IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 4, 2019, accepted March 21, 2019, date of publication March 27, 2019, date of current version April 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907720

Human Action Recognition Using Multilevel Depth Motion Maps

XU WEIYAO[1], WU MUQING[1], ZHAO MIN[1], LIU YIFENG[2], LV BO[2], AND XIA TING[3]

[1]Beijing Laboratory of Advanced Information Networks, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]China Academy of Electronics and Information Technology, Beijing 100041, China
[3]College of Opto-electronic Engineering, Zaozhuang University, Zaozhuang 277160, China

Corresponding authors: Xia Ting (xiayuxue121@126.com) and Xu Weiyao (xuweiyao_2008@126.com)

This work was supported in part by the 111 Project under Grant B17007, and in part by the Director Funds of the Beijing Key Laboratory of Network System Architecture and Convergence under Grant 2017BKL-NSAC-ZJ-01.

ABSTRACT The advent of depth sensors opens up new opportunities for human action recognition by providing depth information. The main purpose of this paper is to present an effective method for human action recognition from depth images. A multilevel frame select sampling (MFSS) method are proposed to generate three levels of temporal samples from the input depth sequences first. Then, the proposed motion and static mapping (MSM) method is used to obtain the representation of MFSS sequences. After that, this paper exploits the block-based LBP feature extraction approach to extract features information from the MSM. Finally, the fisher kernel representation is applied to aggregate the block features, which is then combined with the kernel-based extreme learning machine classifier. The developed framework is evaluated on three public datasets captured by depth cameras. The experimental results demonstrate the great performance compared with the existing approaches.

INDEX TERMS Human action recognition, depth image, ELM classifier, fisher kernel.

I. INTRODUCTION

Human action recognition has become a new research hot topic which integrates computer vision, machine learning and pattern recognition, and has been widely used in virtual reality, intelligent monitoring [1], motion analysis and human-computer interaction [2]. Its main goal is to analyze human activities in video correctly by extracting human motion features. Early research work on human action recognition mainly focuses on RGB video sequences obtained from ordinary cameras. However, human action recognition using RGB images is often disturbed by various lighting conditions, shadows and environmental changes.

Compared with RGB images, depth images are not disturbed by illumination, chroma, shadows and other factors. Even when the light is very dark, high resolution depth images can still be obtained [3]. Fortunately, recent advances in imaging devices, such as Microsoft Kinect, is able to get the depth images and estimate 3D positions easily. Human action recognition based on depth image has attracted wide attention in the past few years [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Jafar A. Alzubi.

Various detection and representation methods, e.g., bag of 3D points [5], skeleton joints [6], depth motion maps (DMMs) [7], have been explored to improve the human action recognition performance using depth images. Recently, Kamel *et al.* [8] propose an action-fusion method for human action recognition from depth maps and posture data using convolutional neural networks. Farooq *et al.* [9] calculate the body part of the action (BPoA) by bounding box with an optimal window size for each DMM to get the action recognition. Cui *et al.* [10] propose a skeleton-based end-to-end model that can simultaneously implement both person identification and action recognition and strengthen the learning of hard samples. Ding et al. [11] propose the Spatio-Temporal Feature Chain (STFC) to recognize human actions from sequences of 3D joint positions. Rahmani and Bennamoun [12] propose a deep model which efficiently models human-object interactions and intra-class variations under viewpoint changes. Kerola *et al.* [13] propose a framework which leverages a novel graph representation of an action as a temporal sequence of graphs. However, most methods are based on the whole depth sequences, which may lose the time and detail information of human action recognition. In addition, actions performed with different

speeds may result in descriptors differently. These challenges lead to increase the intra-class variations, and reduce the recognition accuracy.

In this paper, we focus on recognizing human actions using depth images. In order to make the method robust to different execution rates of each action, the MFSS-MSM method is proposed. The entire input depth map sequences are sampled by calculating key frames according to the proposed MFSS strategy. And then, multiple frame sequences of different scales are generated. However, detailed temporal motion may not be captured using the entire depth sequences, and previous motion may be covered. To overcome the shortcoming, the depth sequences is divided into many sets of depth segments with a fixed length of N. Then we use the proposed MSM model to represent the motion and static information in three-dimensional space. Different with DMM model [7], the proposed MSM model can obtain both motion information and static information in the depth sequences. At last, the local binary pattern (LBP) [14] is adopted to represent the human action, which is a effective texture and powerful descriptor.

The key contributions of this work are as follows:

- The MFSS model based on frame selection strategy is proposed, and generates multiple frame sequences of different scales. To avoid covering motion information, these frame sequences are divided into many sets of frame segments.
- A novel MSM model is proposed by projecting the depth images onto three planes, and generates the motion and static information.
- By dividing all MSM into dense blocks, local rotation invariant texture information in these blocks is characterized by local binary pattern (LBP). Then the extracted feature information is encoded by Fisher [15] and classified by extreme learning machine (ELM) [16].
- The proposed method in this paper has been evaluated on three public datasets, and a comprehensive comparison is provided with the state-of-the-art methods.

The remainder of this paper is organized as follows. In Section II, related works are reviewed briefly. Section III presents the details of our proposed MFSS-MSM method. Experimental results on three datasets are given in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK

With the popularization of low-cost depth cameras, such as Microsoft Kinect, more and more research on motion recognition based on depth maps and skeleton joints has been carried out. According to the features extracted for action recognition, these methods based on depth information can be roughly categorized into three categories: skeleton-based, depth image-based and fusion of different features-based. This section first reviews these three methods briefly, and then explains the motivation of this paper.

For feature extraction based on skeleton features, the existing skeleton-based methods can be roughly divided into joint-based and body-based methods. In [6], Smedt *et al.* present 3D Hand gestures as a set of trajectories of relevant joints of hand-parts in the Euclidean space. In the work of [17], Liu et al. present an enhanced skeleton visualization method for view invariant human action recognition. Du *et al.* [18] propose an end-to-end hierarchical RNN for skeleton based action recognition. Zhu *et al.* [19] propose an end-to-end fully connected deep LSTM network for skeleton based action recognition. Chao *et al.* [20] propose a novel convolutional neural networks (CNN) based framework for both action classification and detection. However, there are some shortcomings limit the usage of skeleton features for action recognition, and joint estimates are unreliable and even fails in the case of self-occlusion.

For depth image-based methods, Chen *et al.* [7] utilize DMM and collaborative representation classifier to achieve real-time action recognition. In [21], Chen *et al.* employ the depth motion maps (DMMs) from three projection views (front, side and top) to capture motion cues, and use the local binary patterns (LBPs) to gain a compact feature representation. In [22], 3DMTM-PHPG model is proposed to represent the actions of depth maps. The spatio-temporal cuboid pyramid (STCP) [23] is proposed to subdivide the DMS volumes into a set of spatial cuboids on scaled temporal levels. Recently, a multilevel temporal sampling (MTS) [24] method is proposed, which is based on the motion energy of key-frames of depth sequences, and the histogram of gradient (HOG) and local binary pattern (LBP) are employed to extract features. In [25], a convolutional neural network method is developed for action recognition.

For the direction of depth and skeleton information fusion methods, Wang *et al.* [26] propose a model to associate local occupancy pattern features from depth images with skeleton joints. In [27], Liu *et al.* extract the dense action trajectories to encode the motion information, and pass them through a deep network to get the viewpoint invariant features. Aiming at the poor recognition performance caused by insufficient two-dimensional information, a human action recognition method by fusing multiple depth information is proposed in [28]. Since different types of features could share some similar hidden structures, and different actions may be well characterized by properties common to all features and those specific to a feature, Meng *et al.* [29] propose a joint group sparse regression-based learning method to model each action. Tang *et al.* [30] combine these two features to better represent actions, i.e. depth map-based features (hon4d) and skeleton-based features (Fourier time pyramid). In [31], motion history image (MHI), depth motion maps (DMM) and skeleton image are obtained from RGB-D sensor firstly, and then these images are then separately trained on ConvNets and respective softmax scores are fused at the decision level.

The method proposed in this paper fall into the depth image based category. Due to the new motion may cover the old motion history, the DMM based on the whole depth sequences may not capture the detailed temporal motion in the depth images. To this end, Chen *et al.* [32] propose a

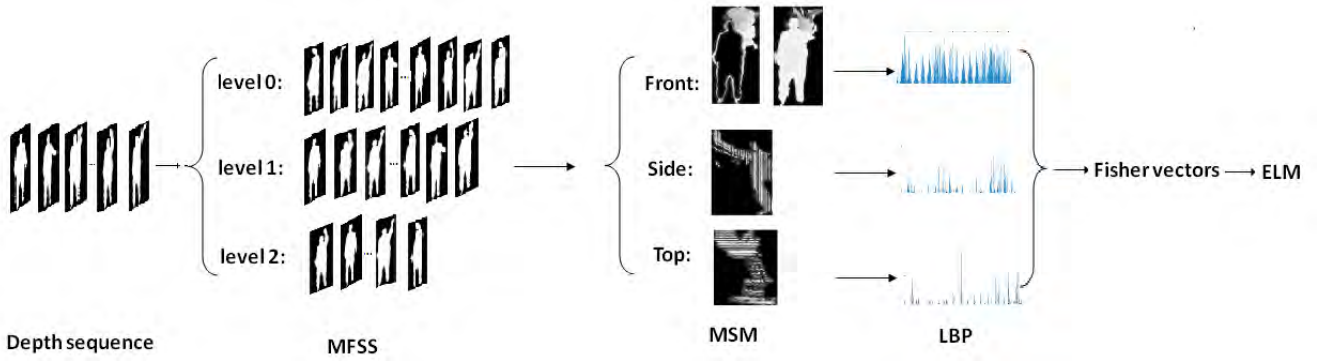


FIGURE 1. General overview of proposed method.

multi-temporal DMMs, which divides the depth sequences into overlapping segments and generate multiple sets of DMMs. Azad *et al.* [24] propose the multilevel temporal sampling method to solve the same problem. On the other hand, DMM can only extract motion information from depth sequences, while ignoring static information. Static information may also be very important for action recognition. In this paper, a more efficient multilevel temporal sampling method MFSS is proposed, and then, the MSM model which can capture both motion information and static information in the depth sequences is also proposed.

III. PROPOSED METHOD

A. GENERAL FRAMEWORK

The purpose of this paper is to design a robust human action representation method, and the proposed framework is illustrated in Figure 1.

B. MULTILEVEL FRAME SELECT SAMPLING (MFSS)

As mentioned earlier, motion maps based on the entire depth sequences may not capture detailed motion clues. Therefore, in order to overcome the shortcoming and obtain more motion information, the MFSS method is proposed. Different with the multilevel temporal sampling (MTS) [24] method which is based on the motion energy of key frames, MFSS method is a simple joint motion detection and frame selection operation [33].

Let I as the depth image sequences.

$$I = (I_1, I_2, I_3, \ldots, I_T) \tag{1}$$

where T is the total number of frames.

Let D_t be the difference image sequence which is calculated by:

$$D_t = \begin{cases} I_1 & if \quad t = 1 \\ \sqrt{(I_t - I_{t-1})^2} & otherwise \end{cases} \tag{2}$$

where t is the temporal index. The calculation of $(I_t - I_{t-1})^2$ is element-wise square, and $D_t \in R^{m \times n}$ denotes the matrix of L_2 norm difference at the pixel level from frame t-1 to frame t. m and n denote the frame resolution (rows and columns).

For t = 1, . . . , T, the sum of all elements of D_t is stored in the vector $d \in R^T$, whose t-th element is given by:

$$d_t = \sum_{i=1}^{m} \sum_{j=1}^{n} D_t(i, j) \tag{3}$$

where D_t(i,j) is the matrix element.

The vector D_t is normalized between 0 and 1 by:

$$d_m = \frac{d_t - d_{\min}}{d_{\max} - d_{\min}} \tag{4}$$

where d_{max} and d_{min} denote the maximum value and the minimum value of the vector d.

We can get the derivative of d_m by :

$$d' = \frac{d}{dt} d_m \tag{5}$$

The index of relevant frames is given by:

$$y = \begin{cases} 1 & if \ |d' - u| > \tau \\ 0 & otherwise \end{cases} \tag{6}$$

where u denotes the mean value of the d'. τ is the threshold operator.

In this way, redundant frames in depth sequences can be eliminated by setting different thresholds τ. In order to sample different levels of depth sequences, It is essential to select frames with relevant motion information by setting different thresholds. Particularly, frames with higher rate of change are selected. Here, we extract three frame sequence samples of level 0, level 1 and level 2 from the original depth sequences, which can be determined by setting different thresholds. To keep the depth sequences completely, all the frames in the sequence are select in the level 0. Then, we can get a new sequence by integrating the three sequences level 0, level 1 and level 2. In order to capture more motion information and avoid previous motion being covered, the depth sequences are divided into many sets of depth segments with a fixed length of N. Next, the MSM are computed for each clip. The proposed Multilevel Frame Select Sampling framework is shown in Figure 2.

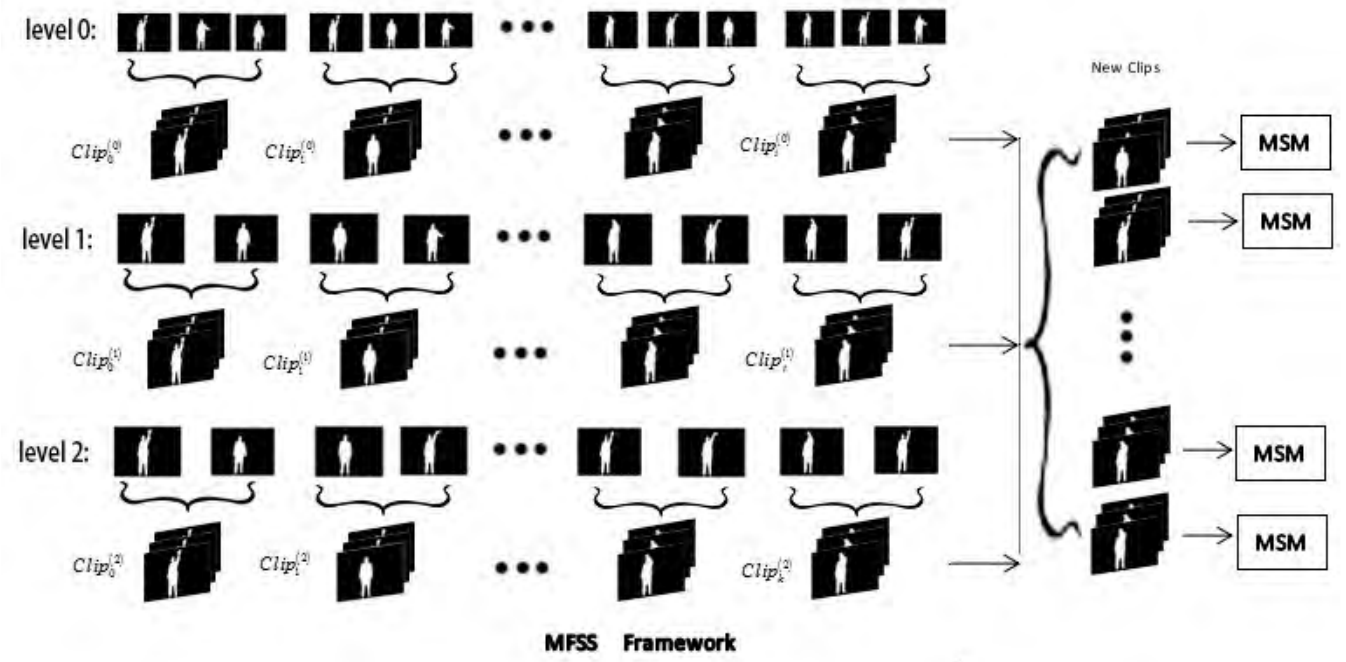


FIGURE 2. Multilevel frame select sampling Framework.

C. MOTION AND STATIC MAPS (MSM)

As human motion is carried out in three-dimensional space, for a depth action sequences, it contains three-dimensional depth information. Firstly, we project the three-dimensional depth frame onto three orthogonal Cartesian planes. Each plane is a view, denoted by MSMv, where v ∈ front, side, top.

DMM [7] can fully capture the shape information and three-dimensional local motion information represented by depth images in the depth sequences. Because of the simplicity of calculation and good performance, many researchers extract DMM from depth images for the action recognition. However, DMM only obtains the motion information from the depth clips. In this subsection, a novel model (MSM) is proposed, which uses motion history image (MHI) and static history image (SHI) to represent motion posture and static posture in the depth sequences respectively.

In order to obtain the information of motion and static state, we define the motion binary function M(x,y,t) and the static binary function S(x,y,t).

$$M(x,y,t) = \begin{cases} 1 & if \quad D_t > S_M \\ 0 & otherwise \end{cases} \tag{7}$$

$$S(x,y,t) = \begin{cases} 1 & if \quad I_t - D_t > S_S \\ 0 & otherwise \end{cases} \tag{8}$$

where S_M and S_S are the thresholds for motion and static information between consecutive frames respectively. When S_M is set as 50 and S_S is set as 10, better performance could be obtained. I_t is the depth image sequences, and D_t is the difference between two consecutive image sequences.

Motion History Image (MHI) is proposed by Bobick [34], which can describe the position of motion. It could encode the motion information of all the frames in one depth clips into a single static image [35]. The MHI is expressed as follows:

$$MHI(x,y,t) = \begin{cases} \sigma & if \quad M(x,y,t) = 1 \\ \max(0, MHI(x,y,t-1) - 1) & otherwise \end{cases} \tag{9}$$

where MHI(x,y,t) is the motion history image, x and y represent pixel position and t is the temporal index, M(x,y,t) is the binary difference image, σ is a threshold for extracting moving patterns in depth sequences.

Furthermore, we utilize the static binary function SHI(x,y,t) to get the static history image (SHI), which is expressed as follows:

$$SHI(x,y,t) = \begin{cases} \sigma & if \quad S(x,y,t) = 1 \\ \max(0, SHI(x,y,t-1) - 1) & otherwise \end{cases} \tag{10}$$

The information from front plane is dominant for the action. However, the side plane and top plane may be very coarse, so only MHI templates are generated from the front plane and top plane respectively. Therefore, one depth image sequences can be modeled as four templates (F_{MHI}, F_{SHI}, S_{MHI}, T_{SHI}) using the proposed MSM. The MSM of one depth sequence is demonstrated in Figure 3.

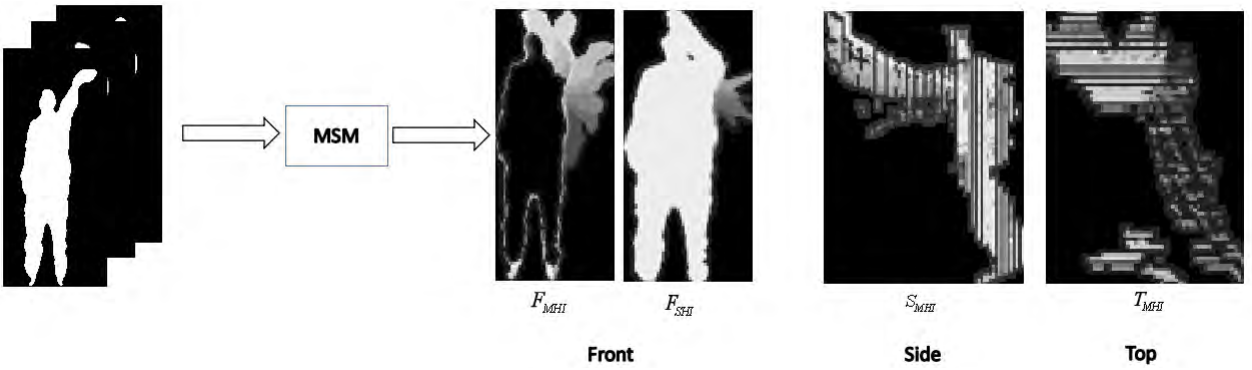


FIGURE 3. MSM generation for one depth sequence.

D. BLOCK-BASED LBP FEATURES

The local binary patterns (LBP) [14] is a simple and effective gray and rotation invariant texture operator, which describes the local texture pattern of an image by labeling the image pixels with binary code.

As histogram of gradient (HOG) [36] is a gradient-based feature extraction method, it considers the change rate of adjacent pixel values in different directions. Nevertheless, LBP features are based on simultaneous comparisons with all adjacent pixels in all directions. Therefore, LBP is more powerful than HOG.

For each pixel q, in an depth image, a set of m neighbors contains these pixels that are equally spaced on a circle of radius r. As in [32], the LBP for q can be expressed in decimal form as follows:

$$LBP_{m,r}(q) = \sum_{i=0}^{m-1} U(q_i - q)2^i \tag{11}$$

$$U(q_i - q) = \begin{cases} 1 & if \quad q_i \geq q \\ 0 & otherwise \end{cases} \tag{12}$$

where the q_i is the i_{th} neighbor around pixel q with a circle of radios r centered at q. In this work, $r = 1$ and $m = 4$ are used.

Therefore, MHI mapping and SHI mapping are first generated for the depth sequence, and then LBP operators are applied to the mappings.

E. ENCODING AND CLASSIFICATION

Fisher Vector outperforms other compared encoding methods showing that the encoding of second order information indeed benefits classification performance [37]. Therefore, the Fisher kernel is employed to encode the block-based LBP features.

For each projection view, Gaussian mixture model (GMM) parameters are estimated by expectation-maximization (EM) algorithm using the corresponding feature matrices of the training data. Then, the four Fisher Vector are simply concatenated as the final feature representation.

Extreme learning machine (ELM) [38] is a neural network with only one hidden layer and one linear output layer, and the computing cost is much lower than other methods based on neural network. The weights between the input layer and the hidden layer are randomly assigned, and then the weights of the output layer are calculated by the least square method.

Compared with ELM, kernel-based ELM (KELM) [16] has been proposed by extending explicit activation functions in ELM, which provides a better generalization performance and is more stable.

IV. EXPERIMENTAL RESULTS

In order to demonstrate the recognition performance of the algorithm, we have carried out the following experiments and analysis on three public datasets. Notice that although these datasets contains both the color and depth frames, only depth frames are used in the experiments.

Due to the background of the depth sequences would introduce noise to the recognition, we employ the method of image smoothing to preprocess the depth sequences, and remove the salt and pepper noise in the depth sequences.

A. EVALUATION CRITERION

In order to evaluate our proposed method, precision, recall and accuracy are used.

Precision is ration between true positive and sum of positive data. This can be interpreted as what portion of predicted targeted class is relevant, i.e. are from this class. The formula for precision is

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{13}$$

Recall is ration between true positive and sum of data from target class. The formula for recall is

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{14}$$

Accuracy is the ratio of the number of samples correctly classified to the total number of samples. The formula is

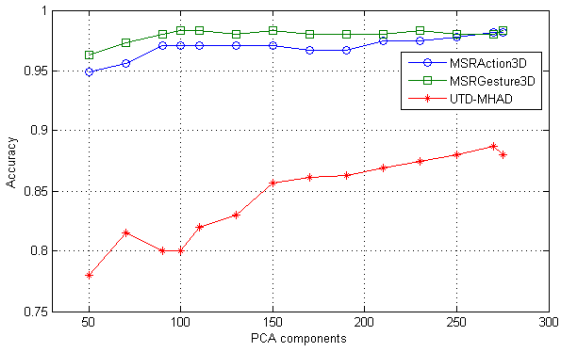


FIGURE 4. Accuracy vs PCA components on on three datasets.

as follows:

$$accuracy = \frac{true\ positive + true\ negative}{the\ total\ number\ of\ samples} \tag{15}$$

B. MSRACTION3D DATASET

MSRAction3D dataset [5] is a public depth datasets for human action recognition, which consists of 20 actions. Each action is performed 2 or 3 times by 10 subjects facing the depth camera. 20 kinds of actions include "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", "pick up & throw". It has a total of 556 depth sequences, and the size of each frame is 240×320. This dataset is quite challenging due to the high similarity between many actions.

We follow the same experimental settings in [39] (subjects 1, 3, 5, 7, 9 for training and the rest for testing). The sizes of MSM of front, side and top are normalized to be 102 × 54, 102 × 75 and 75 × 54 respectively. The frame selection parameter τ is chosen experimentally for each level: $\tau = 0.04$ for level 1, $\tau = 0.09$ for level 2, and all the frames in the sequence is selected in the level 0. The input frame is divided into many blocks for LBP, and the blocks are selected with 50% of overlap. The length of LBP feature vector is 59. The radial basis function (RBF) kernel is employed in KELM.

Figure 4 shows the accuracy on the validation set when selecting different PCA components. For the range of 90 to 190, the accuracy is almost the same. By increasing this number from 200 to 270, the accuracy of the method grows. After 270, the accuracy does not change. It can be seen that 270 is the best value for the MSRAction3D datasets.

The confusion matrix of our method for MSRAction3D dataset is shown in Figure 5. Figure 6 shows two other evaluation indicators: precision and recall. The incorrect recognitions mainly appear on some partly similar actions, e.g., highThrow, drawX and drawTick. But for these three actions, 92%, 81%, 94% are achieved by the proposed method. Using only the single LBP feature, the cross test accuracies are 98.2%.

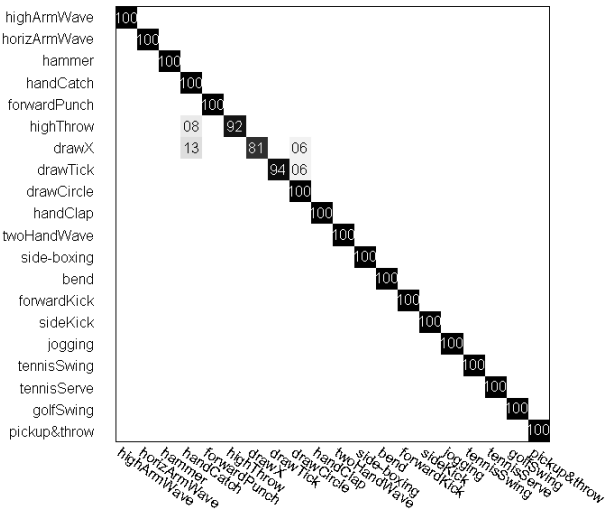


FIGURE 5. Confusion matrix on MSRAction3D.

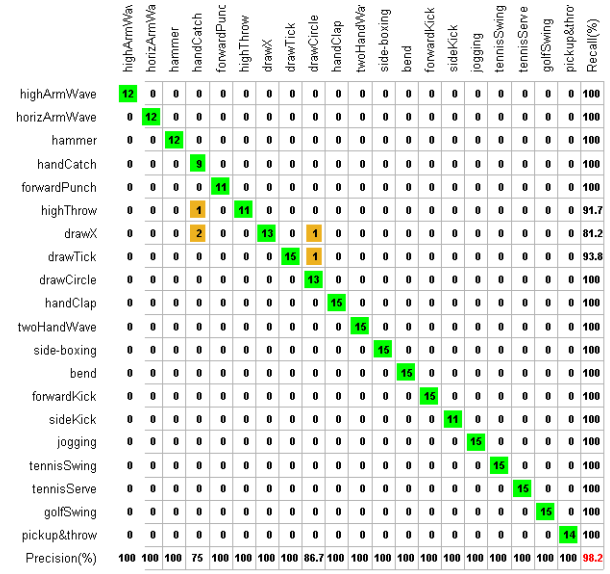


FIGURE 6. Precision and recall on MSRAction3D. The number in red indicates the overall accuracy.

TABLE 1. Performance comparison on MSRAction3D.

Method	Year	Accuracy
Bag of 3D Points [5]	2010	74.7%
Random Occupancy Pattern [40]	2012	86.2%
DMM-HOG [41]	2012	88.7%
Actionlet Ensemble [26]	2014	88.2%
DMM-LBP-DF [21]	2015	93.9%
Extended SNV [39]	2017	93.4%
Multi-Temporal DMMS [32]	2017	94.5%
Deep Convolutional Neural Networks [8]	2018	94.5%
Multilevel Temporal Sampling [24]	2018	95.2%
Proposed MFSS-MSM	**2019**	**98.2%**

Table 1 shows the results compared with other methods under the same test conditions, and our proposed method achieves the best result.

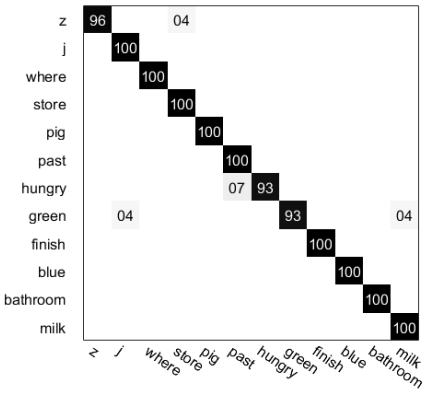


FIGURE 7. Confusion matrix on MSRGesture3D.

TABLE 2. Performance comparison on MSRGesture3D.

Method	Year	Accuracy
Random Occupancy Pattern [40]	2012	88.5%
DMM-HOG [41]	2012	89.2%
HON4D [43]	2013	92.5%
DMM-LBP-DF [21]	2015	94.6%
Extended SNV [39]	2017	94.7%
3DHoTs [44]	2017	94.7%
Multi-Temporal DMMS [32]	2017	98.2%
Multilevel Temporal Sampling [24]	2018	98.1%
Multilevel Temporal Sampling + CNN [24]	2018	97.2%
Proposed MFSS-MSM	**2019**	**98.3%**

C. MSRGESTURE3D DATASET

MSRGesture3D dataset [42] is a dynamic hand gesture dataset of depth sequences captured by a depth camera. This dataset contains 12 gestures defined by American sign language, which is considered to be more challenging than MSRAction3D due to more self-occlusion issues. The 12 standard gesture include ''bathroom'', ''blue'', ''finish'', ''green'', ''hungry'', ''milk'', ''past'', ''pig'', ''store'', ''where'', ''j'', and ''z''. Each gesture is performed 2 or 3 times by each one of the 10 subjects. The sizes of MSM of front, side and top are normalized to be 118×133, 118×29 and 29×133 respectively. For this dataset, the leave-one-subject-out evaluation scheme is performed [32]. Other parameter settings are the same as MSRAction3D dataset.

Figure 4 shows the accuracy on the validation set when selecting different PCA components. By increasing this number from 50 to 100, the accuracy of the method grows. For the range of 100 to 280, the accuracy is almost the same. It can be seen that 100 is the best value for the MSRGesture3D dataset.

The confusion matrix of our method for MSRGesture3D dataset is shown in Figure 7. Figure 8 shows two other evaluation indicators: precision and recall. The incorrect recognitions mainly appear on some partly similar actions, e.g., where, hungry and green. For these three actions, 93% are achieved by the proposed method. The cross test accuracies are 98.3%. Table 2 shows the results compared with other methods.

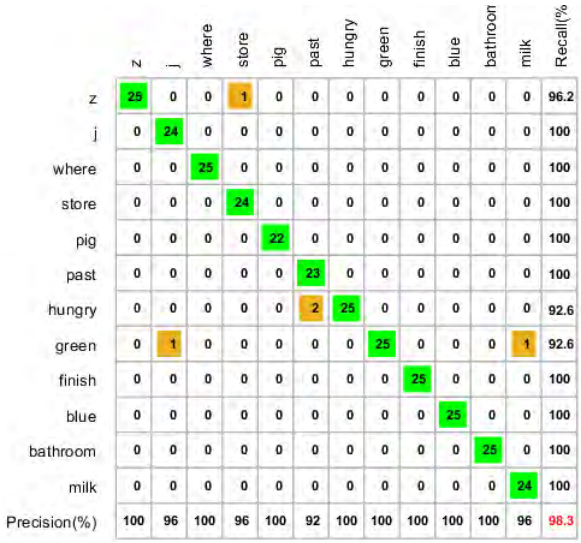


FIGURE 8. Precision and recall on MSRGesture3D. The number in red indicates the overall accuracy.

D. UTD-MHAD DATASET

The UTD-MHAD dataset [45] was collected using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. The dataset contains 27 actions performed by 8 subjects (4 females and 4 males). Each subject repeated each action 4 times. After removing three corrupted sequences, the dataset includes 861 data sequences. The 27 actions are as follows: ''right arm swipe to the left'', ''right arm swipe to the right'', ''right hand wave'', ''two hand front clap'', ''right arm throw'', ''cross arms in the chest'', ''basketball shoot'', ''right hand draw X'', ''right hand draw circle'', ''right hand draw circle'', ''draw triangle'', ''bowling'', ''front boxing'', ''baseball swing from right'', ''tennis right hand forehand swing'', ''arm curl'', ''tennis serve'', ''two hand push'', ''right hand knock on door'', ''right hand catch an object'', ''right handpick up and throw'', ''jogging in place'', ''walking in place'', ''sit to stand'', ''stand to sit'', ''forward lunge'', ''squat''.

We follow the same experimental settings in [45]. The data from the subject numbers 1, 3, 5, 7 are used for training, and the data for the subject numbers 2, 4, 6, 8 are used for testing. In our experiments, the sizes of MSM of front, side and top are normalized to be 150×75, 150×100 and 100×75 respectively. Other parameter settings are the same as MSRAction3D dataset.

Figure 4 shows the accuracy on the validation set when selecting different PCA components. By increasing this number from 50 to 250, the accuracy of the method grows. After 270, the accuracy drops slightly. It can be seen that 260 is the best value for the UTD-MHAD dataset.

The confusion matrix of our method for UTD-MHAD dataset is shown in Figure 9. Figure 10 shows two other evaluation indicators: precision and recall. The cross test accuracies are 88.7%. Table 3 shows the results compared with other methods that also use only depth images.

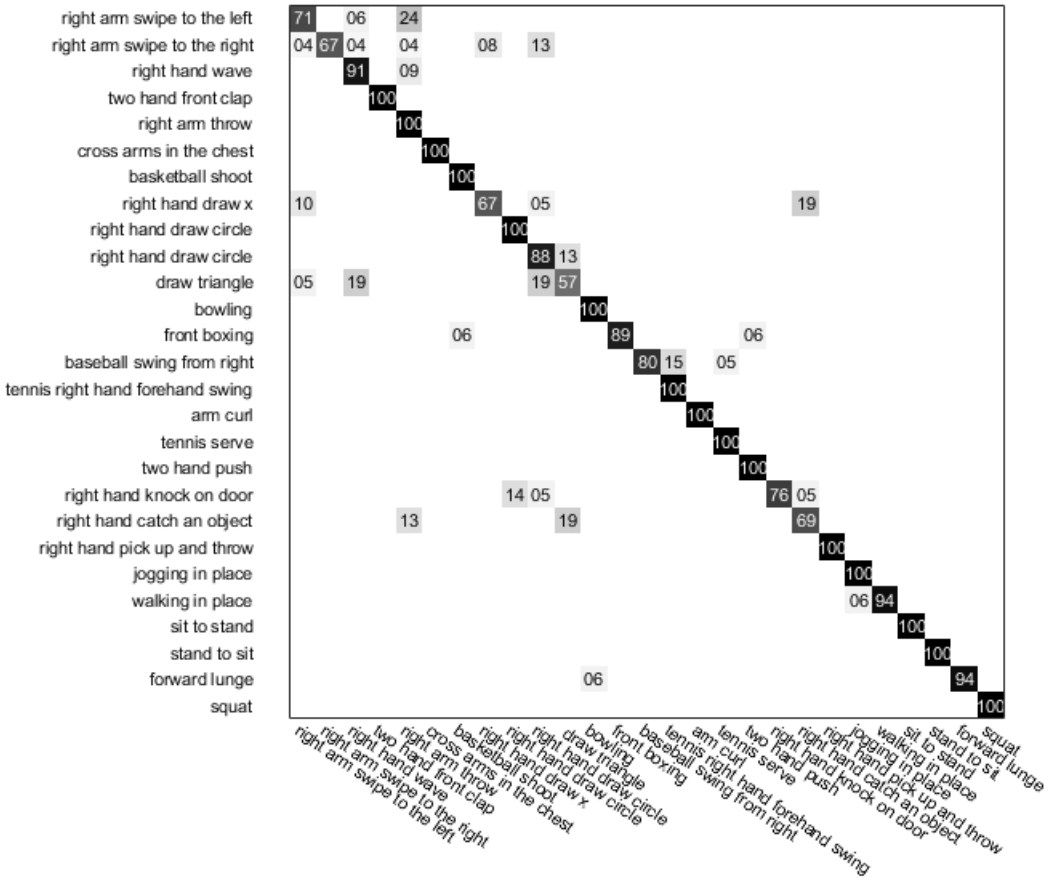


FIGURE 9. Confusion matrix on UTD-MHAD.

TABLE 3. Performance comparison on UTD-MHAD.

Method	Year	Accuracy
DMM-HOG [41]	2012	81.5%
Multi-Temporal DMMS [32]	2017	82.0%
3DHoT-MBC [44]	2017	84.4%
Two-stream Entirety Net [46]	2018	85.4%
Multilevel Temporal Sampling [24]	2018	81.1%
Proposed MFSS-MSM	**2019**	**88.7%**

The results clearly demonstrate the superior performance of our method.

E. PARAMETER ANALYSIS

Some key parameters used in this paper are analyzed in this section. We use the MSRAction3D dataset as the benchmark, and other datasets could also get similar results.

1) PARAMETER τ ANALYSIS

As the value of parameter τ increases, fewer image frames is selected. Therefore, the value of parameter τ would affect the recognition results. In order to make the experiment more comparable, we use the method of fixing one parameter and changing another parameter.

As shown in Figure 11(a), we fix the parameter τ of level 2 as 0.09, and check the recognition accuracy by changing the parameters τ of level 1. When the parameter τ of level 1 is 0.04, we can get the highest accuracy.

On the other hand, as shown in Figure 11(b), we fix the parameter τ of level 1 as 0.04, and check the recognition accuracy by changing the parameter τ of level 2. With the increase of parameter τ of level 2, more frames will be deleted. After 0.16, the accuracy decreases slightly, and then remains unchanged. When the parameter τ of level 2 is 0.09, we can get the highest accuracy.

Therefore, while the frame selection parameter τ of level 1 is 0.04 and that of level 2 is 0.09, we can get the best accuracy.

2) MSM THRESHOLD ANALYSIS

S_M and S_S are the thresholds for motion and static information between consecutive frames respectively. Experiments show that better performance can be obtained when S_M is set as 50 and S_S is set as 10. With the increase of S_M, fewer pixels will be judged as motion information. As shown in Figure 12, our experimental method is similar to parameter τ's.

3) MULTILEVEL TEMPORAL SAMPLING ANALYSIS

We also test the algorithm on the MSRAction3D dataset by using different temporal levels. As shown in table 4, the accuracy of using all three temporal levels is the best.

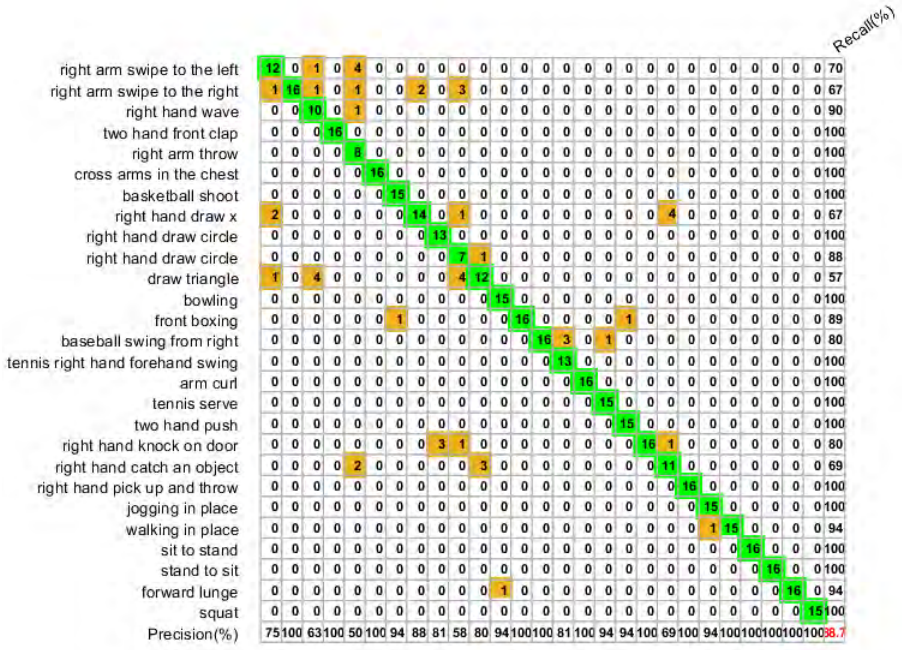


FIGURE 10. Precision and recall on UTD-MHAD. The number in red indicates the overall accuracy.

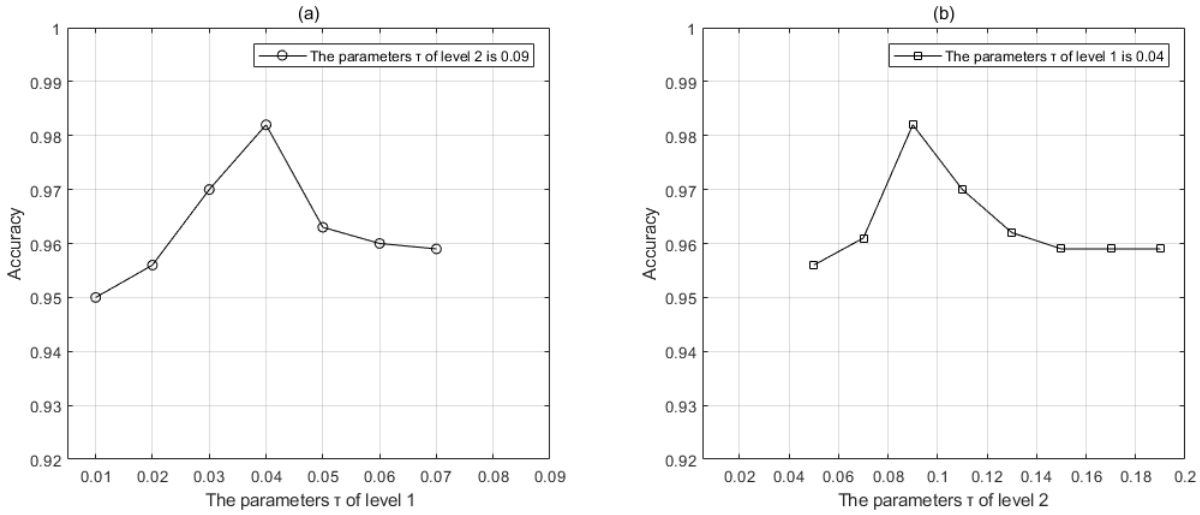


FIGURE 11. The parameter τ analysis on MSRAction3D.

TABLE 4. Recognition accuracy of our method with different temporal levels on the MSRAction3D dataset.

Temporal levels	Accuracy
1 level(level 0)	94.5%
2 levels(levels 0,1)	95.7%
3 levels(levels 0,1,2)	98.2%

TABLE 5. Comparison of different combinations of descriptors on the MSRAction3D dataset.

Descriptors	Accuracy
HOG	95.2%
LBP	98.2%
HOG + LBP	98.3%

4) COMBINING DIFFERENT FEATURES ANALYSIS

In addition, table 5 shows the accuracy of different combinations of HOG and LBP descriptors. It can be seen that the combination of HOG and LBP features is a little better than using LBP alone. In order to reduce the computational cost, only LBP features is used in this paper.

F. COMPUTATION TIME ANALYSIS

It is worth mentioning that our algorithm is implemented in MATLAB and executed on CPU platform with an Intel(R)Core(TM) i5-5200U CPU @2.20GHz and 8GB of RAM.

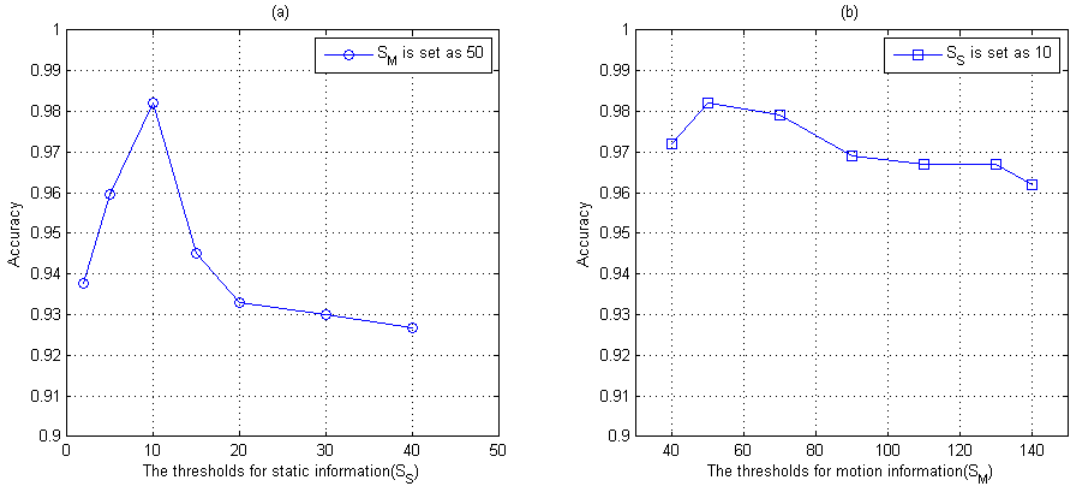


FIGURE 12. The MSM threshold analysis on MSRAction3D.

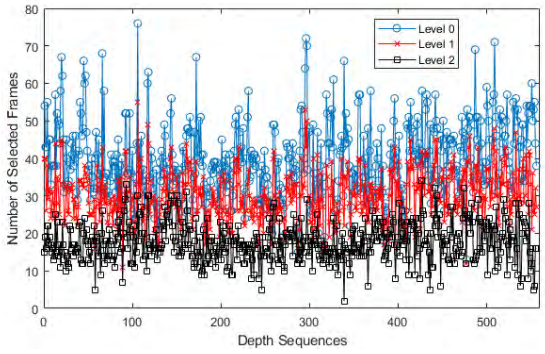


FIGURE 13. Illustration of MFSS operation.

TABLE 6. Average run-time on MSRAction3D.

Step	Run Time(%)	Time/sequence(sec)
MFSS	5.99	0.536568
MSM	70.2	6.283341
LBP	17.96	1.607247
Fisher Vector	5.15	0.461127
ELM	0.66	0.058696

For MFSS operation, three levels representation is used in our method. All the frames in the sequence is selected in the level 0, and the thresholds of level 1 and level 2 is selected to be 0.04 and 0.09 respectively. Figure 13 illustrates the MFSS operation on the MSRAction3D dataset, which include 556 depth sequences. About 75.74% of frames are selected in level 1 and 44.35% in level 2.

Table 6 shows the percentage of time spent on each step of the proposed method. Computing MSM takes the most part of time with 70.2%.

V. CONCLUSION

In this paper, based on only the depth images, we proposed a new effective framework for human action recognition. The motivation of this paper comes from two problems in the process of action recognition. Firstly, the new motion may cover the old motion history, while the DMM based on the whole depth sequence may not capture the detailed temporal motion in the depth image. Secondly, DMM can only extract motion information from depth sequence, but ignore static information, which is also very important for motion recognition. Using multilevel frame select sampling (MFSS) model, we successfully capture three levels of temporal samples from the input depth images firstly. Then, we project each depth image onto three orthogonal Cartesian planes, and then using motion and static maps (MSM) method to get the motion history image and static history image to represent motion posture and static posture respectively. After that, the block-based LBP feature extraction approach is employed to extract texture information. In order to aggregate the block features, the fisher kernel representation is applied. At the end, kernel-based extreme learning machine (KELM) is used as the classifier. In addition, the key parameters used in the framework are analyzed. The best parameter τ and MSM thresholds are analysised in detail. It is proved that the three temporal level can achieve better recognition accuracy compared with other temporal levels. Using LBP features alone can achieve good performance at lower computational cost. Finally, The calculation time of the framework is calculated and analyzed in detail.

The proposed method is extensively evaluated on three public datasets. The experimental results demonstrate that the proposed framework has shown many competitive and attractive characteristics for depth based action recognition. Due to neural networks shows better experimental results than classical feature extraction methods, which can do both feature extraction and classification. As our future work, we plan to apply Convolutional Neural Network for complex recognition system, and larger RGB-D action recognition dataset such as NTU dataset [47] will be considered to analyze.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this manuscript.

REFERENCES

[1] A.-L. Yussiff, S.-P. Yong, and B. B. Baharudin, "Human action recognition in surveillance video of a computer laboratory," in *Proc. Int. Conf. Comput. Inf. Sci.*, Aug. 2016, pp. 418–423.

[2] X. Liu, T. You, X. Ma, and H. Kuang, "An optimization model for human activity recognition inspired by information on human-object interaction," in *Proc. Int. Conf. Measuring Technol. Mechatronics Automat.*, Feb. 2018, pp. 519–523.

[3] Y. Kong and Y. Fu. (2018). "Human action recognition and prediction: A survey." [Online]. Available: https://arxiv.org/abs/1806.11230

[4] Z. Zhang *et al.*, "Deep learning based human action recognition: A survey," in *Proc. Chin. Autom. Congr.*, Oct. 2018, pp. 3780–3785.

[5] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 9–14.

[6] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "3D hand gesture recognition by analysing set-of-joints trajectories," in *Proc. Int. Workshop Understand. Hum. Activities Through 3D Sensors*, 2016, pp. 86–97.

[7] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *J. Real-Time Image Process.*, vol. 12, no. 1, pp. 155–163, Aug. 2013.

[8] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[9] A. Farooq, F. Farooq, and A. V. Le, "Human action recognition via depth maps body parts of action," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 5, pp. 2327–2347, 2018.

[10] R. Cui, G. Hua, A. Zhu, J. Wu, and H. Liu, "Hard sample mining and learning for skeleton-based human action recognition and identification," *IEEE Access*, vol. 7, pp. 8245–8257, 2017.

[11] W. Ding, L. Kai, C. Fei, and Z. Jin, "STFC: Spatio-temporal feature chain for skeleton-based human action recognition," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 329–337, Jan. 2015.

[12] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5833–5842.

[13] T. Kerola, N. Inoue, and K. Shinoda, "Cross-view human action recognition from depth maps using spectral graph sequences," *Comput. Vis. Image Understand.*, vol. 154, pp. 108–126, Jan. 2017.

[14] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[15] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[16] C. M. Wong, C. M. Vong, P. K. Wong, and J. Cao, "Kernel-based multilayer extreme learning machines for representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 757–762, Mar. 2018.

[17] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[18] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1110–1118.

[19] W. Zhu *et al.* (2016). "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks." [Online]. Available: https://arxiv.org/abs/1603.07772

[20] L. Chao, Q. Zhong, X. Di, and S. Pu. (2017). "Skeleton-based action recognition with convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1704.07595

[21] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 1092–1099.

[22] B. Liang and L. Zheng, "3D motion trail model based pyramid histograms of oriented gradient for action recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1952–1957.

[23] X. Ji, J. Cheng, and W. Feng, "Spatio-temporal cuboid pyramid for action recognition using depth motion sequences," in *Proc. 8th Int. Conf. Adv. Comput. Intell.*, Feb. 2016, pp. 208–213.

[24] R. Azad, M. Asadi-Aghbolaghi, S. Kasaei, and S. Escalera, "Dynamic 3D hand gesture recognition by learning weighted depth motion maps," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[25] P. Wang, W. Li, Z. Gao, C. Tang, and P. Ogunbona, "Depth pooling based large-scale 3D action recognition with convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, p. 1.

[26] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3D human action recognition," in *Human Action Recognition with Depth Cameras*. Cham, Switzerland: Springer, 2014.

[27] J. Liu, N. Akhtar, and A. Mian. (2017). "Viewpoint invariant action recognition using RGB-D videos." [Online]. Available: https://arxiv.org/abs/1709.05087

[28] X. Yan, H. Zhenjie, and L. Jiuzhen, "Action recognition using weighted fusion of depth images and skeleton's key frames," *J. Comput.-Aided Des. Comput. Graph.*, vol. 30, no. 7, 2018.

[29] L. Meng, H. Leung, and P. H. S. Hubert, "Human action recognition via skeletal and depth based feature fusion," in *Proc. Int. Conf. Motion Games*, 2016, pp. 123–132.

[30] N. C. Tang, Y.-Y. Lin, J.-H. Hua, M.-F. Weng, and H.-Y. M. Liao, "Human action recognition using associated depth and Skeleton information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4608–4612.

[31] P. Khaire, P. Kumar, and J. Imran, "Combining CNN streams of RGB-D and skeletal data for human activity recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 107–116, Nov. 2018.

[32] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition," *IEEE Access*, vol. 5, pp. 22590–22604, 2017.

[33] A. Sobral, T. Bouwmans, and E.-H. Zahzah, "Comparison of matrix completion algorithms for background initialization in videos," in *Proc. Int. Conf. Image Anal. Process.*, vol. 9281, 2015, pp. 510–518.

[34] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[35] M. Vafadar and A. Behrad, "Human hand gesture recognition using motion orientation histogram for interaction of handicapped persons with computer," in *Proc. Int. Conf. Image Signal Process.*, vol. 37, no. 37, 2008, pp. 378–385.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[37] M. Seeland, M. Rzanny, N. Alaqraa, J. Wäldchen, and P. Mäder, "Plant species classification using flower images—A comparative study of local feature representations," *PLoS ONE*, vol. 12, no. 2, 2017, Art. no. e0170629.

[38] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[39] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.

[40] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 872–885.

[41] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 1057–1060.

[42] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Proc. 20th Eur. Signal Process. Conf.*, 2012, pp. 1975–1979.

[43] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 716–723.

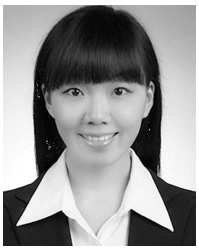
[44] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, "Action recognition using 3D histograms of texture and a multi-class boosting classifier," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4648–4660, Oct. 2017.

[45] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 168–172.

[46] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1044–1048, Jul. 2018.
[47] A. Shahroudy, J. Liu, T.-T. Ng, and W. Gang. (2016). "NTU RGB+D: A large scale dataset for 3D human activity analysis." [Online]. Available: https://arxiv.org/abs/1604.02808

XU WEIYAO was born in 1984. He received the master's degree in communication and information systems from the Nanjing University of Posts and Telecommunications, in 2011. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. His current research interests include around machine learning and human action recognition.

WU MUQING was born in 1963. He received the Ph.D. degree. He is currently a Professor with the Beijing University of Posts and Telecommunications and a Senior Member of the China Institute of Communications. His current research interests include mobile ad hoc networks, UWB, highspeed network traffic control and performance analysis, and GPS locating and services.



ZHAO MIN received the Ph.D. degree in information and telecommunication systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014, where she is currently a Lecturer with the Laboratory of Network System Architecture and Convergence. Her research interest includes new network architecture.

LIU YIFENG received the Ph.D. degree in electronic engineering from Wuhan University, Wuhan, China, in 2016. He is currently the Principal Investigator of machine intelligence with the Innovation Center, China Academy of Electronics and Information Technology, Beijing, China. His current research interests include machine learning and computer vision.

LV BO received the Ph.D. degree in information and communication from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014. He was a Researcher with the Innovation Center, China Academy of Electronics and Information Technology. He was a Visiting Scholar with the Department of Computer Science, Worcester Polytechnic Institute (WPI), Worcester, MA, USA, in 2016. His research interests include urban computing and data visualization.

XIA TING was born in 1986. She received the master's degree in communication and information systems from the Hangzhou University of Electronic Science and Technology, in 2011. She is currently a Lecturer with Zaozhuang University. Her current research interests include machine learning and human action recognition.