The Journal of Engineering

The 2nd 2018 Asian Conference on Artificial Intelligence Technology (ACAIT 2018)

Multi-feature consultation model for human action recognition in depth video sequence

eISSN 2051-3305
Received on 18th July 2018
Revised 9th August 2018
Accepted on 9th August 2018
E-First on 25th September 2018
doi: 10.1049/joe.2018.8301
www.ietdl.org

Xueping Liu[1] ✉, *Yibo Li*[2], *Xiaoming Li*[3], *Can Tian*[4], *Yueqi Yang*[4]

[1]*College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, People's Republic of China*
[2]*College of Automation Engineering, Shenyang Aerospace University, Shenyang, People's Republic of China*
[3]*Institute of Information Systems Engineering, Concordia University, Montreal, Canada*
[4]*Faculty of Aerospace Engineering, Shenyang Aerospace University, Shenyang, People's Republic of China*
✉ *E-mail: liuxueping024@hotmail.com*

Abstract: In the field of computer vision research, the research on human action recognition of depth video sequence is an important research direction. Herein, considering the characteristics of temporal and spatial depth video sequence, the authors propose a framework of the consultation model of several action sequence features to solve the classification problem in-depth video sequence. According to the characteristics of the 3D human action space, a variety of action sequence feature data is obtained, and then these data is projected to three coordinate planes, the acquired fusion features are used to train the consultation model, and finally the model is validated through the data. The authors have achieved good results by comparing the two publicly available datasets with the other methods. Experimental results demonstrate that the model performs well in existing identification methods.

1 Introduction

Human action recognition [1, 2] is a very important research direction in the field of computer vision. The research has an important application value in real life, from video surveillance to robot vision, human–computer interaction [3–5] and so on. At present, most of the research takes RGB video data and depth video data [6] as the research object and many methods have been put forward in the action recognition of RGB video [7]. RGB video contains rich colour, edge, and other information, but the accuracy of action recognition algorithm is low due to the existence of complex background, variable light, visual angle change, and so on. In recent years, the research on human action recognition in-depth video [8] has attracted a lot of attention. It is easy to segment objects from the complex background through depth video. Compared with RGB video, segmentation of the human body from depth video can simplify the process of action recognition. For example, the method [9] can quickly and accurately estimate the 3D spatial position of the joints from the depth images provided by Kinect. In addition, depth information can be obtained in real time and has perspective independence.

This paper mainly focuses on human action recognition for depth video data, and scholars have proposed different methods of feature extraction and classification for 3D human action recognition. The developed major feature representation techniques for human action recognition based on depth sequences include a bag of 3D points [10], projected depth maps [11], surface normals [12,13], space–time occupancy patterns [14], skeleton data [15, 16], and spatio-temporal depth cuboid [17]. These above papers include more details of different feature representation techniques based on depth images. In recent years, there are many ways to solve the action recognition of depth video. Vemulapalli *et al.* [18] represented the skeleton configuration and action as points and curves in a Lie group, respectively. Albert Haque *et al.* [19] represented a combination of recurrent and convolutional neural networks with the goal of identifying small, discriminative regions indicative of human identity. Du *et al.* [15] first design a recurrent neural network model, and realise the recognition of depth skeleton data through this model. Wang *et al.* [16] employ a two-stream recurrent neural network architecture to model temporal dynamics and spatial configuration of human action.

In this paper, three methods are used to extract the depth human action video. The features of the motion history image (MHI), motion accumulation image, and motion subtraction image (MSI) are extracted, respectively. Then the three features are projected to *xoy*, *yoz*, and *xoz* coordinate plane, respectively. Their Hu moment [20] features are extracted from the feature of the action image after projection, and then Hu moment features are connected together to form a complete action sequence feature. Then, we build a consultation model, which combines the commonly used basic classifier models (such as KNN, SVM, and GBDT). Then, we classify the features of the action sequence, vote for the results after the classification, and finally, realise the process of complete classification. Fig. 1 is the framework of the model.

The remainder of this paper is organised as follows. In Section 2, we introduce three extraction methods of action sequence feature. In Section 3, we introduce the consultation model. Experimental results and discussion are presented in Section 4. Finally, we conclude the paper in Section 5.

2 Extraction method of action sequence feature

In order to better describe our model, we first describe the extraction method of action sequence feature. We first introduce the feature extraction of three kinds of action sequences in-depth video, and then introduce the projection of the extracted features to three coordinate planes, and introduce the feature extraction method of eigenvector of seven invariant moments.

2.1 Three kinds of extraction feature methods

2.1.1 Motion history image: MHI is used to model the depth action sequence. In the action representation based on the silhouette of human action, MHI has unique advantages, because it takes both spiral and temporal correlation of an action sequence. The MHI image is shown in Fig. 2*a*. To describe a sequence, the MHI can be calculated by a simple replacement and attenuation operation. It is as shown in the following formula:

$$H_\tau(x, y, i) = \begin{cases} t & \text{if } h(x, y, i) = 1 \\ \max\,(0, H_\tau(x, y, i-1) - \delta) & \text{otherwise} \end{cases} \tag{1}$$

Here, (x, y) represents the pixel location, i denotes the time, $h(x, y, i)$ shows the object presence (or motion) in the current video image, the duration τ governs the temporal extent of the movement, and δ is the decay parameter.

2.1.2 Motion subtraction image: The concept of MSI is proposed to capture the 3D structure and depth information. The depth images in the entire depth video sequence are added after subtraction. It is achieved by subtracting a sequence of action video from every two frames and adding the result of the subtraction. At last, the formula (2) is the formula required for calculation, and the MSI image is shown in Fig. 2*b*:

$$\mathrm{MSI} = \sum_{i=2}^{n} \mathrm{image}_i - \mathrm{image}_{i-1} \tag{2}$$

Here, the image_i is the image of ith frame in an action video sequence.

2.1.3 Motion accumulative image: The concept of motion accumulative image (MAI) is proposed to capture the 3D structure and depth information. The depth images in the entire depth video sequence are added. It is implemented to take cumulative action images of all the frames of a sequence of action video. At last, the formula (3) is the formula required for calculation, and the MAI image is shown in Fig. 2*c*:

$$\mathrm{MAI} = \sum_{i=1}^{n} \mathrm{image}_i \tag{3}$$

Here, the image_i is the image of ith frame in an action video sequence.

2.2 Projection of depth image

In reality, seeing the same behaviour from different perspectives will see different effects. This provides a new way of thinking about different people's behaviour from different perspectives. Fig. 3 is the projection of 3D human action to *xoy*, *yoz*, and *xoz*. From Fig. 3, the results of each view are different, and then the features of the corresponding extraction are different. We can get more 2D human action features through 3D human action projection in different directions.

2.3 Seven invariant moments

As the eigenvector of human action, the seven invariant moments have good performance, the seven invariant moments are as follows:

For a $M \times N$ MHI, the $p + q$ order moment m_{pq} is defined as follows:

$$m_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} f(x, y) x^p y^q \tag{4}$$

Among them, $p, q = 0, 1, 2, \ldots$.

The $p + q$ order central moment μ_{pq} is defined as follows:

$$\mu_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} f(x, y)(x - \bar{x})^p (y - \bar{y})^q \tag{5}$$

(x, y) represents the object image point, $(\bar{x} - \bar{y})$ is the object centroid:

$$\bar{x} = \frac{m_{10}}{m_{00}}, \quad \bar{y} = \frac{m_{01}}{m_{00}} \tag{6}$$

Then the normalised central moment of MHI is obtained by normalising the central moments of zero centre moments μ_{00}

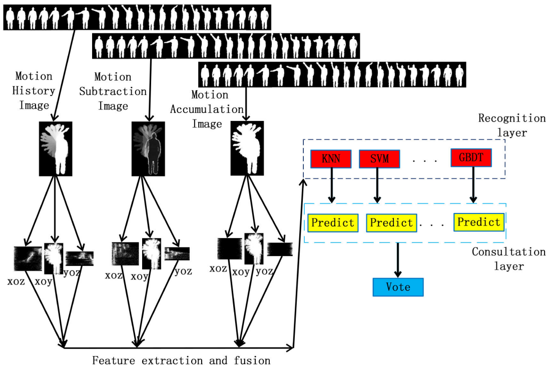

Fig. 1 *Framework of multi feature consultation model*

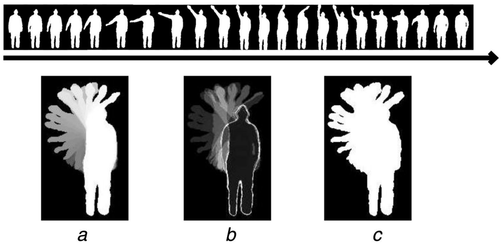

Fig. 2 *Three kinds of extraction feature images*

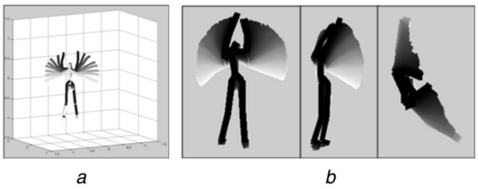

Fig. 3 *Projection of 3D human action image*

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{r}}, r = \frac{p+q+2}{2}, \quad p + q = 2, 3, 4, \ldots \tag{7}$$

The invariant moments are as follows:

$$\begin{cases} v_1 = \eta_{20} + \eta_{02} \\ v_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ v_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ v_4 = (\eta_{30} - \eta_{12})^2 + (\eta_{21} - \eta_{03})^2 \\ v_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} - \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2] \\ +(\eta_{03} - 3\eta_{21})(\eta_{03} + \eta_{21})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{12} - \eta_{03})^2] \\ v_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ +4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ v_7 = (3\eta_{21} - \eta_{03})[(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ -(3\eta_{12} - \eta_{30})(\eta_{03} + \eta_{21})[(\eta_{03} - \eta_{21})^2 - 3(\eta_{12} - \eta_{03})^2] \end{cases} \tag{8}$$

3 Consultation model for human action recognition

3.1 Introduction of basic classifiers in the consultation model

3.1.1 k-Nearest neighbour (KNN) algorithm: KNN algorithm is a supervised learning classification algorithm based on statistical ideas. The working mechanism of this algorithm is given a test sample, based on some distance measurement methods such as generalised Hamming distance, we find out the k training samples which are closest to the training centre, and then make the corresponding prediction based on the information of the k 'neighbours'. In general, the voting method can be used to deal with classification problems, and select the most category labels in the k samples as the prediction results.

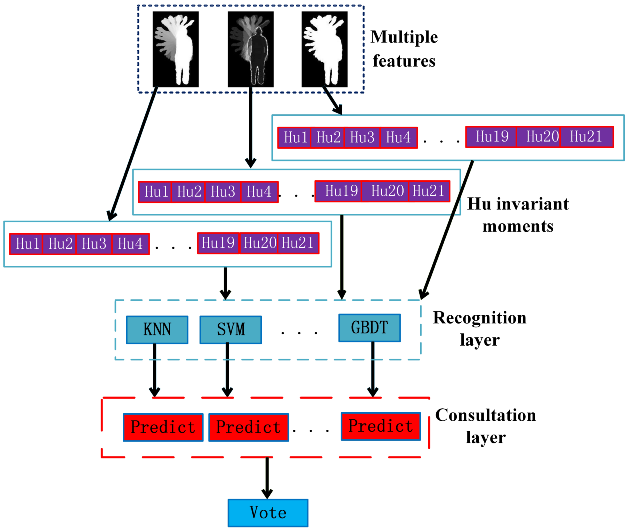

Fig. 4 *Implementation process of the consultation model*

3.1.2 Support vector machine (SVM) algorithm: SVM algorithm is originally designed for two valued classification problems, but in the field of pattern recognition, most of them are multi-classification problems. Therefore, when dealing with pattern recognition problems, it is usually necessary to construct suitable multi-class classifiers. The parameters of different classification planes are solved directly by modifying the objective function parameters. In order to solve this problem, we can combine two classifications into one multi-classifier to realise the multi-classification function of SVM. There are many kernel functions of SVM. In this paper, the radial basis function is selected as the kernel function, and the radial basis kernel function is as follows:

$$k(x, y) = \exp\left(\frac{\| x - y \|^2}{2\sigma^2}\right), \sigma > 0 \tag{9}$$

3.1.3 Classification and regression tree algorithm (CART): CART algorithm is a classical decision tree model which can solve the problem of regression and classification. The model is based on tree structure, which usually contains a root node and multiple internal nodes and leaf nodes. The root node is the complete set of samples, the leaf nodes store the classification results, the intermediate nodes correspond to a property test, and the collection of the samples is divided into the corresponding subnodes according to the test results. The path from root node to the leaf node is the decision process of the decision tree. In training, the sample purity (including the proportion of the same kind of samples) is continuously iterated. The key to the training of a decision tree is how to better divide attributes. There are many ways to select attributes, in which CART models use Gini coefficients to measure node purity. The Gini coefficient can be expressed as

$$\text{Gini}(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{10}$$

The Gini coefficient is the probability of randomly extracting two heterogeneous samples in the sample, which shows that it is inversely proportional to the purity. Therefore, CART algorithm usually divides attributes by reducing Gini coefficients.

3.1.4 Ensemble learning algorithm: Ensemble learning algorithm is a classical method in machine learning theory. Bagging and boosting method are most commonly used. The bagging method, also known as the guide aggregation, is effective only when the potential model can produce different changes, that is, the potential data is introduced to change, and it produces a variety of models with slight changes. The bagging method generates m different datasets and builds a model for each dataset. For classification problems, the final output depends on the voting. Random forest (RF) algorithm is an integrated learning technology formed by improving the decision tree by bagging method.

Boosting method is another widely applied thought in ensemble learning theory. Gradient boosting decision tree (GBDT) algorithm is one of them. GBDT is an additive model using M decision tree as a weak classifier. Its mathematical expression is as follows:

$$f(x) = \omega_0 + \sum_{m=1}^{M} \omega_m \Phi(x) \tag{11}$$

3.2 Introduction of consultation model

In any matter, when we make choices and judgments, we will listen to many opinions, and finally, make judgments. Based on this idea, the concept of ensemble learning method in machine learning theory is introduced, and a machine learning consultation model is established. However, unlike the ensemble learning algorithm, the consultation model integrates the five-strong classifiers based on KNN, SVM, CART, RF, and GBDT to implement the consultation model.

The thought of the consultation model is simple. The idea is that there is a certain number of excellent models, and each model produces little discrimination results on the training set, but the recognition performance of different models for different actions is not the same. So, the output results obtained from multiple models can produce the final recognition results by voting, which is better than the results obtained from single model training. The implementation process of the machine learning consultation model is shown in Fig. 4.

4 Experimental results

In the experiments, we evaluate our model on two datasets: MSR Action3D dataset [10] and DHA dataset [21]. We first describe the datasets and then report our experimental results and analysis.

4.1 Experimental dataset

4.1.1 MSR Action3D dataset: The MSR Action3D dataset is generated by a Microsoft Kinect depth sensor, which is widely used in action recognition. This dataset consists of 567 videos of 20 actions and each action is performed by ten subjects for two or three times. This dataset is challenging for quite similar actions such as ‘draw x’ and ‘draw tick’, both of which have similar movements of hands.

4.1.2 DHA dataset: The dataset is obtained by Kinect, including the depth video sequence and the RGB video sequence. The resolution of the two video sequences is 480 × 640. Although the depth video sequence provided by this dataset has removed the background, it is still challenging, because many behaviours in this dataset are highly similar (such as walking, running, jumping, and skipping). There are 23 kinds of behaviours in this dataset. Each action is recorded by 21 people (12 men, 9 women), 1 time per person, the total is 357 × 2 video sequences.

4.2 Experimental results and analysis

4.2.1 MSR Action3D: In order to verify whether each classifier can accurately judge the similar action, we divided the 20 actions into three subsets, each with eight actions. Table 1 lists the three action subsets used in the experiments. The MS1 and MS2 were intended to group the actions with similar movement, while MS3 was intended to group the complex actions together.

The confusion matrices of human action recognition are illustrated in Fig. 5.

As seen from Fig. 5, almost the most of the actions on MS1 and MS3 have a 100% recognition rate. In MS1, a distinct confusion was noted between the action hammer and forward punch, both tennis serve and pick up and throw in the forward action. This is reasonable as the two actions with identical gestures are similar. In the MS2 subset, the recognition rates of the hand catch, draw x,

Table 1 Three subsets of actions used on the MSR Action-3D dataset

MS1	MS2	MS3
horizontal arm wave	high arm wave	high throw
hammer	hand catch	forward kick
forward punch	draw x	side kick
high throw	draw tick	jogging
hand clap	draw circle	tennis swing
bend	two hand wave	tennis serve
tennis serve	forward kick	golf swing
pick-up and throw	side boxing	pick-up and throw

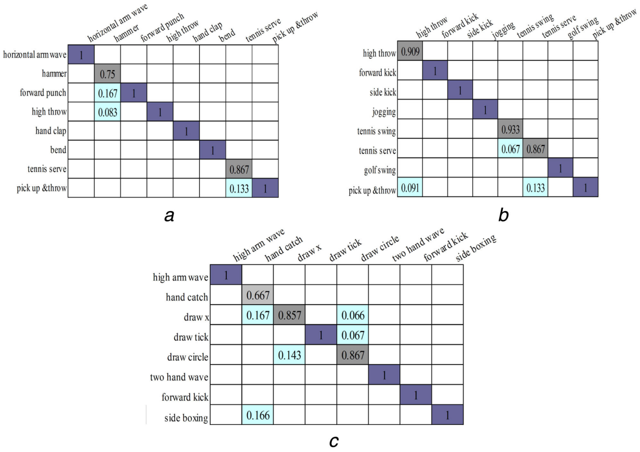

Fig. 5 *Confusion matrices of human action recognition on MSR Action-3D dataset*

Table 2 Experimental results on the MSR Action3D dataset

Methods	Accuracy
Li *et al.* [10]	74.70
Xia *et al.* [17]	89.30
Vemulapalli *et al.* [18]	89.48
Wang *et al.* [22]	91.40
Yang *et al.* [23]	93.45
Gori *et al.* [24]	95.38
our method	94.65

and draw circle were all <90%, because the three actions with small range of action were single-handed movements in the chest; thus, the video sequence of the first and the last few frames were nearly identical, especially for the draw x and draw circle. MS3 revealed a clear recognition error in action tennis serve and pick-up and throw; however, high classification rates (> 90%) have been obtained for 7/8 action classes. We obtain an average accuracy of 94.65%. We compare our model with other methods, and the results are as follows. From Table 2, we can see that our model has achieved good results.

4.2.2 DHA: The DHA action dataset acquired by the Kinect has both visibility and depth video sequences. Although the background of this action dataset was removed, several behaviours continued to have a high similarity, and therefore, the DHA action dataset is challenging. The evaluation setting adopted in DHA dataset is Leave-One-Subject-Out setting [25]. The experimental results are shown in Fig. 6.

As observed, our model recognises 23 actions that achieved an average of 94.17% recognition rate. Moreover, the actions ‘bend,’ ‘jack,’ ‘walk,’ ‘one-hand-wave,’ two-hand-wave,' ‘side-box,’ ‘leg-curl,’ and ‘tai-chi’ have reached 100% recognition rate, indicating that our model has a strong generalisation ability. Similarly, we compare our model with other methods. The results are as follows. From Table 3, we can see that our model has achieved good results.

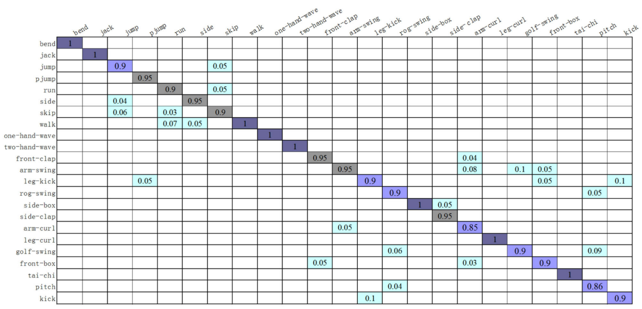

Fig. 6 *Confusion matrix of human action recognition on DHA dataset*

Table 3 Experimental results on the DHA dataset

Methods	Accuracy
Gao *et al.* [26]	86.00
Lin *et al.* [21]	86.80
Liu *et al.* [25]	89.95
our method	94.17

5 Conclusion

In this paper, we propose a framework of multi-feature consultation model to solve the classification problem in-depth video sequence. We first obtain a variety of action sequence feature data and then project these data to three coordinate planes. The fusion features are used to train our consultation model. Experimental results on two publicly available datasets demonstrate the effectiveness of the proposed model. As we analysed on the MSR Action3D and DHA datasets, the similar human actions are very difficult to be distinguished. In the future, we will consider combining more basic classifiers into our consultation model, and we will improve our model to solve the problem of distinguishing similar actions.

6 Acknowledgments

This work was supported by the National Nature Science Foundation of China (nos. 61203170, 61602320) and the Scientific Research General Project of Education Department of Liaoning Province, China (no. L201608).

7 References

[1] Cai, Z., Han, J., Liu, L., *et al.*: ‘RGB-D datasets using microsoft kinect or similar sensors: a survey’, *Multimed. Tools Appl.*, 2017, **76**, (3), pp. 4313–4355

[2] Ramamurthy S, R., Roy, N.: ‘Recent trends in machine learning for human activity recognition – a survey’, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2018, **8**, (1), p. 1254

[3] Haria, A., Subramanian, A., Asokkumar, N., *et al.*: ‘Hand gesture recognition for human computer interaction’, *Procedia Comput. Sci.*, 2017, **115**, pp. 367–374

[4] Guler, A., Kardaris, N., Chandra, S., *et al.*: ‘Human joint angle estimation and gesture recognition for assistive robotic vision’. Computer Vision – ECCV 2016 Workshops, Springer, Cham, 2016, pp. 415–431

[5] Aziz N, N A., Mustafah Y, M., Azman A, W., *et al.*: ‘Features-based moving objects tracking for smart video surveillances: a review’, *Int. J. Artif. Intell. Tools*, 2018, **27**, (02), p. 1830001

[6] Chen, C., Jafari, R., Kehtarnavaz, N.: ‘A survey of depth and inertial sensor fusion for human action recognition’, *Multimed. Tools Appl.*, 2017, **76**, (3), pp. 4405–4425

[7] Herath, S., Harandi, M., Porikli, F.: ‘Going deeper into action recognition: A survey’, *Image and Image Vis. Comput.*, 2017, **60**, pp. 4–21

[8] Gao, Z., Zhang, H., Liu A, A., *et al.*: ‘Human action recognition on depth dataset’, *Neural Comput. Appli.*, 2016, **27**, (7), pp. 2047–2054

[9] Shotton, J., Fitzgibbon, A., Cook, M., *et al.*: ‘Real-time human pose recognition in parts from single depth images’. Computer Vision and Pattern Recognition, 2011, pp. 1297–1304

[10] Li, W., Zhang, Z., Liu, Z.: ‘Action recognition based on a bag of 3D points’. Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 9–14

[11] Chen, C., Liu, K., Kehtarnavaz, N.: ‘Real-time human action recognition based on depth motion maps’, *J. Real-Time Image Process.*, 2016, **12**, (1), pp. 155–163

[12] Oreifej, O., Liu, Z.: ‘HON4D: histogram of oriented 4D normals for activity recognition from depth sequences’. Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 716–723

[13] Yang, X., Tian Y, L.: 'Super normal vector for activity recognition using depth sequences'. Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 804–811

[14] Wang, J., Liu, Z., Chorowski, J., *et al.*: 'Robust 3d action recognition with random occupancy patterns'. Computer Vision – ECCV, Florence, Italy, 2012, pp. 872–885

[15] Du, Y., Wang, W., Wang, L.: 'Hierarchical recurrent neural network for skeleton based action recognition'. Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 1110–1118

[16] Wang, H., Wang, L.: 'Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks'. Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 499–508

[17] Xia, L., Aggarwal J, K.: 'Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera'. Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 2834–2841

[18] Vemulapalli, R., Arrate, F., Chellappa, R.: 'Human action recognition by representing 3D skeletons as points in a lie group'. Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 588–595

[19] Haque, A., Alahi, A., Li, F.F.: 'Recurrent attention models for depth-based person identification'. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1229–1238

[20] Hu, M.K.: 'Visual pattern recognition by moment invariants', *IRE Trans. Information Theory*, 1962, **8**, (2), pp. 179–187

[21] Lin Y, C., Hu M, C., Cheng W, H., *et al.*: 'Human action recognition and retrieval using sole depth information'. Proc. the 20th ACM Int. Conf. Multimedia. ACM, Nara, Japan, 2012, pp. 1053–1056

[22] Wang, C., Flynn, J., Wang, Y., *et al.*: 'Recognizing actions in 3D using action-snippets and activated simplices'. AAAI, Feinikesi, AZ, USA, 2016, pp. 3604–3610

[23] Yang, X., Tian Y, L.: 'Super normal vector for human activity recognition with depth cameras', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **39**, (5), pp. 1028–1039

[24] Gori, I., Aggarwal J, K., Matthies, L., *et al.*: 'Multitype activity recognition in robot-centric scenarios', *IEEE Robot. Autom. Lett.*, 2016, **1**, (1), pp. 593–600

[25] Liu, H., Tian, L., Liu, M., *et al.*: 'Sdm-bsm: A fusing depth scheme for human action recognition', *Image Processing. IEEE*, 2015, pp. 4674–4678

[26] Gao, Z., Song, J., Zhang, H., *et al.*: 'Human action recognition via multi-modality information', *J. Electr. Eng. Technol.*, 2014, **9**, (2), pp. 739–748