Action Recognition Based on Sub-action Motion History Image and Static History Image

Shichao Zhang ,Enqing Chen ,Lin Qi , Chengwu Liang

School of Information Engineering, Zhengzhou University, Zhengzhou 450000, China

Abstract. 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 proposition of the Sub-action Motion History Image (SMHI) and Static History Image (SHI) in a depth sequence. We evenly subdivide the normalized motion energy into a set of segments which corresponding frame indices are used to partition a video into different sub-actions segments. The Local Binary Patterns (LBP) descriptor is then computed from the SMHI and SHI for the representation of an action. We evaluate the proposed framework on MSR Action3D dataset. Experimental results indicate that the proposed approach outperforms the most of the art methods and demonstrate the effectiveness of the proposed approaches.

1 INTRODUCTIONS

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. While significant efforts, 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 RGB images in activity recognition. First, depth cameras 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 illumination changes, which bring great benefits to the activity recognition.

In this paper, we propose an effective and robust approach to recognize actions by extracting Local Binary Patterns (LBP) descriptors from Sub-action Motion History Images (SMHI) and Static History Images (SHI). The SMHI are generated by stacking sub-action motion energy of depth maps projected onto three orthogonal Cartesian planes. We propose an adaptive sub-action segment based on motion energy. The stacked motion energy of each sub-action represents particular appearances and shapes, which can be used to recognize corresponding actions. The SHI are obtained by accumulating static information simultaneously, which presents the static posture history by retaining the static part of body. It also contains the motion information which helps accurately identify the action.

The rest of the paper is organized as follows. Section II discusses the related work on depth-based action recognition. Section III, we describe the detailed procedures of the computing the feature descriptor. Section Ⅳ is the experiment and comparison. Finally, the conclusion of this paper is presented in Section Ⅴ.

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. [2] sample a bag of 3D points from the depth maps 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. What's more, the sampling scheme is view dependent and more accurate than using 2D silhouettes. Xia et al. [3] 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 Discriminate 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 translates each action into a series of symbols. After that, a discrete HMM classifier used for action recognition. Yang et al. [4] employ a HOG [5] 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.

As the spatial-temporal interest points can provide a compact representation of the image content by describing local parts of the scene, so it can be robust to clutter, occlusions, and intra-class variations. Recently, many methods are proposed for action recognition based on spatial-temporal interest points. Zhang et al. [6] 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.

The Sparse Representations 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. Zheng et al. [7] proposed a dictionary learning framework for cross-view action recognition with the assumption that sparse representations of videos from different views of the same action should be strictly equal. However, this assumption is too strong to flexibly model the relationship between different views.

3 PROPOSED METHOD

This section offers a detailed description for human action recognition from depth maps. The framework is demonstrated in Figure.1, there are two main segments in this part: sub-action motion history image (SMHI) and static history image (SHI) for action representation, and feature descriptor.

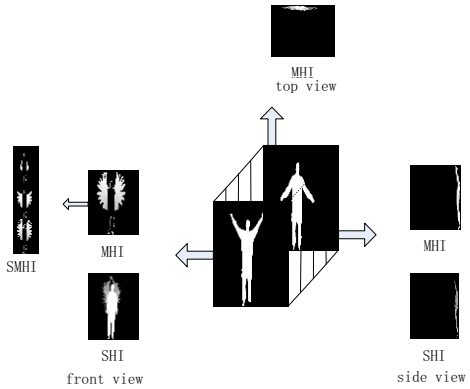

Figure 1. The framework of the proposed method.

3.1. Sub-action Motion History Images and Static History Images

Motion history image (MHI) and motion energy image (MEI) templates proposed by Bobick and Davis [8] 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. However, the traditional MHI method has the limitation of scalability because only lateral motion of the action is analysed. Human activities are performed in 3D space, which means MHI performed in 2D 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, just like Liang [9]. So each depth image generates three 2D maps, which is front, side and top view, respectively. On 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.

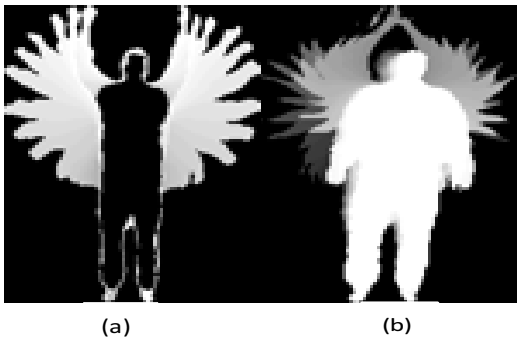

Figure 2. (a) MHI and (b) SHI from front side of one sample action

The motion update function $\varphi_m(x,y,t)$ and the static update function $\varphi_s(x,y,t)$ are defined to represent the regions of motion information and static posture with action performing [9]:

$$\varphi_M(x,y,t)=\begin{cases}1 & if\ D_t>\varepsilon_M\\ 0 & otherwise\end{cases} \quad (1)$$

$$\varphi_S(x,y,t)=\begin{cases}1 & if\ I_t-D_t>\varepsilon_S\\ 0 & otherwise\end{cases} \quad (2)$$

where x, y and t represent pixel position coordinates and time. $I_t(t\in(1,t))$ is a depth sequence and $D_t(t\in(1,t))$ is a 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 $\varepsilon_S=50$ in our experiments. T is the total number of frames in actions.

The motion history image (MHI) $H_M(x,y,t)$ and static history image (SHI) $H_S(x,y,t)$ can be generated by using motion update functio $\varphi_m(x,y,t)$ and $\varphi_s(x,y,t)$:

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

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

Additionally, to different people, they could have varied motion speed or frequency when they are required to perform the same activity. For the same action, different people to complete may cause intra-class variations increase; there is no help in classification. Just like the action Draw circle, the second image and the seventh image have the same posture (in fact, it has a little same action sequence). So, condensing the whole action sequence into a single image may lose some detailed information. It is not beneficial to recognize the actions. In order to handle this difficulty, we propose an adaptive sub-action segments based on the motion energy. To divide the whole action sequence into different sub-actions, the key information of each action fragment can be preserved, and the intra-class variations also can be reduced. For a depth sequence , we first obtain the projected maps $\left(I_1^i, I_2^i, I_3^i\right), i=(1,\cdots,T)$ according to front, side, and top views, i.e. I_1, I_2, I_3 . After then, we compute the motion energy by accumulating summations of non-zero elements value of each MHI as:

$$E(p,t)=sum(H_M(x,y,t)) \quad (5)$$

Where $E(p,t), p=(1,2,3)$ is the motion energy of the each MHI which obtained by three projected maps at t moment; $sum(\cdot)$ represents the summations of non-zero elements value of each MHI.

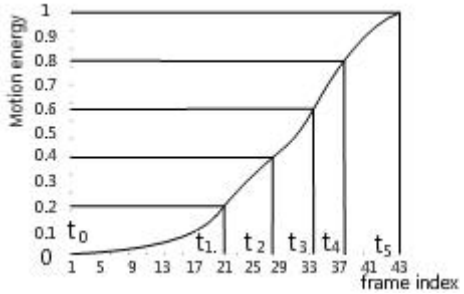

Figure 3. The sample of Sub-action segments.

Laptev et al. [10] use the temporal pyramid to take into account the temporal order of a video. Inspired by this, we propose adaptive sub-action segments approach as shown in Fig.3. We evenly subdivide the normalized motion energy into a set of segments. The different energy segments corresponding frame indices are used to partition a video into different sub-actions segments. The definition of the sub motion segmentation is shown as follows:

$$E(p,t)/E(p,T)=N/M \quad (6)$$

$E(p,t), p=(1,2,3)$ is the motion energy of MHI at t moment and $E(p,T), p=(1,2,3)$ is the motion energy of the whole MHI; M represents to partition a video into M sub-actions segments; $N=(1,2,\cdots,M-1)$.

In this paper, we divide the front view into 3 sub-action segments to form respective motion history image. Which means $E(p,t_1)/E(p,T)=1/3$, $E(p,t_2)/E(p,T)=2/3$, and $E(p,t_3)/E(p,T)=3/3$. And we use the five different t moment to partition a video into 3 sub-actions, which is t_0t_1, t_0t_2, t_0t_3 , respectively; t_0 represent the opening time and T is the total number of frames.

Considering the motion information is mainly concentrated on the front view. For the front view, we use a 3 sub-action MHI and SHI. For the side view, we use MHI and SHI, respectively. For the top view, we only use MHI. Thus, each action depth sequence can be modelled as a seven templates.

3.2 Feature Descriptor and Recognition

In order to make our feature has a better discriminative power; we further extract the LBP feature [11] for each template. The LBP operator labels the pixels of an image with decimal numbers that encode the local structure around each pixel. A subset of these patterns named uniform patterns is able to describe image texture. The histogram of LBP labels calculated over a region can be used as a texture descriptor. Since seven templates are used, 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. The dimensionality of total feature is D=11151. It is relatively high and not conducive to classification. We employ the widely used principal component analysis (PCA) due to its effectiveness and simplicity.

After feature extraction, we use support vector machine (SVM) to classify the actions, due to its strong discriminative power.

Table 1. Evaluation of method on the on three subsets.

Method (%)	Test One				Test Two				Cross Subject Test			
	AS1	AS2	AS3	Overall	AS1	AS2	AS3	Overall	AS1	AS2	AS3	Overall
Bag of 3D Points [2]	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 [3]	98.5	96.7	93.5	96.2	98.6	97.9	94.9	97.2	87.9	85.5	63.5	79.0
3DMTM-PHOG [9]	97.3	97.4	98.7	97.8	100.0	100.0	100.0	100.0	93.4	82.3	96.4	90.7
EigenJoints [12]	94.7	95.6	97.3	97.3	97.3	98.7	97.3	97.8	74.5	76.1	96.4	82.3
Our method	97.3	96.7	98.7	97.6	100.0	100.0	100.0	100.0	95.3	90.3	95.5	93.7

4. EXPERIMENTS RESULTS

In this section, we evaluated the proposed method onthe MSR Action3D dataset [2]. In all experiments, we select the SMHI-HOG and SHI-HOG with 90% principal components for PCA. The empirical results show that the proposed method outperforms the most of the art methods. Especially in similarity actions, we can achieve a good recognition rate.

MSR-Action3D dataset [2] is a set of depth videos captured by a Kinect device. The dataset contains 567 depth sequences and 20 action types. Commonly, the dataset is divided into three actions subsets: AS1, AS2 and AS3. For each subset, there are three different tests model: Test One (T1), Test Two (T2), and Cross Subject Test (CST).

The performance of SMHI-HOG and SHI-HOG in terms of accuracies on all tests is shown in Table I. We also compare our proposed method with other methods on the Cross Subject Test in Table II. The proposed method achieves an accuracy of 93.7% which significantly outperforms the existing methods. The approach [2] uses a bag of 3D points to characterize a set of salient postures based on the original depth maps. The Histograms of 3D Joints [3] and EigenJoints [12] mainly depends on the accurate estimation of the joints positions, so it cannot achieve a good recognition rates. The DSTIP [13] 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 [9] by 3%, 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 90.3%, which can outperform the method by 8%. This may use the sub-action motion history images which can capture more key information. Using the motion history images directly, this key information may be covered by other coarse information. Thus, the proposed method is quite robust.

Table 2. Evaluation of method on the Cross Subject Test.

Method	Accuracy (%)
Bag of 3D Points [2]	74.70
HOJ3D [3]	79.00
EigenJoints Depth Motion Maps [4]	88.73
3DMTM-PHOG [9]	90.70
EigenJoints [12]	82.30
DSTIP [13]	89.30
Random Occupancy Pattern [14]	86.50
STOP [15]	84.80
Action Let Ensemble [16]	88.20
HON4D [17]	88.89
Pose Set [18]	90.00
Our method	93. 70

5 CONCLUSIONS

In this paper puts forward an effective and robust method to perform human action recognition on depth sequences. We propose the adaptive sub-action segments which can effectively preserve detailed action information. The SMHI and SHI can represent human actions in a discriminative way. It is able to capture the key motion information and static posture information. The Local Binary Patterns descriptor is used to encode the feature. To divide the whole action sequence into different sub-actions, the intra-class variations can be reduced; it is conducive to classification. The experimental results on MSR Action3D dataset demonstrate the effectiveness and robustness. Especially for the similarity actions, our method achieves a better recognition rate. And the recognition rate is better than most of methods. It is proves that our sub-action segmentation method is effective.

ACKNOWLEDGMENT

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

References

1. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, 2011. In: Proc. of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1297–1304.
2. W.Q. Li, Z.Y. Zhang, Z.C. Liu. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2010, pp. 9–14.
3. L. Xia, C. Chen, J. Aggarwal. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.
4. X.D. Yang, C.Y. Zhang, Y.L. Tian. Proc. of the 20th ACM Int'l Conf. on Multimedia,2012, pp. 1057–1060, (MM '12).
5. N. Dalaland, and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.
6. Zhang, L.E. Parker. 4-dimensional local spatial-temporal features for human activity recognition. IEEERSJ Int'l Conf. on Intelligent Robots and Systems, 2011, pp.2044–2049.
7. Zheng, Z. Jiang, J. Phillips, and R. Chellappa. Cross-view action recognition via a transferable dictionary pair. In BMVC, 2012.
8. A. F. Bobick and J. W. Davis. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(3):257–267, 2001.
9. B. Liang and L.H. Zhang, 3D Motion Trail Model based Pyramid Histograms of Oriented Gradient for Action Recognition. 22nd International Conference on Pattern Recognition 2014.
10. J. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In CVPR, 2008.
11. T. Ojala, M. Pietikaimen, and T. Maenpaa. Multi-resolution gray-scale and rotation invariant texture classifications on Pattern Analysis and Machine Intelligence, 24(7):971-987,(2002).1,3,5.

12. X.D. Yang, Y.L. Tian. Effective 3d action recognition using EigenJoints. J. Vis. Commun. Image Represent. 25 (2014) 2–11.
13. L. Xia, J. Aggarwal. Spatial-temporal depth cuboids similarity feature for activity recognition using depth camera. IEEE Conf. On Computer Vision and Pattern Recognition, 2013, pp, 2834-2841.
14. J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3D Action Recognition with Random Occupancy Patterns. In ECCV，2012.
15. A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. Campos. STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences. In CIARP，2012.
16. J. Wang, Z.C. Liu, Y. Wu, 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.
17. O. Oreifej and Z. Liu. HON4D: Histogram of Oriented 4DNormals for Activity Recognition from Depth Sequences. In CVPR, 2013.
18. C. Wang, Y. Wang and A. Yuille. An Approach to Pose based Action Recognition. In CVPR ， 2013.