International Journal of Multimedia and Ubiquitous Engineering
Vol.12, No.1 (2017), pp.203-214
http://dx.doi.org/10.14257/ijmue.2017.12.1.17

Action Recognition Using Motion History Image and Static History Image-based Local Binary Patterns

Enqing Chen[1], Shichao Zhang*[1] and Chengwu Liang[1]

[1]*School of Information Engineering, Zhengzhou University, Zhengzhou 450000, China*
[1]*E-mail: zhangshichao816816@163.com*

Abstract

Human action recognition is an important yet challenging task. In this paper, we propose a robust and effective framework to largely improve the performance of human action recognition using depth maps. The key contribution is the Motion History Image (MHI) and Static History Image (SHI) is used to represent depth sequence. And we optimize the condition to construct the MHI and SHI; it allows us to capture more critical information. The local binary pattern (LBP) are then computed to gain the compact feature representation of an action. We evaluate the proposed framework on MSR Action3D dataset. Experimental results indicate that the proposed approach outperforms the state-of-the-art methods and demonstrate the effectiveness of the proposed approaches.

Keywords*: motion history image, static history image, local binary patterns*

1. Introduction

Human action recognition has been an active field in computer vision, due to its extensive application in real-word, such as human computer interaction, medical health care, and video retrieval. In the past few decades, research has been mainly focused on recognizing actions from videos taken by ordinary RGB cameras. There are intrinsic defects of this type of data source, *e.g.* it is sensitive to illumination changes, occlusions, and background clutters. While significant effort, recognizing actions accurately still remain a challenging task.

As the imaging techniques advance, the release of the Microsoft Kinect provides a new possibility to address these issues. Using Kinect, depth information can be captured simultaneously with RGB videos. Depth maps have several advantages with regard to color images in activity recognition. First, they provide an efficient and powerful human motion capturing technology which can accurately estimate the 3D skeleton joint positions from a single depth map [1]. Second, depth cameras are robust to the change in color and illumination, which brings great benefits to the activity recognition.

There is massive literature in action recognition in the research field of computer vision and pattern recognition [2-3]. Most of the existing methods can accurately recognize some certain actions. But for similar actions, the recognition accuracy and robustness are not satisfactory. And the average recognition rate is also not very well.

In this paper, to overcome this problem, we use an effective and robust approach to recognize actions by extracting Local Binary Pattern (LBP) descriptors from Motion History Images (MHI) [23] and Static History Images (SHI) [23]. In our work, we optimize the construct condition of MHI and SHI. It makes we can capture more motion information. After extract LBP feature, our feature has a better discriminative power. And the feature dimension is smaller, which reduces the computational cost. We evaluate our

*Shichao Zhang, E-mail: zhangshichao816816@163.com

ISSN: 1975-0080 IJMUE

method according to the standard experimental protocols on the MSRAction3D dataset. From the experimental results and comparisons with state-of-the-art approaches, we show the effectiveness and robustness of our method. Especially to similar actions, we can also achieve a high recognition rate.

The rest of the paper is organized as follows. Section 2 discusses the related work on depth-based action recognition. Section 3, we describe the detailed procedures of the computing MHI-LBP and SHI-LBP. Section 4 is the experiment and comparison. Finally, the conclusion of this paper is presented in Section 5.

2. Related Work

With the introduction of the low-cost RGB-D cameras (Kinect), action recognition in depth videos has become a very active topic. In this field, different approaches have been proposed. In this section, we give a review of the research efforts for depth-based action recognition.

Li *et al.* [11] sample a bag of 3D points from the depth map to describe a set of salient postures that correspond to the nodes in the action graph. And the action graph was used to model the dynamics of actions. In addition, the sampling scheme is view dependent and more accurate than using 2D silhouettes.

Yang and Tian [10] used a Naïve-Bayes-Nearest-Neighbor (NBNN) classifier to recognize human actions. They combined static posture, motion, and offset information to form an action feature descriptor called EigenJoints. In order to eliminate noisy frame and reduce computational cost, Accumulated Motion Energy (AME) was performed to select informative frame. What's more, this method is not always available due to inaccuracies in skeleton estimation.

Xia *et al.* [12] suggest a compact posture representation through a histogram of 3D joint locations (HOJ3D). They aligned the spherical coordinates with the person's specific direction. Then, the Linear Discriminant Analysis (LDA) was used to extract the dominant features. The K-means clustering was performed to represent each posture as a visual word, and then a Bag of Words (BOW) model was used to translate each action into a series of symbols. After that, a discrete HMM classifier was used for action recognition.

Yang *et al.* [13] employ a HOG feature extraction after projecting the depth maps into three orthogonal planes and accumulating the depth maps throughout each posture into a motion image. A linear SVM classifier used to recognize actions.

Wang and Liu [14] treat an action sequence as a 4D shape and propose random occupancy pattern features, which are extracted from randomly sampled 4D sub-volumes with different sizes and at different locations. These features are robust to noise and insensitive to occlusions. In order to deal with the errors of the skeleton tracking, they defined action-let as a conjunction structure on base features. An Elastic-Net regularization algorithm was performed to find discriminative action-lets. Finally a SVM classifier is used to action classification.

Sung *et al.* [16] used RGB, depth maps and skeleton joint positions for action recognition. The proposed method consists of two steps. First, they compute the Histogram of Oriented Gradients (HOG) feature in both the RGB and depth maps within the bounding box of the person. Second is to get the bounding boxes for the head, torso, left arm, and right arm, used the skeleton joint positions. Then compute the HOG in RGB and depth with each of the four bounding boxes. A two-layered maximum-entropy Markov model was trained for action recognition.

As the spatial-temporal interest points can provide a compact representation of the image content by describing local areas of the scene thus offer robustness to clutter, occlusions, and intra-class variations [7, 9, 27]. Recently, many methods are proposed for action recognition based on spatial-temporal interest points. Zhang *et al.* [17] proposed a 4D local spatial-temporal feature which combines both intensity and depth information.

They first used separate filters along the 3D spatial dimensions and the temporal dimension to detect interest point. After that, they computed the intensity and depth gradients with a 4D hyper cuboids to obtain features for action sequence, and the Latent Dirichlet Allocation with Gibbs sampling was used to classify the action.

Xia *et al.* [18] proposed a spatiotemporal interest point detector on depth map, which used a correction function based on different nature noise in depth videos. It can effectively eliminates the noise ('signal flip' and 'holes') appear on depth maps. They extended the cuboids detector [9] to the fourth dimension. A depth cuboids similarity descriptor is proposed to describe the local feature. A feature selection process based on F-score is applied to generate the feature vector. Super Vector Machine is used for action classification.

The Sparse Representations [19-20] is the classic representation scheme and been applied in the field of action recognition. There have been several approaches for action recognition that use sparse representations. A zary and Savakis [21] applied sparse representations to construct the scale and position invariant features, which used spatial-temporal kinematic joint features and raw depth features. They create an over complete dictionaries and use both L1-norm and L2-norm minimization to classification the actions.

Zheng and Jiang [22] proposed a dictionary learning framework for cross-view action recognition. In the method, they assumption the sparse representations of videos from different views of the same action is strictly equal. And the assumption is too strong to flexibly model the relationship between different views.

These methods can recognition action performed by same person, but for different people to do the action recognition rate is low. The robustness of these methods is not very good. Different from these approaches, our proposed method in this paper is good at capturing the detail information. Especially to similar actions, we can also achieve a high recognition rate. In the following, we introduce the proposed method in detail.

3. Proposed Method

This section offers a detailed description for human action recognition from depth maps. The framework is demonstrated in Figure 1. The proposed method consists of two components. One is feature extraction from SMHI and SHI template, while the other is feature representation and classification.

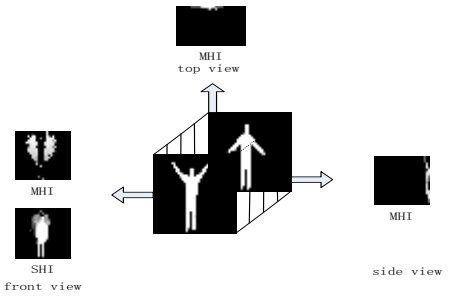

Figure 1. The Framework of Our Method

3.1. Feature Extraction

There are two main segments in this part: motion history image (SMHI) and static history image (SHI) templates for action representation, and Local binary pattern(LBPs) extracted from the template.

3.1.1. Motion History Images and Static History Images

Motion history image (MHI) and motion energy image (MEI) templates proposed by Bobick and Davis [5] describe where the motion happens and how the object moves, which presents the motion history from stacking the action sequence into a single gray scale image and preserving dominant motion information. Therefore, the MHI is not so sensitive to silhouette noises, like holes, shadows, and missing parts. However, the traditional MHI method has the limitation of scalability because only lateral motion of the action is analyzed. Human activities are performed in 3D space, which means MHI performed in2D space may miss some motion information of the action performed in the real world. In order to make full use of body shapes and motion information from depth maps, each depth frame is projected onto three orthogonal Cartesian planes. So each depth image generates three 2D maps, which is front, side and top view, respectively. To each projected map, we obtain its motion history image by computing the absolute difference between two consecutive maps and calculating its sum. The motion energy is obtained by accumulating summations of non-zero elements of MHI.

Static history image (SHI) presents the static posture history by retaining the static part of body. As an action performs, the body has both parts of the movement and the static. When computing the differences, the stationary parts and the moving parts are preserved simultaneously. It also contains the motion information which helps accurately identify the action. Figure.2 shows the MHI and SHI generated from the front view of the action Two Hand Wave.

The motion update function $\varphi_{\mathrm{m}}(x,y,t)$ and the static update function $\varphi_{\mathrm{s}}(x,y,t)$ are defined to represent the regions of motion information and static posture with action performs [23]:

$$\varphi_{\mathrm{m}}(x,y,t)=\begin{cases}1 & if\ D_t>\varepsilon_m\\ 0 & else\end{cases} \tag{1}$$

$$\varphi_{\mathrm{s}}(x,y,t)=\begin{cases}1 & if\ I_t-D_t>\varepsilon_s\\ 0 & else\end{cases} \tag{2}$$

where x, y and ε_{s} represent pixel position coordinates and time. $D_{\mathrm{t}}(t\in(1,T))$ is an absolute difference between two frames. ε_m is the motion threshold and ε_{s} is the static threshold, and we empirically set $\varepsilon_m=15$ and $\varepsilon_{\mathrm{s}}=50$ in our experiments. This can makes us to filter out more useless information, which is caused by a camera shake or a non-regular wobble of the human body. T is the total number of frames in actions.

The MHI $H_M(x,y,t)$ can be generated by using motion update function $\varphi_{\mathrm{m}}(x,y,t)$:

$$H_M(x,y,t)=\begin{cases}T & if\quad \varphi_m(x,y,t)=1\\ H_M(x,y,t-1)-1 & else\end{cases} \tag{3}$$

What's more, the SHI $H_S(x,y,t)$ can be obtained in the similar way as MHI:

$$H_S(x,y,t)=\begin{cases} T & if \quad \varphi_s(x,y,t)=1 \\ H_S(x,y,t-1)-1 & else \end{cases} \tag{4}$$

In 3DMTM-PHOG [23], they used motion history images (MHI), static history images (SHI), average motion images (AMI) and average static posture image (ASI) to represent depth sequence. In the experiment, we find that the AMI and ASI were not conductive to recognize action. When we use the AMI and ASI, the recognition rate can be reduced in some parameter settings. In our method, we only use MHI and SHI to represent the action sequences. After our optimization, the MHI and SHI can capture more information. And our feature has a better discriminative power and a smaller size.

Figure 2. (a) MHI and (b) SHI from Front Side of One Sample Action

3.1.2. Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) was originally proposed by T. Ojala [8] and applied in texture analysis, which can describe local structures in a simple but powerful way. The most important features of LBP are its tolerance regarding illumination changes. In recent years, LBP is widely used in image processing and pattern recognition. The LBP operator labels the pixels of an image with decimal numbers that encode the local structure around each pixel. The formation of LBP is shown in Figure 3.In order to create LBP, each pixel is compared with its eight neighbors in a 3×3 neighborhood by subtracting the threshold (center pixel value).The resulting strictly negative values are encoded with 0, and the positive values are encoded with 1.Thus, a binary number are obtained by concatenating all these binary values in a clockwise direction, which starts from the one of its top-left neighbor. The corresponding decimal value of the generated binary number is then used for labeling the given pixel. The pixel values are bilinear interpolated whenever the sampling point is not in the center of a pixel.

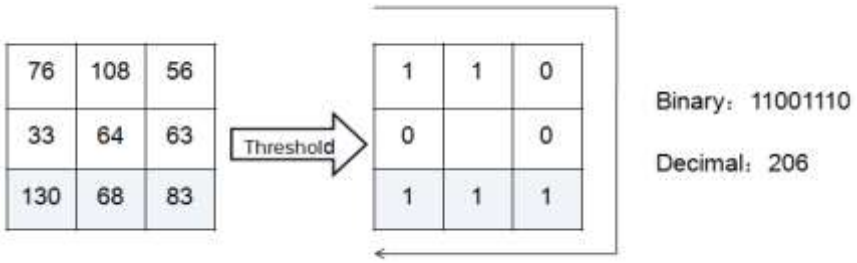

Figure 3. Example of the Basic LBP Operator

As then 3×3 neighborhood is too small to capture dominant features with large-scale structures. The operator was extended to use neighborhoods of different sizes [8]. A local neighborhood set is defined as a set of sampling points evenly spaced on a circle, which is centered at the pixel to be labeled. If the sampling points do not fall within the pixels, the bilinear interpolation was used to estimate the values of these points. Thus, we can obtain any radius and any number of sampling points in the neighborhood set.Figure4 shows some examples of the extended LBP operator. For neighborhoods set we use the notation (P,R) which means Psampling points on a circle ofradius of R.

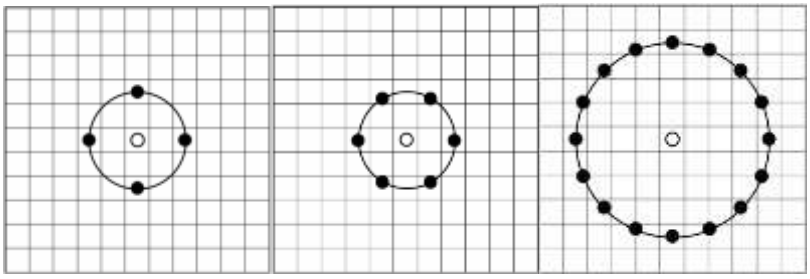

Figure 4. Three Circular Neighborhoods: (4, 1), (6, 1), (16, 2)

Formally, given a pixel g_c in an image, the LBP can be obtained in decimal form as follows:

$$LBP_{P,R}(g_c)=\sum_{p=0}^{P-1} f(g_p-g_c)2^p \tag{5}$$

Where g_p are pixel values of central pixel and P surrounding pixels in the circle neighborhood with a radius R. The function $f(x)$ is defined as:

$$f(x)=\begin{cases}1 & if\ x\geq 0\\ 0 & else\end{cases} \tag{6}$$

Under the definition, the LBP operator is invariant to monotonic gray-scale transformations, which preserve pixel intensity order in the local neighborhoods set. The LBP operator $LBP_{P,R}$ produces 2^p different values, corresponding to 2^p different binary patterns formed by P pixels in the neighborhood. A subset of these patterns named uniform patterns is able to describe image texture [8]. The histogram of LBP labels calculated over a region can be used as a texture descriptor. After obtain the MHI and SHI templates for the action sequence, the LBP operator is applied to overlapped blocks of the templates, this can make full use of texture information. The LBP histograms of the blocks for each template are concatenated to form the feature vector.

3.2. Feature Representation and Classification

In this paper, we use 4 templates to extract feature. For the front view, we use MHI and SHI. Considering the motion information is mainly concentrated on the front view, only MHI is used from the projections onto top view and side view. So, each action sequence can be modeled as four templates.

For feature extraction, we enforce the largest bounding box of the human body and resize the different action templates to the same size (fixed size). This step can reduce the intra-class variation and computational cost. The fixed size of each template was set to the

$1/2$ mean value of all of the bounding box sizes. The sizes of MHI_F, SHI_F, MHI_S, and MHI_T is 97×54, 97×76 and 76×54,respectively. The block sizes of the templates was sized to 20×27, 20×27, 20×25 and 25×27.The overlapped size between two blocks was set to the $1/2$ block size. Thus, we can obtained 27 blocks for MHI_F, 27 blocks for SHI_F, 45 blocks for MHI_S and 15 blocks for MHI_T.The number of total blocks are 114. Note that the dimensionality of the LBP histogram feature [8] based on uniform patterns is $P(P-1)+3$, so the dimensionality of the total feature $(P(P-1)+3)\times114$. In this paper, we use $R=1$ and $P=6$, so the dimensionality of total feature is $D=3762$. And we also considering the different parameter set (P,R) of the LBP features. The detail recognition result with different parameter settings is presented in Section IV.

After we get the total features, we use support vector machine (SVM) [6] to classify the actions. The SVM is widely used in pattern recognition and computer vision, due to its strong discriminative power. In our experiments, the optimal parameters of the Radial Basis Function (RBF) kernel are obtained by 5-fold cross-validation procedure over the training actions.

And in [23], the Pyramid Histograms of Oriented Gradient (PHOG) was used to encode human figures. In fact, it extracts the HOG features in three dimensions. There is no effect on the recognition results; if we only extract HOG one times. And the feature dimension is increase, and the computational cost is also increase.

4. Experiments

In this section, we evaluated the proposed method on the MSR Action3D dataset [11]. We also list the experimental results of different parameter settings. In all experiments, we select the MHI-LBP and SHI-LBP features with 90% principal components for PCA [4]. The experimental results show that our algorithm significantly outperforms the state of the art methods on this dataset.

4.1. MSR-Action3D Dataset

MSR-Action3D dataset [11] is a set of depth videos captured by a Kinect device. The dataset contains 567 depth sequences and 20 action types: 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. Ten subjects perform each action two or three times. It provides two sources of data: depth sequences at 15 frames per second with resolution of 320×240 and skeleton joint positions in each frame. Some examples of the dataset are shown in Figure5.

All the 20 actions were selected to cover various movements of arms, legs, torso and their combinations. Once the action is performed by a single arm or leg, the subjects were required to use their right arm or leg. Although the background of the dataset is very clean, the dataset is still challenging as many of the actions in the dataset are highly similar to each other.

Commonly, the dataset is divided into three actions [11] subsets and each having 8 actions as shown in Table 1. All the subsets (AS1, AS2 and AS3) are intended to group similar actions. For each subset, there are three different tests model: Test One (T1), Test Two (T2), and Cross Subject Test (CST). In Test One, 1/3 of the subset is used as training and the rest as testing; In Test Two, 2/3 of the subset is used as training and the rest as

testing; In Cross Subject Test, subject 1,3,5,7,9 are used for training and 2,4,6,8,10 are used for testing.

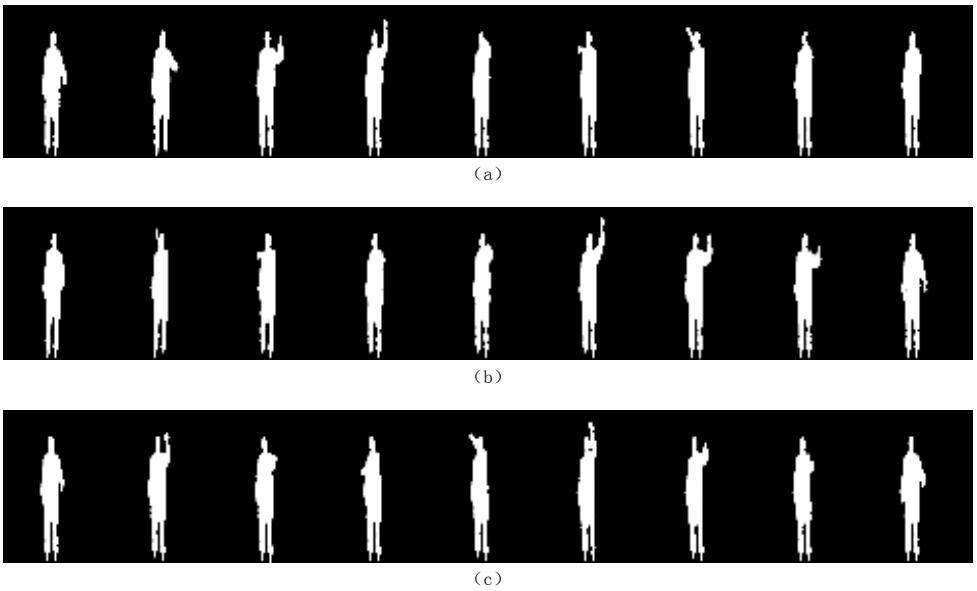

Figure 5. Sample Action Frames: (a) Draw X, (b) Draw Tick and (c) Draw C

Table 1. The Three Action Subsets of MSR Action3D Dataset

Action Set 1(AS1)	Action Set 2(AS2)	Action Set 3(AS3)
Horizontal arm wave (HoW)	High arm wave (HiW)	High throw (HT)
Hammer (H)	Hand catch (HC)	Forward kick (FK)
Forward punch (FP)	Draw x (Dx)	Side kick (SK)
High throw (HT)	Draw tick (DT)	Jogging (J)
Hand clap (HC)	Draw circle (DC)	Tennis swing (TSw)
Bend (B)	Two hand wave (THW)	Tennis serve (TSr)
Tennis serve (TSr)	Froward kick (FK)	Golf swing (GW)
Pickup & throw (PT)	Side boxing (SB)	Pickup & throw (PT)

4.2. Various Parameter Settings

We present the various parameter sets in this part. The average recognition accuracies associated with different parameter sets are shown in Table.2. According the recognition rate, we can choose the best P and R values, namely P=6, R=1.

Table 2. Recognition Accuracy (%) with different Parameters of LBP Operator

R \	P=4	P=6	P=8	P=10
R=1	94.37	97.68	96.14	96.37
R=2	95.97	96.85	96.85	96.22
R=3	95.92	96.47	95.98	96.31

4.3. Comparisons

The performance of SMHI-HOG and SHI-HOG in terms of accuracies on all tests is shown in TABEL 3. And when compared with other methods on the three subsets, our method also provide the overall accuracies for each test. As shown in Table 3, the performance of our method is superior to other methods. The approach [11] uses a bag of 3D points to characterize a set of salient postures based on the original depth maps. The Histograms of 3D Joints [12] and EigenJoints [10] mainly depends on the accurate estimation of the joints positions, so it cannot achieve a good recognition rates. The 3DMTM-PHOG [23] also used motion history images (MHI), static history images (SHI), average motion images (AMI) and average static posture image (ASI) to recognition action. In his method, the AMI and ASI is no help to recognition but increased computational costs. In our method, we only use MHI and SHI and reached the same recognition rate in Test One and Test Two.

We also compare our proposed method with other methods on the Cross Subject Test in Table 4. The proposed method achieves an accuracy of 95.2% which significantly outperforms the existing methods. The DSTIP [18] proposed a spatiotemporal interest point detector and a depth cuboids similarity descriptor to recognize actions. It can effectively eliminate the noise in the depth maps but it is very complex, and the accuracy of which is 89.3%. The proposed method outperforms 3DMTM-PHOG [23] by 4.5%, though both methods are based upon motion history images and static history images. Especially in Cross Subject Test AS2, we can achieve a recognition rate of 91.2%, which can outperform the method by 9%.

Table 3. The Performance Evaluation of Proposed Method on Three Subsets

Method (%)	Test One				Test Two				Cross Subject Test			
	AS 1	AS 2	AS 3	Over all	AS 1	AS 2	AS 3	Over all	AS 1	AS 2	AS 3	Over all
Bag of 3D Points [11]	89.5	89.0	96.3	91.6	93.4	92.9	96.3	94.2	72.9	71.9	79.2	74.7
HOJ3D [12]	98.5	96.7	93.5	96.2	98.6	97.9	94.9	97.2	87.9	85.5	63.5	79.0
EigenJoints [10]	94.7	95.6	97.3	95.8	97.3	98.7	97.3	97.8	74.5	76.1	96.4	82.3
3DMTM-PHOG[23]	97.3	97.4	98.7	97.8	100.0	100.0	100.0	100.0	93.4	82.3	96.4	90.7
Our method	98	97.4	98	97.8	100.0	100.0	100.0	100.0	99.1	91.2	95.5	95.2

What' more, the confusion matrices of our method on Cross Subject Test are shown in Figure 5.

Table 4. Evaluation of Method on the Cross Subject Test

Method	Accuracy (%)
Bag of 3D Points [11]	74.70
HOJ3D [12]	79.00
EigenJoints [10]	82.30
STOP [24]	84.80
Random Occupancy Pattern [25]	86.50
Actionlet Ensemble [14]	88.20
HON4D [15]	88.89
DSTIP [18]	89.30
Pose Set [26]	90.00
3DMTM-PHOG[23]	90.70
Our method	95. 2

The confusion matrices of our method on Cross Subject Test are shown in Figure6. In AS1, only one action was confused; the action Tennis serves (TSr) was classified into the action Pick up& throws (PT). In AS2, most actions were confused because the action in this subset with high similarity; Hand catch (HC) were confused with Draw x (Dx), Draw tick (DT) and Froward kick (FK); Draw x (Dx) were confused with Draw tick (DT) and Draw circle (DC), as they have highly similar movements; Tennis swing (TSw) was confused with Side boxing (SB).In AS3, Side kick (SK) were confused with Jogging (J)and Golf swing (GW); Tennis serves (TSr) and Tennis swing (TSw) was confused with each other.

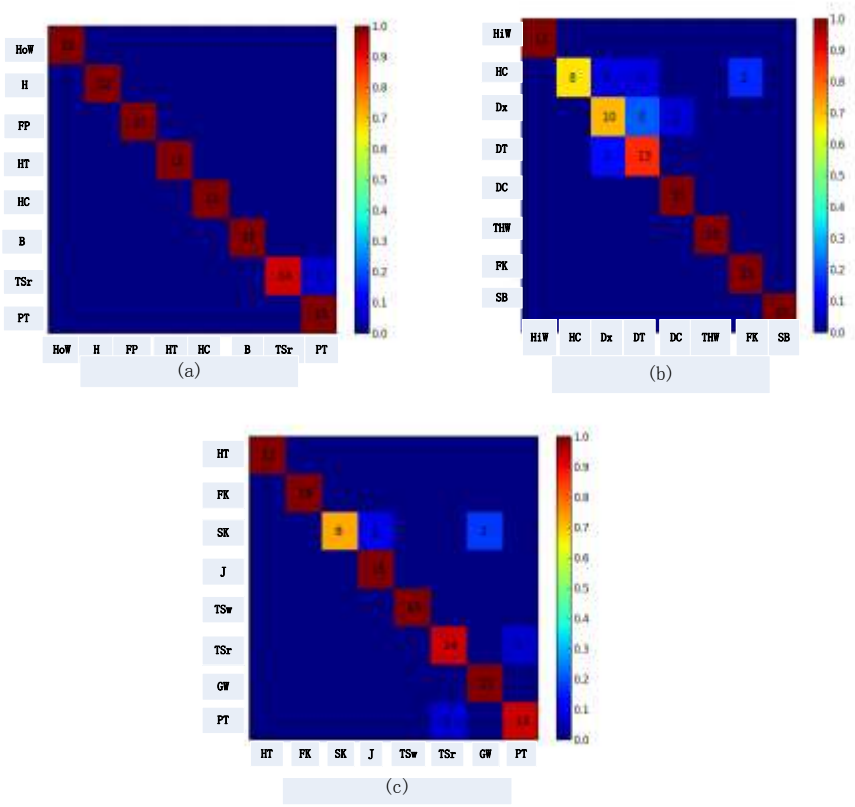

Figure 6. Confusion Matrices of our Method: (a) AS1, (b) AS2, and (c) AS3 on Cross Subject Test

5. Conclusion

This paper puts forward an effective and robust method to perform action recognition based on depth sequences. We use the motion history images (MHI) and static history images (SHI) to represent action sequences; due to the template can represent the actions in a compact and discriminative way. The MHI is able to capture the motion information from front/side/top view, so it can make full use of motion information. The SHI is obtained by stacking the static posture information. In order to make our feature has a better discriminative power; we further extract the Local Binary Patterns (LBP) feature for each template. Then, we use PCA to adopt the descriptor onto its principal subspaces to reduce the redundancy and computational cost. The experimental results on MSR Action3D dataset demonstrate the effectiveness and robustness of the proposed method. On Test two, we can achieve 100% recognition accuracy, and on the most challenging Cross Subject Test, the recognition rate is 95.2, which significantly outperforms the existing methods.

Acknowledgements

This work was supported partly by the National Natural Science Foundation of China under Grants no. 61331021, no.61210005and no. 61201251.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, and M. Finocchio, “Real-time human pose recognition in parts from single depth images”, In: Proceding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **(2003)**, pp. 1297-1304.

[2] J. Aggarwal, and M. S. Ryoo, “Human activity analysis: A review”, In ACM Computing Surveys (CSUR), vol. 43, no. 3, **(2011)**, pp. 16.

[3] A. Janoch, S. Karayev, Y. Jia, and J. T. Barron, “A category-level 3d object dataset: Putting the kinect to work”, In Consumer Depth Cameras for Computer Vision. Springer, **(2013)**, pp. 141-165.

[4] J. Shlens, “A tutorial on principal component analysis: Systems Neurobiology Laboratory”, In University of California at San Diego, **(2005)**.

[5] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates”, In Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 3, **(2001)**, pp. 257-267.

[6] C. C. Chang and C. J. Lin, “Lib-svm: a library for support vector machines”, In ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, **(2011)**, p. 27.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning Realistic Human Actions from Movies”, In CVPR, **(2008)**.

[8] T. Ojala, M. Pietik¨ainen, and T. Maenpaa, “Multi-resolution gray-scale and rotation invariant texture classification with local binary patterns”, In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, **(2002)**, pp. 971-987.

[9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio temporal features”, In 2nd Joint IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,**(2005)**, pp. 65-72.

[10] X. D. Yang, and Y. L. Tian, “Effective 3d action recognition using EigenJoints”, J. Vis. Commun. Image Represent, vol. 25, **(2014)**, pp. 2-11.

[11] W. Q. Li, Z. Y. Zhang and Z. C. Liu, “Action recognition based on a bag of 3d points”, IEEE Conf. on Computer Vision and Pattern Recognition Workshops, **(2010)**, pp. 9-14.

[12] L. Xia, C. Chen, and J. Aggarwal, “View invariant human action recognition using histograms of 3d joints”, In IEEE Conf. on Computer Vision and Pattern Recognition Workshops, **(2012)**, pp. 20-27.

[13] X. D. Yang, C. Y. Zhang and Y. L. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients”, In Proceeding of the 20th ACM Int'l Conf. on Multimedia, **(2012)**, pp. 1057–1060.

[14] J. Wang, Z. Liu, Y. Wu, and J. S. Yuan, “Mining action-let ensemble for action recognition with depth cameras”, IEEE Conf. on Computer Vision and, Pattern Recognition, **(2012)**, pp. 1290-1297.

[15] O. Oreifej, and Z. Liu, “HON4D: Histogram of Oriented 4DNormals for Activity Recognition from Depth Sequences”, In CVPR, **(2012)**.

[16] J. Sung, C. Ponce, B. Selman, and A. Saxena, “Unstructured human activity detection from rgb-d images”, In IEEE Int'l Conf. on Robotics and Automation, **(2012)**, pp. 842-849.

[17] Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition", In IEEERSJ Int'l Conf. on Intelligent Robots and Systems, **(2011)**, pp. 2044-2049.
[18] L. Xia, and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera", In IEEE Conf. On Computer Vision and Pattern Recognition, **(2013)**, pp. 2834-2841.
[19] J. Luo, and W. Wang, "Spatio-temporal feature extraction and representation for RGB-D human action recognition", In Pattern Recognition Letters, vol. 50, **(2014)**, pp. 139-148.
[20] C. Liu, Y. Yang, and Y. Chen, "Human action recognition using sparse representation", In Proceeding IEEE International Conference on Intelligent Computing and Intelligent Systems, **(2009)**, 184-188.
[21] Azary, and Savakis, "3D action classification using sparse spatio-temporal feature representations", In Advances in visual computing, **(2012)**.
[22] J. Zheng, Z. Jiang, J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair", In BMC, **(2012)**.
[23] B. Liang, and L. Zhang, "3D Motion Trail Model based Pyramid Histograms of Oriented Gradient for Action Recognition", In 22nd International Conference on Pattern Recognition, **(2014)**.
[24] A. Vieira, E. Nascimento, Z. Liu, and M. Campos, "STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences", in CIARP, **(2012)**.
[25] J. Wang, Z. Liu, J. Chorowski, and Y. Wu, "Robust3D Action Recognition with Random Occupancy Patterns", In ECCV, **(2012)**.
[26] C. Wang, Y. Wang, and A. Yuille, "An Approach to Pose based Action Recognition", In CVPR, **(2013)**.
[27] Y. Zhu, and W. Chen, "Evaluating spatiotemporal interest point features for depth-based action recognition", In Image and Vision Computing, vol. 32, **(2014)**, pp. 453-464.

Authors

Shichao Zhang, received his bachelor degree in Jiangxi University of Science and Technology, China, in 2013. He is currently studying for a postgraduate in Zhengzhou University. His research interests include human-computer interaction, pattern recognition and computer vision.

Enqing Chen, is an associate professor at the Zhengzhou University. He received his PhD degree in communication and information system engineering from Beijing Institute of Technology in 1991. His current research interests include signal and information processing technology, wireless sensor network technology and application, multimedia information processing and identifying. He is a member of IEEE.

