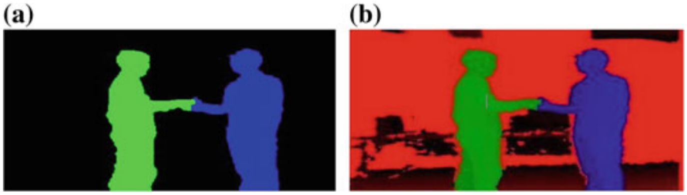

Fig. 4 **a** Original image of handshaking. **b** Noise-free background removed image of Fig. 4a

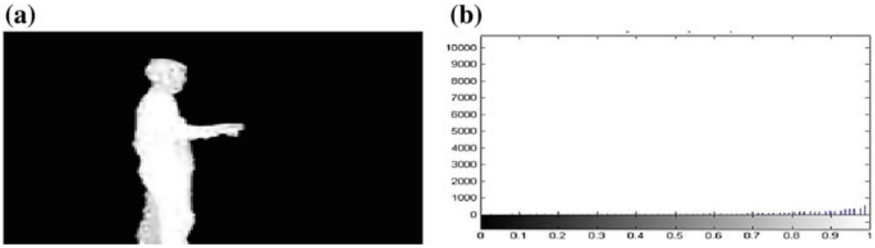

Fig. 5 **a** Depth of green object (left). **b** Feature green histogram (right)

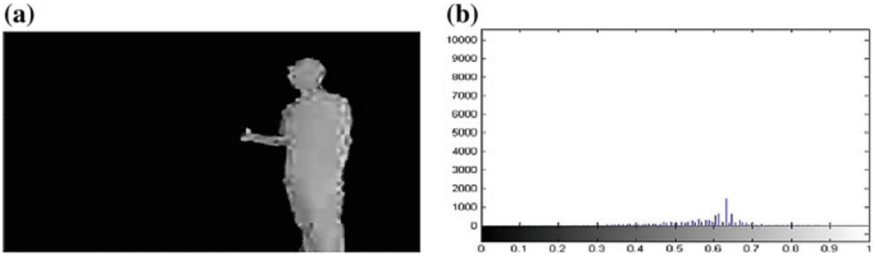

Fig. 6 **a** Depth map blue object (left). **b** Feature blue histogram (right)

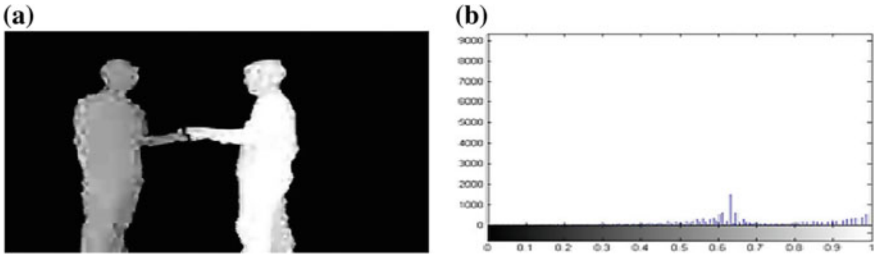

Fig. 7 **a** Grayscale depth object (left). **b** Feature gray histogram (right)

4.2 Feature Extraction for Centroid and Spatiotemporal Points

The enhancement of noise-free objects follows feature extraction phase for spatio-depth and spatiotemporal feature extraction. Region of Interest (ROI) containing the information of actors is segmented into 8 × 5 grids. We take five bins and count the number of pixels in each bin of all the grids. Spatiotemporal features are computed as the difference of no. of pixels representing foreground object in several changing modes. Figure 9b represents the outcome features of spatiotemporal parameters (Fig. 8).

This is computed for all activities, in the successive difference of frames. The computation of the histogram of dense sampling of oriented optical flow (HOOF) features highlights this works as a promising accuracy (Fig. 9).

The metadata of feature sets to get classified by support vector machine is given in Table 2. The computation of feature vector utilized 200 grids for per-pixel information. Again in Tables 3 and 4, it is shown that the results of single-person activity like the action of discussing are very low (60%) since such activities are very much similar to other activities. This results the demand to work more efficient feature computation techniques.

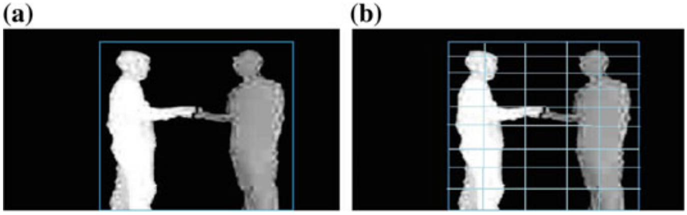

Fig. 8 **a** RoI handshaking image (left). **b** Grid image (right)

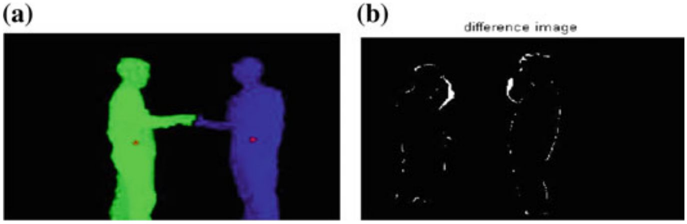

Fig. 9 **a** Centroid feature of blue and green objects (left). **b** Spatiotemporal features points

5 Results and Comparative Analysis

Table 3 presents the confusion matrix of single-person activities such as ‘dozing’, ‘typing’, and reading books, etc. From the diagonal matching values of eight actions, it can be observed that some of activities are very hard to distinguish like discussing (6) in which accuracy is 60%; on the other hand, dozing (34) and typing (36) show success benchmark of our results with 100% accuracy.

Similarly, dense sampling of HOOF features is classified for two-person communication activities as represented in Table 2. The activity of passing bottle by person A to person B drinking is misclassified by (3) with waving for bye vice versa waving

Table 2 Meta data feature vector computation

Features	Values
Avg. depth of blue, green subject	2 (1 each)
Distance between the centroids	1
No. of pixels in foreground	1
Frame differencing	1
ROI grids pixel distribution	200 (8*5*5)

Table 3 Confusion matrix single-person activity

S. No.	Labeled activity class	No. of video frames	Accuracy (%)
1	Dozing	34	**100**
2	Drinking	35	91.4
3	Mobile device wait	30	83.3
4	Talking phone	35	88.5
5	Reading book	34	82.3
6	Streteching	11	81.8
7	Typing	35	**100**
8	Discussing	10	60

Actions class	1	2	3	4	5	6	7	8
1	**34**	0	0	0	0	0	0	0
2	0	32	0	2	0	1	0	0
3	0	2	25	1	2	0	0	0
4	0	3	0	31	1	0	0	0
5	1	3	0	1	28	1	0	0
6	0	2	0	0	0	9	0	0
7	0	0	0	0	0	0	**35**	0
8	0	0	0	0	0	**4**	0	**6**

Table 4 Multiple action frame accuracy and confusion matrix

S. No.	Labeled activity class	No. of video frames	Accuracy (%)
1	Fighting	500	**100**
2	Person passes bottle Person B drinks	385	98.7
3	Waving bye	445	66.3
4	Japanese greetings	500	**100**
5	Handshakes	500	91
6	Person A drops object Person B picks	313	**100**
7	Person A reads Person B stands	301	67.8
8	Sitting and talking	724	100
9	Person A walks to Person B	400	89.7

Actions class	1	2	3	4	5	6	7	8	9
1	**500**	0	0	0	0	0	0	0	0
2	0	**330**	5	0	0	0	0	0	0
3	0	135	**295**	0	0	7	0	0	8
4	0	0	0	**500**	0	0	0	0	0
5	45	0	0	0	**455**	0	0	0	0
6	0	0	0	0	0	**313**	0	0	0
7	72	0	0	1	24	0	**294**	0	0
8	0	0	0	0	0	0	0	**724**	0
9	0	16	2	0	23	0	0	0	**359**

bye action is misclassified by (135) with bottle passing and drinking between two persons A and B (Table 3).

The action of handshaking is misclassified by (45) with fighting. The action of walking person A to B is misclassified with person A walking and person B drinking, waving bye and handshakes by (16), (2), and (23), respectively. Except all these scope of future research for this research our work is achieving novelty to recognize the actions of fighting, person A drops to person B picks and sitting and talking by 100% accuracy (Table 4). As represented in Table 3, ‘dozing’ and ‘typing’ are the most successfully (100%) classified single-person activities. Similarily, the activity ‘discussing’ requires more efforts to recognize as labeled row-8 have only 60% performance. The bold figures in Table 4 denote the better performance of our model for the specified multiple-person activities.

6 Conclusion and Future Directions

The conclusion can be drawn as a result of this research work and evaluations of depth map-based human activities recognition. The rate of recognition is higher than natural color vision-based recognition and computationally efficient for feature extraction. The standard SVM x-fold classifier with Gaussian kernel is used which assumed better than all other Bayesian classifiers, neural network, and decision tree for the feature classification.

The proposed architecture is invariant of lighting and illumination changes of environment. Moreover, we get 3D information which is not affected by varying texture and color. Deep learning features exploration for the group activities in unconstraint environment inspires this research for future directions.

The deep learning techniques are outperforming against highly complex and weakly unsupervised data. Moreover, this is highly stressed to cope up with dimensionality and various complexities of multiple actions in video data by exploiting deep network libraries with the help of recently developed fast optimization technologies efficient.

References

1. M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications* (Springer, Berlin, Heidelberg, 2013), pp. 149–187
2. C. Chen, K. Liu, N. Kehtarnavaz, Real-time human action recognition based on depth motion maps. J. Real-Time Image Proc. **12**(1), 155–163 (2016)
3. D.G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
4. C. Lu, J. Jia, C.K. Tang, Range-sample depth feature for action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 772–779

5. X. Yang, Y. Tian, Super normal vector for human activity recognition with depth cameras. IEEE Trans. Pattern Anal. Mach. Intell. **39**(5), 1028–1039 (2017)
6. C. Chen, R. Jafari, N. Ketharnavaz, A survey of depth and inertial sensor fusion for human action recognition. Multimedia Tools Appl. **76**(3), 4405–4425 (2017)
7. N. Kumar, A scheme of visual object tracking for human activity recognition in social media analytics, *in International Conference on Information, Communication and Computing Technology* (Springer, Singapore, 2017), pp. 194–204
8. I. Laptev, On space-time interest points. Int. J. Comput. Vis. **64**(2–3), 107–123 (2005)
9. G. Somasundaram, A. Cherian, V. Morellas, N. Papanikolopoulos, Action recognition using global spatio-temporal features derived from sparse representations. Comput. Vis. Image Underst. **123**, 1–13 (2014)
10. A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
11. A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, M.F. Campos, Stop: Space-time occupancy patterns for 3D action recognition from depth map sequences, in *Iberoamerican Congress on Pattern Recognition* (Springer, Berlin, Heidelberg, September 2012), pp. 252–259
12. C. Chen, R. Jafari, N. Kehtarnavaz, Action recognition from depth sequences using depth motion maps-based local binary patterns, in *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, New York, January 2015), pp. 1092–1099
13. X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in *Proceedings of the 20th ACM International Conference on Multimedia* (ACM, New York, October 2012), pp. 1057–1060
14. J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3D action recognition with random occupancy patterns, in *Computer Vision–ECCV 2012* (Springer, Berlin, Heidelberg, 2012), pp. 872–885
15. L. Xia, J.K. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 2834–2841
16. O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4D normals for activity recognition from depth sequences, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 716–723
17. X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 804–811
18. I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008* (IEEE, New York, June 2008), pp. 1–8
19. L. Bo, K. Lai, X. Ren, D. Fox, Object recognition with hierarchical kernel descriptors, in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, June 2011), pp. 1729–1736
20. Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, June 2011), pp. 3361–3368
21. A. González, D. Vázquez, S. Ramos, A.M. López, J. Amores, Spatiotemporal stacked sequential learning for pedestrian detection, in *Iberian Conference on Pattern Recognition and Image Analysis* (Springer International Publishing, Berlin, June 2015), pp. 3–12
22. Y. Guo, Y. Li, Z. Shao, RRV: A Spatiotemporal Descriptor for Rigid Body Motion Recognition. arXiv preprint arXiv:1606.05729 (2016)