

A novel gradient based method for RGB-D human action recognition

Haohao Tang^a, Hanling Zhang

College of Computer Science and Electronic Engineering Hunan University
Changsha, Hunan, China

^athh0120@163.com

Abstract. Most action recognition approaches proposed over the years that were designed for RGB sequences cannot utilize the rich 3D-structural information to reduce large intra-class variations. This paper addresses the issue of handling sole depth structural information for accurate human action recognition. It presents and evaluates the saliency based P-DmHOG features for human action representation. Saliency based P-DmHOG features are depth features inspired by the well-known HOG descriptor. Good results namely 96.78%, 97.78% and 93.13% are achieved on three public benchmark datasets: MSR Actions 3D, MSR Action Pairs 3D, and MSR Daily Activity 3D, which show the efficiency of proposed method to spatio-temporal and shape information.

1. Introduction

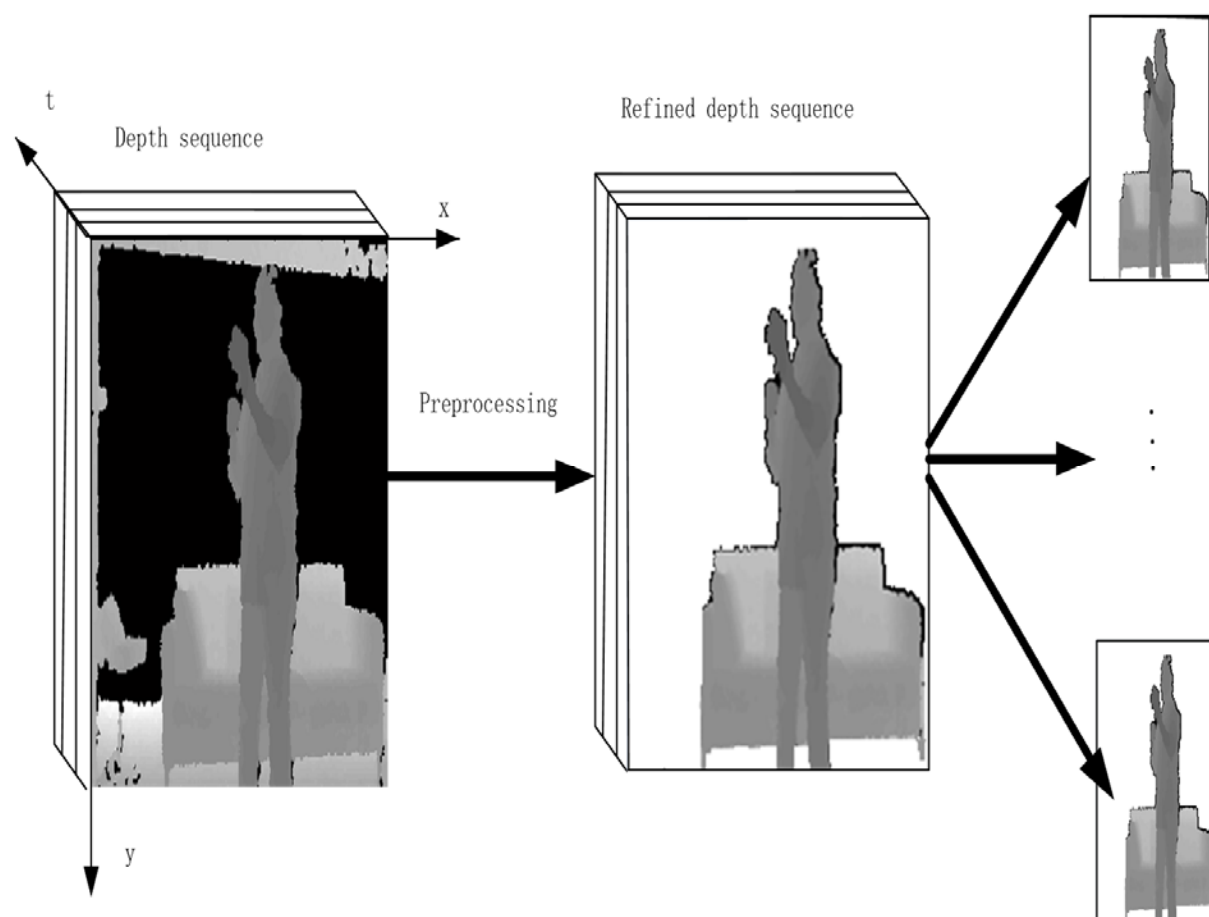
Human action recognition based on video information has been a core topic in computer applications and content-based video information analysis for decades [1], while it is still challenging due to 3D structure changes [2] and other common difficulties in video representation like cluttered backgrounds. Due to the recent advent of the cost-effective Kinect sensor, action recognition from RGB-D cameras has received an increasing interest throughout the computer vision community. As the sensor techniques advance, the recent emergence of low-cost depth sensors facilitates a variety of visual recognition tasks including human action recognition.

Compared with conventional RGB data, depth data is more robust to intensive light changes and color changes, since the depth value is estimated by infrared radiation and is not related to visible light. In addition to this, depth maps also have several advantages with respect to traditional color images in the context of action recognition. On the one hand, It provides additional body shape and structure information, which has been successfully applied to distinguish the actions. On the other hand, color and texture are precluded in depth maps, which makes the problems of human detection and action recognition easier than RGB images involved only. At the same time, Kinect also provide a technique for skeleton tracking, which are more compact than RGB or depth sequences. But some estimated 3D joint positions are noisy and may have significant errors when there are occlusions. On the other hand, although skeleton sequences can represent the temporal dynamics of actions well, the appearance and scene information is still missing.

In this paper, we propose a novel RGB-D human action recognition framework based upon the local and gradient information in depth maps only. The noises in depth data, which are common problems for



3D sensors, are not well handled. To address this problem, we detect saliency region of each depth map by the method mentioned in [14] to generate refined depth maps. Then, saliency based P-DmHOG features are learned on refined depth maps. Finally, to retain only the discriminant features, we train a random decision forest (RDF). The framework of our action recognition method is demonstrated in Fig. 1.



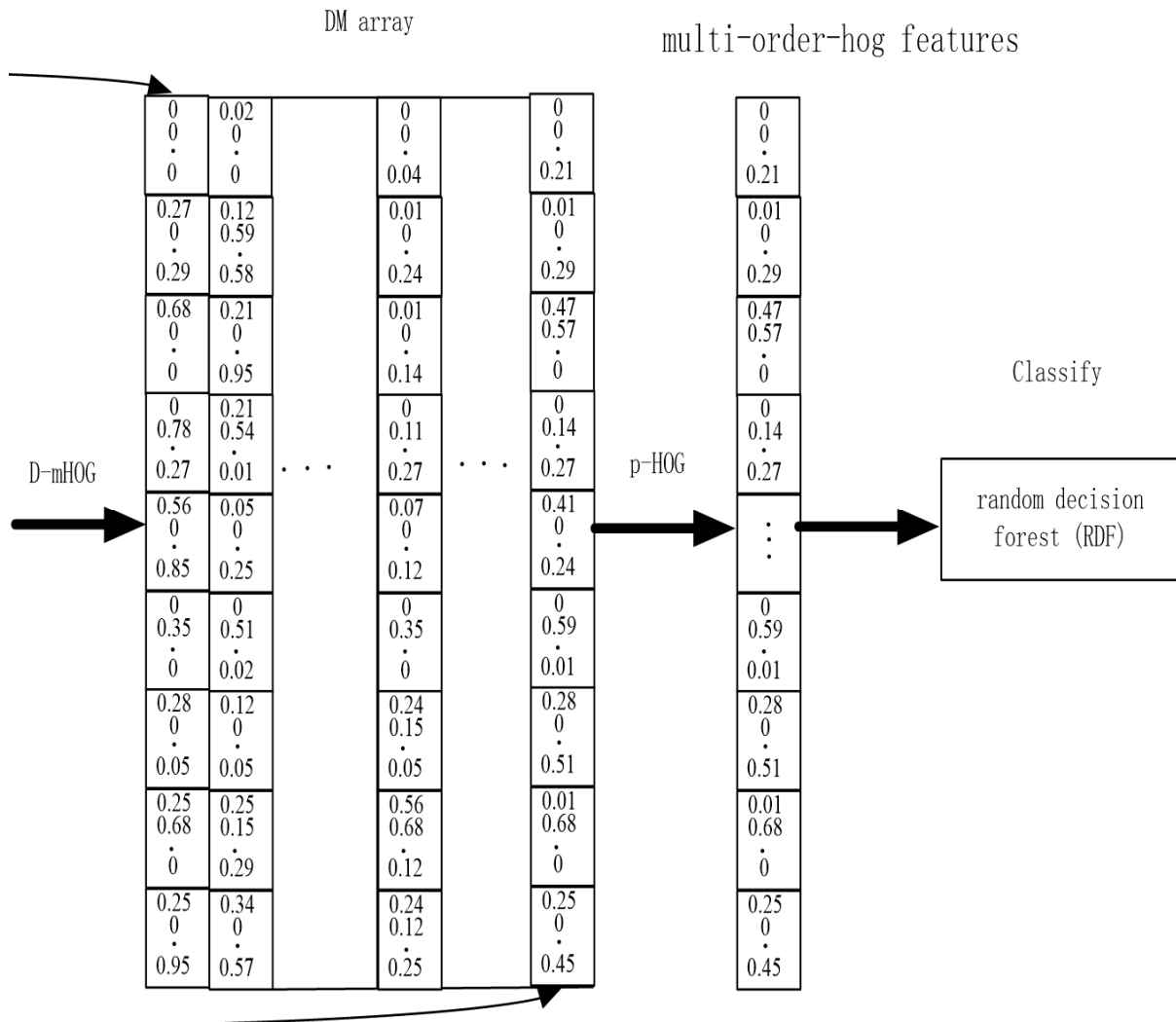


Figure 1. The framework of our system

2. Methods

2.1. Preprocessing of Depth Maps

The depth maps sequence $V^{N \times M \times T} = \{I^1, I^2, \dots, I^T\}$ ($N \times M$ is the size of depth map, T is the frames number of depth maps sequence) is able to provide pivotal body shape and motion information to distinguish actions, but also contain noises. Fig. 2 (a) illustrates the noise situation in depth map. From those images, we can know that action human is most important and salient in a scene for action recognition. Therefore, saliency detection method [14] is used to generate saliency map I_{sm}^t for it firstly. Then, we weight the initial depth map it with Q^t to obtain the refined depth map sequences $V_r^{N \times M \times T} = \{I_r^1, I_r^2, \dots, I_r^T\}$ using.

$$I_r^t = I^t Q^t \quad (1)$$

Q^t is the assuring constraint whose elements are 1 for certain pixels and 0 for all other pixels, and is given by:

$$Q^t(i, j) = \begin{cases} 1 & \text{if } (I_{sm}^t(i, j) > 0) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

If $I_{sm}^t(i, j) > 0$, we assume that the pixel $I^t(I, j)$ belongs to the salient region. Moreover, the remaining $I^t(I, j)$ belongs to the noise region. Some examples of refined depth maps are shown in Fig. 3 (b).

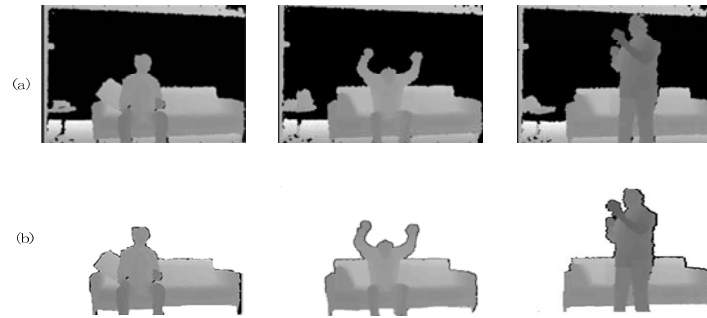


Figure 2. (a) The input depth maps. (b) The refined depth maps

2.2. Depth-mHOG

HOG is a spatial shape descriptor applied to image classification, human detection and human action recognition [9]. For a RGB image, HOG is proved to be effective in detecting silhouette contours in previous works. On the other hand, depth map is an image or image channel that contains information relating to the distance of the surfaces of scene objects from a viewpoint. Therefore, depth maps have sharper edges which arise due to the strong discontinuities at object boundaries. To sum up, HOG is also suitable for the depth map to extract local and gradient information.

In this paper, we propose a novel descriptor named Depth-modified HOG to extract shape and local information of each refined map. It represents the spatial distribution of action and is formulated as a vector representation. This descriptor is mainly inspired by two sources: (1) the use of Depth Gradients [10], and (2) the modified Histogram of Orientation Gradients (mHOG) the details of extracting D-mHOG descriptors are as follows.

Step 1: Extracting temporal Gradient map sequences. Giving a sequence of refined depth maps $V_r^{N \times M \times T} = \{ I_r^1, I_r^2, \dots, I_r^T \}$, temporal gradient maps are extracted firstly for further processing using Eq.(3).

$$I_{rt}^t = \frac{\partial V_r}{\partial t} = I_r^{t+1}(i, j) - I_r^t(i, j) \quad (3)$$

According to Eq. (3), the athletic parts receive high value and the static regions receive small value, so the change regions are highlighted while the static background (the sofa as shown in Fig. 2 (b)) regions are suppressed. Thus, the temporal gradient maps is a description of shape and temporal information in a more compact way.

Step 2: The modified HOG for temporal gradient maps and original input depth maps is computed. The modified HOG descriptor is created as follows: the gradient image of t^{th} temporal gradient map I_{rt}^t and t^{th} refined depth map I_{rt}^t are divided into n rectangular cells. A 50% overlap between the cells is used. Spatial gradients of the refined depth maps $V_r^{N \times M \times T}$ along x,y dimensions as.

$$\nabla V_r = \left(\frac{\partial V_r}{\partial x}, \frac{\partial V_r}{\partial y} \right) \quad (4)$$

$$\begin{cases} \frac{\partial V_r}{\partial x} = I'(i+1, j) - I'(i-1, j) \\ \frac{\partial V_r}{\partial y} = I'(i, j+1) - I'(i, j-1) \end{cases} \quad (5)$$

Where the gradient along each dimension is computed using the finite difference approximation. From Eq. (5), we can find that spatial gradients considers the current and the neighbors temporal information for each pixel in the temporal gradient map. Then the gradient magnitude g and the gradient orientation θ of x, t dimensions are computed for all the values in the cell using respectively Eq.(6) and Eq.(7).

$$g(i, j) = \sqrt{\frac{\partial V_r}{\partial t}^2 + \frac{\partial V_r}{\partial x}^2} \quad (6)$$

$$\theta(i, j) = \arctan \frac{\frac{\partial V_r}{\partial t}}{\frac{\partial V_r}{\partial x}} \quad (7)$$

Within each cell, an orientation histogram is generated by quantifying the angle of each gradient vector into m bins. Lastly, we normalize the descriptor by using L2-norm. The modified HOG of y, t dimensions is computed as the same. The final D-mHOG descriptor $dt \in K \times 1$ for t th temporal gradient map is a concatenation of all the HOG vectors from each cells.

Step 3: 2D array representation for action video. We collect the D-mHOG descriptors over time to form a 2D array $D^{K \times (T-1)} = \{d^1, d^2, \dots, d^{T-1}\}$ named depth-modified(DM) array (shown in Fig. 2). Changes in the DM array correspond to changes in the local shape and depth data.

2.3. P- HOG

In order to capture the temporal order and changes, the P-HOG [9] for DM array D is computed. Firstly, the array D is divided into cells at several pyramid level L . As shown in Fig. 3, the grid at level L has $2^{2(L-1)}$ cells. Then the HOG for each grid at each pyramid resolution level is computed. Local and entire temporal information is represented by a histogram of orientations within a sub region quantized into k bins. The final saliency based depth P-DmHOG for each DM array D is a concatenation of all the HOG vectors at each pyramid resolution. The concatenation of all the HOG vectors introduces the order and temporal information of action.

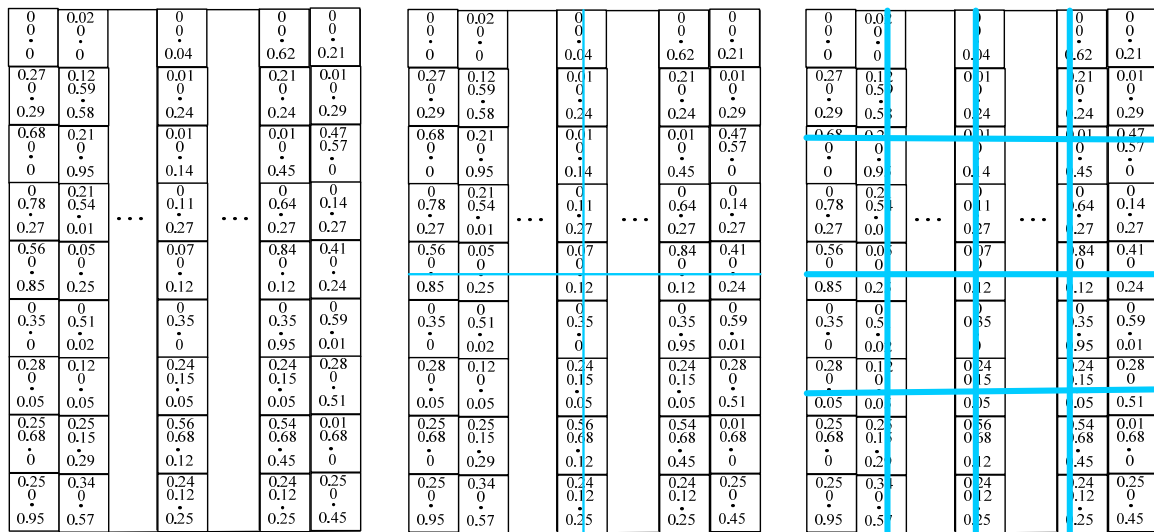


Figure 3. Grids at three pyramid resolution in the DM array

Such an approach bridges the spatial and temporal information extraction in a way that has implications to other fields in computer vision. Depth-mHOG stress the local and shape information between adjacent frames and introduce the DM array for further temporal and local information collection.

2.4. Random Decision Forests

In order to capture the temporal order and changes, the P-HOG [9] for the same feature, different classifiers [10, 16] result in different recognition accuracy. The human action recognition problem tackled in this paper fits well in the random decision forest (RDF) framework as the number of actions to be classified is quite large.

A random decision forest [10] is an ensemble of weak learners which are decision trees. For a given feature vector, each tree independently predicts its label and a majority voting scheme is used to predict the final label of the feature vector. The procedure of RDF consist of training part and test part. In training part, each decision tree is trained on a randomly selected 2/3 part of the training data. The remaining 1/3 of the training data are out-of-bag (OOB) samples and used for validation. After a set of trees have been trained, Feature pruning process identify discriminative part of feature and skipped the useless one. Thus, the lower dimensional feature vectors are fed to the trained random decision forest for classification.

3. Parameter setting

3.1. Preprocessing

In MSR Action 3D Dataset [15], the background is pre-processed to clear the discontinuities created from undefined depth regions. Hence, the denoising preprocessing is applied on MSR Action Pairs 3D Dataset [6] and MSRDailyActivity3D Dataset [4] only. Saliency detection process [14] is used in our method. From Fig.3, we can know that saliency detection based denoising achieves superior performances to tackle noises and cluttered backgrounds than using depth data from RGB-D sequence. It is because the noises may appear in the same depth distance as action body.

3.2. Parameters in D-mHOG

In D-mHOG part, there are two parameters in computing mHOG, cell number n and the bin number m . As Fig. 4 shows,

Increasing the number of bins improves performance significantly up to about 18. As the increasing of cell number, the recognition accuracy increase at first and then decrease. In order to balance calculation speed and recognition accuracy, we find that when we set $n=10$ and $m=18$, the performance is best.

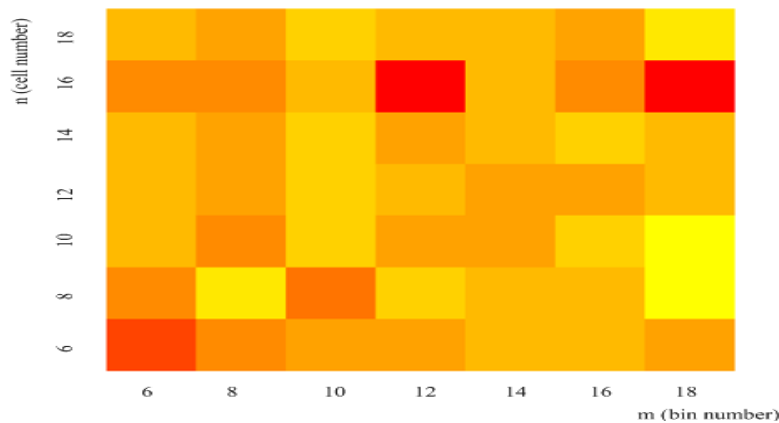


Figure 4. The effects of cell number n and bin number m in D-mHOG on MSR Action Pairs 3D Dataset. Color 'red' indicates that the accuracy is poor, color 'yellow' means the better one.

3.3. Parameter in P-HOG

In P-HOG processing procedure, the pyramid level L and bin number k impair the quality of recognition rate. Experimentally, when the number of L is small, some representative information may be discarded, while if the number of L is too big, the pyramid structure will result in a big increase in computation time. In order to balance the computing time, we find that when we set $L=3$, the performance is best. Fig. 5 (a) shows the results of pyramid level L on MSR Action Pairs 3D Dataset.

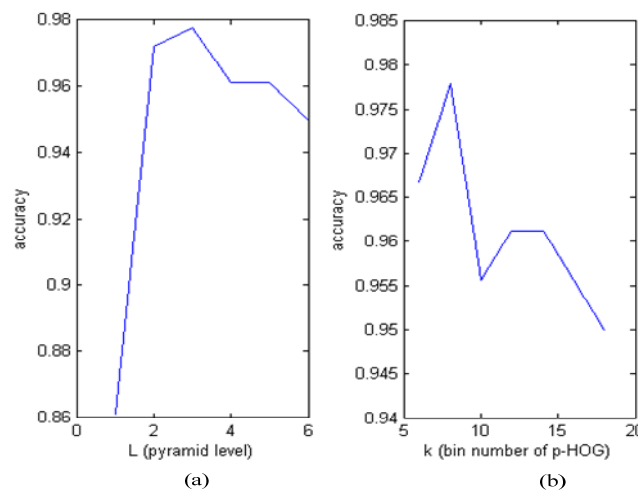


Figure 5. The effects of bin number k and pyramid level L in P-HOG on MSR Action Pairs 3D Dataset

The P-HOG descriptor is applied on the DM array using different parameter configurations. Fig. 5(b) shows the results on the dataset by choosing the various parameter k . Due to the pyramid structure of P-HOG, we discuss the effect of bin number k only. By varying this one parameter, we achieved good results on the dataset MSR Action Pairs 3D. The best results were using 8 bins at 97.78% (shown in Fig. 5(b)).

4. Experiments

We extensively experimented on the proposed ideas using three standard 3D activity datasets including MSR Actions 3D [15], MSR Action Pairs 3D Dataset [6], and MSR Daily Activity 3D [4]. In random decision forest, experiments are performed using five subjects(1,3,5,7,9) as training and five subjects(2,4,6,8,10) as test in [10]. The subjects for training also make a difference on recognition rate. But it makes little difference between combinations of various subject for training. For the sake of justice, like many action recognition models, we select five subjects (1, 3, 5, 7, 9) for training and others for test as in [10].

4.1. MSR Action 3D Dataset

MSR Action 3D dataset [15] is an action dataset of depth sequences captured by a depth camera. This dataset is still challenging as many activities appear very similar. This dataset including 20 action types performed by 10 subjects, each subject performs each action 2 or 3 times. There are 567 depth map sequences in total. We compare our method with 7 action recognition algorithms on MSRAction3D dataset, including Actionlet Ensemble [4], Depth Cuboid [5], HON4D [6], DMM-HOG [8], HOG2 [9], HDG [10] and SNV [11]. In order to make a fair evaluation, we directly use author-provided accuracy results. For HOG2 [9], we run the authors code. The recognition accuracy comparison is shown in Table 1, which indicate the different methods emphasizing the effect of extracted feature, and providing a fair evaluation of different algorithms. As we can see from Table 1, our method is outperformed the seven state-of-the-art methods even though.

Although HOG2 [9] uses HOG twice and is combined with skeleton information, our method obtains 96.73%, 1.89% superior to HOG2. This is because the temporal information in the pixels of each frame is added in our method. Fig. 6 shows a confusion matrix for MSR Action 3D dataset. In this confusion matrix, only two actions, 'high throw' and 'draw x', have more than two misjudgment situation due to these actions have some very similar action such as 'draw tick'.

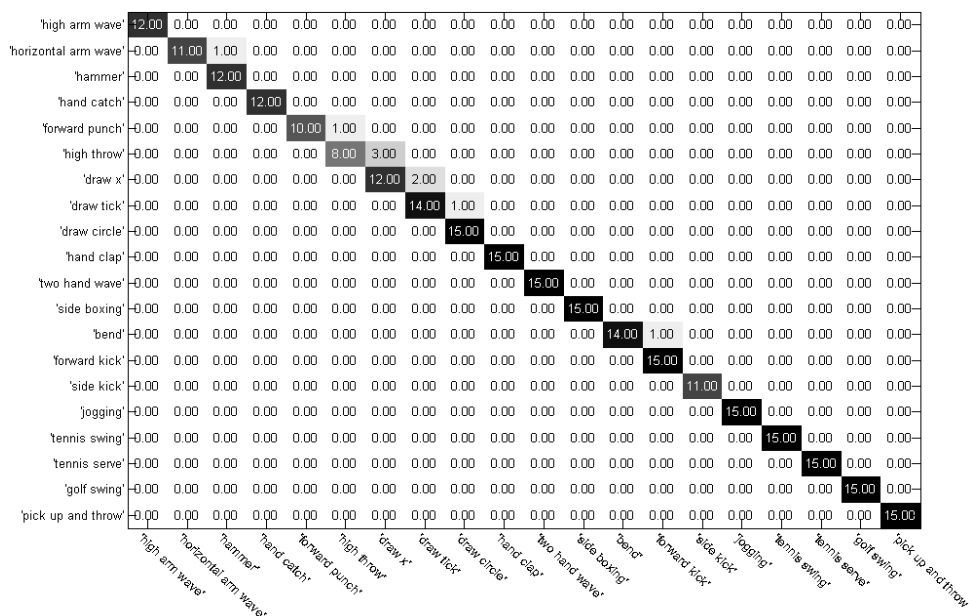


Figure 6. Confusion matrix of our method on MSR Action 3D dataset

To make fair comparison, we divided this dataset into three subsets (AS1, AS2, AS3) has done in [8]. In the Cross Subject Test, half subjects(1,3,5,7,9) are used for training and the rest ones (2,4,6,8,10) used for testing. The results of cross subject test on MSR Action 3D dataset are shown in Table 2. Five existing methods is used to compare with our method in Table 2, it is easy to see that our method achieves very superior results in AS2, AS3 and average. Even though our method is a little bit

lower than the one of DMM-HOG [8] in AS1, it is better than all of the rests and even higher of 8.54% than Actionlet Ensemble [4].

Table 1. The performance of our method on MSR Action 3D dataset, compared to previous approaches

Method	Accuracy%
Actionlet Ensemble [4]	88.2
Depth Cuboid [5]	89.3
HON4D [6]	88.89
DMM-HOG [8]	85.5
HOG2 [9]	94.84
HDG [10]	88.82
SNV [11]	93.09
Ours	96.73

Table 2. The performance of cross subject test on MSR Action 3Ddataset, compared to previous approaches

Method	AS1	AS2	AS3	Average
Action let Ensemble [4]	85.80	82.15	92.13	88.2
DMM-HOG [8]	96.2	84.1	94.6	91.6
SNV [11]	93.56	88.74	95.84	93.09
Ours	94.34	97.35	97.32	96.72

4.2. MSR Action Pairs 3D Dataset

In MSR Action Pairs 3D Dataset, each pair of actions has similar motion and shape but different spatial temporal order. Twelve activities are included in this dataset which are performed by 10 subjects where each subject was asked to do each activity 3 times. Based on the MSR Action Pairs3 D Dataset, we compare our method with four classic recognition models: Actionlet Ensemble [4], HON4D [6], DMM-HOG [8] and SNV [11]. From Table 3, we can see the accuracy comparison with different algorithms, and the comparison reveals that our method is able to significantly enhance the recognition rate and to yield similar or superior performance to state-of-the-art approaches. The accuracy of SNV [19] is higher than ours. This is probably due to the adaptive spatio-temporal pyramid structure [11] can globally capture the spatial and temporal orders while our method collect the pyramid structure of spatial information only. The confusion matrix is highlighted in Fig. 7. From Fig. 7, we can find there are only a little errors in the recognition of “wear a hat” and “stick a poster”. The results have demonstrated that temporal orders in actions can be well distinguished by our method.

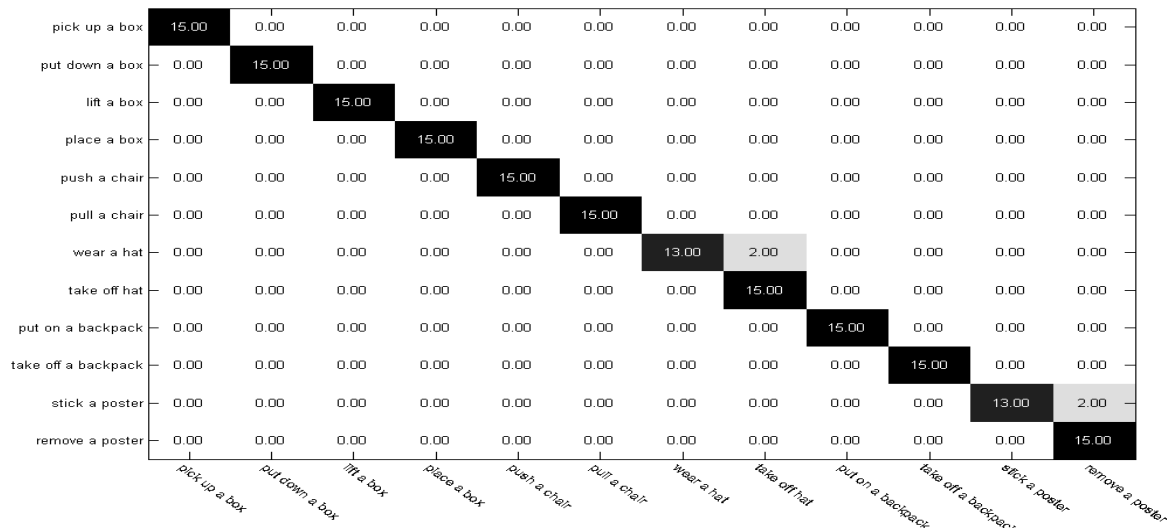


Figure 7. Confusion matrix of our method on MSR Action Pairs 3D dataset

Table 3. The performance of our method on MSRActionPairs3D dataset, compared to previous approaches

Method	Accuracy%
Actionlet Ensemble [4]	82.22
HON4D [6]	96.67
DMM-HOG [8]	66.11
SNV [11]	98.89
Ours	97.78

4.3. MSR Daily Activity 3D

MSR Daily Activity 3D dataset emphasizes the importance of capturing the shape and the motion cues jointly in the activity sequence. In this dataset, 16 activities are included in this dataset which was performed by 10 subjects, each subject was asked to perform each activity twice, one is in sitting position and the other is in standing position. Based on the MSR Daily Activity 3D data set, we compare our method with six action recognition models: Actionlet Ensemble [4], Depth Cuboid [5], HON4D [6], HDG [10] and SNV [11]. Recognition accuracy of different methods is shown in Table. 4, which illustrate that the MSR Daily Activity3D data set is more challenging. From this analysis, the results presented in Table 4 give a further evidence of the effectiveness of the proposed method, which is able to overcome [5], the best accuracy, by a margin of 4.9%. Fig. 8 highlights the Confusion matrix of our method for MSR Daily Activity3D data set. Our method presents a strong ability to discriminate “sit up” and “sit down” even though these action are very similar.

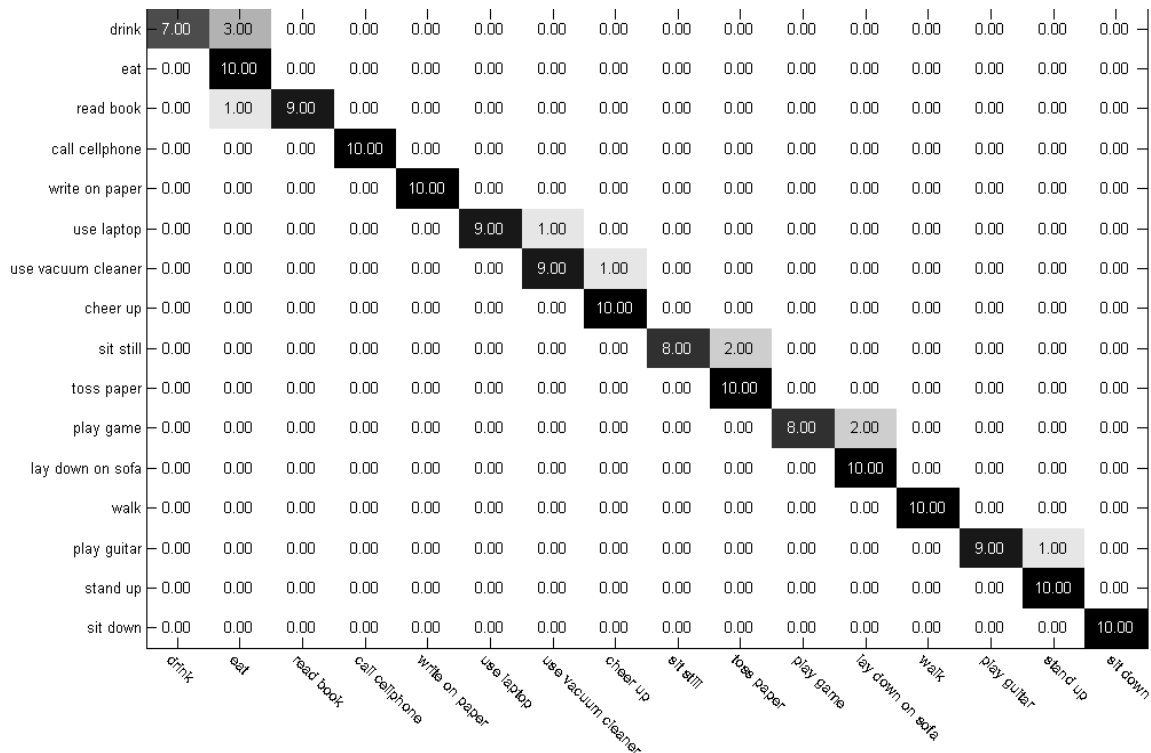


Figure 8. Confusion matrix of our method on MSRDailyActivity3D dataset

Table 4. The performance of our method on MSR Daily Activity 3D dataset, compared to previous approaches

Method	Accuracy%
Actionlet Ensemble [4]	85.75
Depth Cuboid [5]	88.2
HON4D [6]	80
HDG [10]	81.25
SNV [11]	86.25
Ours	93.13

5. Conclusion

The main contribution of our work is a new approach for local and multi-order HOG action recognition using a descriptor derived from depth map sequences without additional information involved. At first, the D-mHOG is used to obtain local shape and temporal information of depth action sequence and generate the DM array for further optimization of D-mHOG descriptor. Then, the multi-scale HOG (P-HOG) is applied on DM array to capture spatio-temporal structure information further. The experiment results on three public datasets show that the proposed approach can effectively improve the results and achieve the start-of-the-art performance without skeleton information involved. In future work, we will discuss real-time measure which integrates efficient computing time and high recognition accuracy for real-time human action recognition system.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.61672222, 61572183, 61472131), Science and Technology Planning Project of Changsha, China (Grant No.kq1706021).

References

- [1] M. Liu, H. Liu, "Depth context: a new descriptor for human activity recognition by using sole depth sequences", *Neuro computing*, Vol. 175, pp. 747 – 758, 2016.
- [2] Weng Z F, Wang S J, Huang L F, "A Virtual 3D Hair Reconstruction Method from a 2D Picture" *Journal of Computers*, Vol. 27 (1), 2016.
- [3] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, "Real-time human pose recognition in parts from single depth images", *Communications of the ACM*, Vol. 56 (1), pp. 116–124, 2013.
- [4] J. Wang, Z. Liu, Y. Wu, J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras", in *Computer Vision and Pattern Recognition (CVPR)*, pp. 1290 – 1297, 2012.
- [5] L. Xia, J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2834–2841, 2013.
- [6] O. Oreifej, Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013.
- [7] C. Lu, J. Jia, C.-K. Tang, "Range-sample depth feature for action recognition", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 772 – 779, 2014.
- [8] X. Yang, C. Zhang, Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients", in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1057–1060, 2012.
- [9] E. Ohn-Bar, M. Trivedi, "Joint angles similarities and hog2 for action recognition", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 465 – 470, 2013.
- [10] H. Rahmani, A. Mahmood, D. Q. Huynh, A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests", in *Applications of Computer Vision (WACV)*, pp. 626–633, 2014.
- [11] X. Yang, Y. Tian, "Super normal vector for activity recognition using depth sequences", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 804–811, 2014.
- [12] Bosch A, Zisserman A, "Pyramid histogram of oriented gradients (phog)", *Univ. Oxford Visual Geometry Group*, 2013.
- [13] X. Yang, Y. Tian, "Effective 3d action recognition using eigenjoints", *Journal of Visual Communication and Image Representation*, Vol. 25 (1), pp. 2 – 11, 2014.
- [14] Zhang H, Xu M, Zhuo L, et al. "A novel optimization framework for salient object detection", *The Visual Computer*, pp. 1-11, 2014.
- [15] W. Li, Z. Zhang, Z. Liu, "Action recognition based on a bag of 3d points", in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 9–14, 2010.
- [16] Lin W S, Lee C P, "A novel distance-based k-nearest neighbor voting classifier", *Journal of Computers*, Vol. 23 (3), pp. 26-34, 2012.