# Recognizing Actions Using Depth Motion Maps-based Histograms of Oriented Gradients

Xiaodong Yang, Chenyang Zhang, and YingLi Tian
Department of Electrical Engineering
The City College, City University of New York
{xyang02, czhang10, ytian}@ccny.cuny.edu

## ABSTRACT
In this paper, we propose an effective method to recognize human actions from sequences of depth maps, which provide additional body shape and motion information for action recognition. In our approach, we project depth maps onto three orthogonal planes and accumulate global activities through entire video sequences to generate the Depth Motion Maps (DMM). Histograms of Oriented Gradients (HOG) are then computed from DMM as the representation of an action video. The recognition results on Microsoft Research (MSR) Action3D dataset show that our approach significantly outperforms the state-of-the-art methods, although our representation is much more compact. In addition, we investigate how many frames are required in our framework to recognize actions on the MSR Action3D dataset. We observe that a short sub-sequence of 30-35 frames is sufficient to achieve comparable results to that operating on entire video sequences.

## Categories and Subject Descriptors
I.4.9 [**Applications**]

## General Terms
Algorithms, Performance, Experimentation

## Keywords
3D Action Recognition, RGBD Camera, Depth Maps, Feature Representation

## 1. INTRODUCTION
Automatic human action recognition has many real-world applications including content-based video search, human-computer interaction, video surveillance, health care, and etc [7-9, 14, 15]. In the past decades, research of human action recognition mainly concentrates on video sequences captured by traditional RGB cameras. The spatio-temporal volume-based methods have been extensively used for recognizing actions through measuring similarities between action volumes. In order to facilitate accurate similarity measurements, various detection and representation methods of spatio-temporal volumes have been proposed [3, 5-6]. Trajectory-based approaches have been explored for recognizing human activities as well [10]. In this case, human actions are interpreted by the motions of a set of key joints of human body. However, in traditional videos it is nontrivial to quickly and reliably detect and track human body joints.


*MM'12*, October 29–November 2, 2012, Nara, Japan.


As the imaging technique advances, e.g. the launch of Microsoft Kinect, it has become feasible to capture color image sequences as well as depth maps in real time by RGBD sensors. The depth maps are able to provide additional body shape and motion information to distinguish actions that generate similar projections from a single view, which motivates recent research work to explore action recognition based on depth maps [7, 12-14]. A Bag-of-3D-Points or 3D Silhouettes method was proposed in [7] to represent postures by sampling 3D points from depth maps. An action graph was then employed to model the sampled 3D points to perform action recognition. Their experimental results on MSR Action3D dataset [7] validated the superiority of 3D silhouettes from depth maps over 2D silhouettes from a single view.

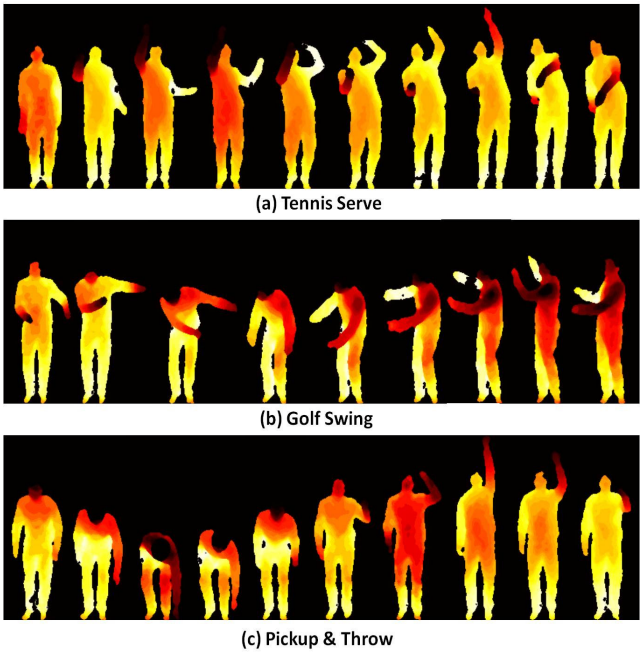


**Figure 1. The sampled sequences of depth maps for actions of (a) Tennis Serve, (b) Golf Swing, and (c) Pickup & Throw.**

In this paper, we focus on recognizing human actions using sequences of depth maps. Fig. 1 illustrates the depth maps for actions *Tennis Serve*, *Golf Swing*, and *Pickup & Throw*. As shown in this figure, depth maps provide additional shape and motion information. However, point cloud of depth maps also incur a

great amount of data that might result in expensive computations. Here, we propose an effective and efficient approach to recognize human actions by extracting Histograms of Oriented Gradients (HOG) descriptors from Depth Motion Maps (DMM). The DMM are generated by stacking motion energy of depth maps projected onto three orthogonal Cartesian planes. The stacked motion energy of each action category produces specific appearances and shapes on DMM, which can be used to characterize corresponding action categories. Motivated by the success of HOG in human detection [2], we adopt HOG descriptors to represent DMM. Compared to the original depth data, the proposed DMM-HOG representation is more compact and more discriminative. In addition, we investigate how many frames are sufficient to perform action recognition using DMM-HOG. The experiments on MSR Action3D dataset [7] demonstrate that a short sub-sequence (e.g. 35 frames) is sufficient to obtain reasonably accurate recognition results for human action recognition. This observation is important to make online decisions and to reduce observational latency when humans interact with computers.

## 2. RELATED WORK

The spatio-temporal volume-based methods are widely used in action recognition from videos captured by traditional RGB camera. These approaches mainly focus on detection and representation of space-time volumes. For example, Bobick and Davis [1] accumulated foreground regions of a person as Motion History Images (MHI) to explicitly track shape changes. Tian *et al*. [11] employed Harris detector and local HOG descriptor on MHI to perform action recognition and detection. Similar to MHI, the proposed DMM also stacks foreground motion regions to record where and how actions are performed. However, there are main differences: 1) MHI only keeps most recent motions to capture the recency of action, while DMM accumulates global activities through entire video sequences to represent the motion intensity; 2) our method stacks motion regions from front/side/top views, i.e. three orthogonal projections of depth maps, while only a single view is used in MHI. In the most recent work, local spatio-temporal features have been extensively used. As object recognition using sparse local features in 2D images, an action system first detects interest points [3, 5-6] and then computes descriptors based on the detected local spatio-temporal volumes. The local features are then combined (e.g. bag-of-words) to model different activities. The fundamental difference between those systems and our method is that they designed features based on 2D video sequences, instead of 3D depth maps that include supplementary information of body shape and motions.

With the release of RGBD sensors, research of action recognition based on depth information has been explored [7, 14]. Li *et al*. [7] used 3D Silhouettes for action recognition. They sampled a set of representative 3D points from depth maps to characterize the posture being performed in each frame. They first projected depth maps onto three orthogonal Cartesian planes and sampled 2D points at equal distance along contours of the three projections. The 3D points were then retrieved in depth maps according to the contour points. However, the sampled 3D points of each frame generated a considerable amount of data which resulted in expensive computations in clustering training videos of all classes. Yang and Tian [14] proposed an EigenJoints-based action recognition system by using a NBNN classifier. The compact representation of EigenJoints employed joints differences to capture action information of static postures, consecutive motions, and overall dynamics. However, the 3D positions of skeleton joints might be complete wrong if there are sever occlusions [9].

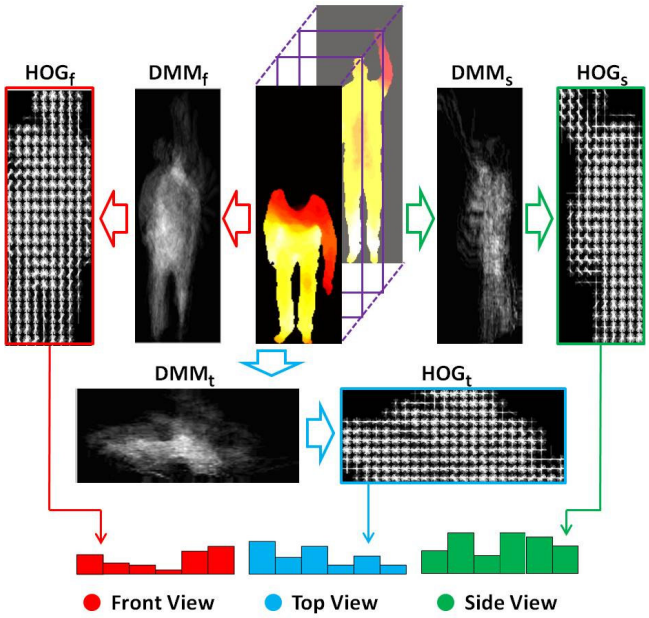


**Figure 2. The framework of computing DMM-HOG. HOG descriptors extracted from depth motion map of each projection view are combined as DMM-HOG, which is used to represent the entire action video sequences.**

## 3. COMPUTAION OF DMM-HOG

The framework to compute action representation of DMM-HOG is demonstrated in Fig. 2. We project depth frames onto three planes and compute associated motion energy, which are then stacked to obtain DMM. HOG descriptors are extracted from three depth motion maps and concatenated as the final action representation of DMM-HOG.

### 3.1 Depth Motion Maps (DMM)

In order to make use of the additional body shape and motion information from depth maps, each depth frame is projected onto three orthogonal Cartesian planes. We then set the region of interest of each projected map as the bounding box of foreground (i.e. non-zero) region, which is further normalized to a fixed size. This normalization is able to reduce intra-class variations, e.g. subject heights and motion extents, of different subjects when they perform the same action. So each 3D depth frame generates three 2D maps according to front, side, and top views, i.e. $map_f$, $map_s$, and $map_t$. As for each projected map, we obtain its motion energy by computing and thresholding the difference between two consecutive maps. The binary map of motion energy indicates motion regions or where movement happens in each temporal interval. It provides a strong clue of the action category being performed. We then stack the motion energy through entire video sequences to generate the depth motion map $DMM_v$ for each projection view:

$$DMM_v = \sum_{i=1}^{N-1} \left(\left|map_v^{i+1} - map_v^i\right| > \epsilon\right)$$

where $v \in \{f, s, t\}$ denotes the projection view; $map_v^i$ is the projected map of the $i$th frame under projection view $v$; $N$ is the number frames; $\left|map_v^{i+1} - map_v^i\right| > \epsilon$ is the binary map of motion energy; and $\epsilon$ is the threshold. We empirically set $\epsilon = 50$ in our experiments. As shown in Fig. 2, the DMM generated from an action video of *Pickup & Throw* demonstrate specific

appearances and shapes, which characterize the accumulated motion distribution and intensity of this action. The DMM representation encodes the 4D information of body shape and motion in three projected planes, meanwhile significantly reduces considerable data of depth sequences to just three 2D maps.

## 3.2 DMM-HOG Descriptor

HOG is able to characterize the local appearance and shape on DMM pretty well by the distribution of local intensity gradients. The basic idea is to compute gradient orientation histograms on a dense grid of uniformly spaced cells and perform local contrast normalization. In each cell, 4 different normalizations, i.e. L1-norm, L2-norm, L1-sqrt, and L2-Hys [2], are computed based on adjacent histograms. As for each depth motion map, we evenly sample $23 \times 10$ non-overlapping cells and 8 gradient orientation bins. So each $DMM_v$ generates a descriptor $HOG_v$ with the dimension of $4 \times 23 \times 10 \times 8 = 7360$. As shown in Fig. 2, we concatenate $[HOG_f, HOG_s, HOG_t]$ as the DMM-HOG descriptor which is the input to a linear SVM classifier to recognize human actions.

# 4. EXPERIMENTS AND DISCUSSIONS

The proposed method is evaluated on the MSR Action3D dataset [7]. We extensively compare our approach with the state-of-the-art methods under a variety of experimental settings. We further investigate how many frames are sufficient to recognize actions using DMM-HOG.

## 4.1 Experimental Setup

The MSR Action3D [7] is a public dataset with sequences of depth maps captured by a RGBD camera. It includes 20 action categories performed by 10 subjects facing to the camera during performance. Each action was performed 2 or 3 times by each subject. The depth maps are with the resolution of 320×240. The 20 action categories are chosen in the context of interactions with game consoles. As illustrated in Fig. 1, actions in this dataset reasonably capture a wide range of motions related to arms, legs, torso, and their combinations.

In order to facilitate a fair comparison, we follow the same experimental settings as [7, 14] to split 20 categories into three subsets as listed in Table 1. As for each subset, there are three different tests, i.e. Test One (One), Test Two (Two), and Cross Subject Test (CrSub). In Test One, 1/3 of the subset is used as training 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, half subjects are used for training and the rest ones used for testing.

**Table 1. Action subsets and tests used in our experiments.**

| Action Set 1 (AS1) | Action Set 2 (AS2) | Action Set 3 (AS3) |
|---|---|---|
| Horizontal Wave | High Wave | High Throw |
| Hammer | Hand Catch | Forward Kick |
| Forward Punch | Draw X | Side Kick |
| High Throw | Draw Tick | Jogging |
| Hand Clap | Draw Circle | Tennis Swing |
| Bend | Hands Wave | Tennis Serve |
| Tennis Serve | Forward Kick | Golf Swing |
| Pickup Throw | Side Boxing | Pickup Throw |

## 4.2 Evaluations of DMM-HOG

We first evaluate the effect of DMM normalization size to recognition performances. As discussed in Section 3.1, we normalize the three depth motion maps to a fixed size. Fig. 3 shows action recognition accuracies of DMM with different normalization sizes under a variety of test sets. The overall recognition rates of most test sets are similar across different DMM normalization sizes. As for AS1One and AS1Two, the size of 200×100 achieves the best results, while for AS2CrSub the size of 50×25 outperforms the others. Although lower resolutions are able to reduce computational cost in computing HOG, we extract HOG descriptors only from the three depth motion maps, instead of each video frame. So for each video, the difference of computation time between different sizes is limited. The following experimental results are based on the size of 200×100. As shown in Fig. 3, while the performances in AS1CrSub are promising, the recognition accuracies in AS2CrSub and AS3CrSub are relatively low. In Cross Subject Test, different subjects perform actions with great variations but the amount of subjects is limited, which results in considerable intra-class variations. Furthermore, some actions in AS2 are quite similar, e.g. *Draw X*, *Draw Tick*, and *Draw Circle*, which generates small inter-class variations. The performances on cross subject test might be improved by adding in more subjects.

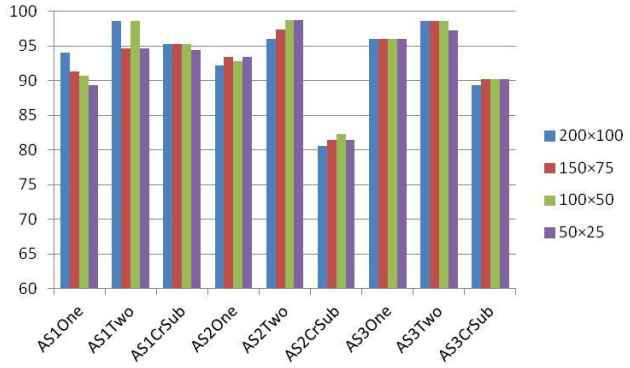


**Figure 3. The recognition rates (%) of DMM with different normalization sizes under a variety of test sets.**

## 4.3 How Many Frames Are Sufficient

Most existing systems [3, 5-7] recognize actions by operating on entire video sequences. We perform experiments to investigate how many frames are sufficient for action recognition with reasonably accurate results in our framework. The recognition rates using different amount of frames under a variety of test sets are demonstrated in Fig. 4. The sub-sequences are chosen from the first $K$ frames of a given video. As shown in this figure, in most cases 30-35 frames are sufficient to achieve comparable results to the ones using entire sequences, with quite limited gains or even some loss as more frames are added in. As affect recognition in [4], the temporal segments of an action can be intuitively approximated by the statuses of neutral, onset, apex, and offset. The most discriminative information is within the status of apex and onset, which are probably covered by the first 30-35 frames of the MSR Action3D dataset. The sequences after apex contribute little or even incur more noise. This observation provides important guides to reduce latency of action recognition systems where decisions have to be made on line. The following results are based on the sub-sequence of first 35 frames.

## 4.4 Comparisons to the State-of-the-art

We compare our DMM-HOG approach with the state-of-the-art methods including 3D Silhouettes [7] and EigenJoints [14] on the MSR Action3D dataset in Table 2. The recognition accuracies of 3D Silhouettes and EigenJoints are obtained from [7, 14]. The

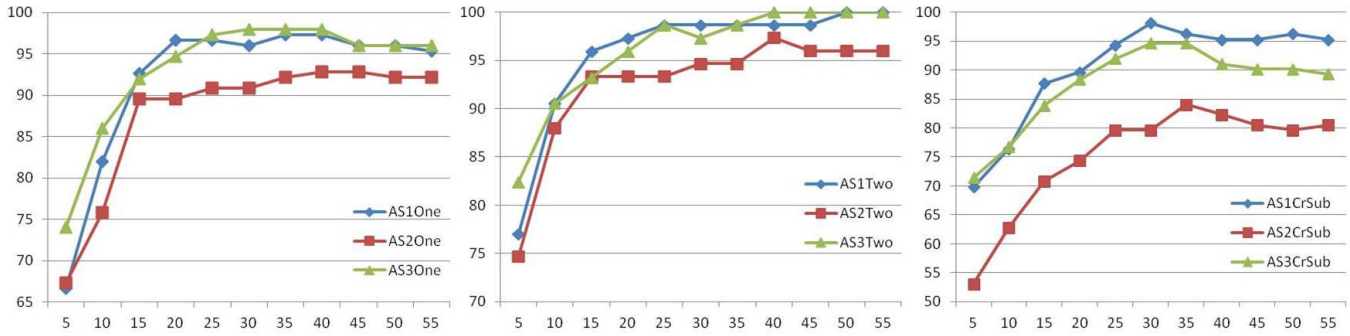


**Figure 4. The recognition accuracies (%) using different number of first frames in Test One (left), Test Two (middle), and Cross Subject Test (right). 30-35 frames are sufficient to enable accurate action recognition in most test sets.**

best results under different test sets are highlighted in bold. As shown in this table, our method consistently and considerably outperforms 3D Silhouettes in all testing cases. The overall accuracies under non-cross-subject tests of our method are comparable to that of EigenJoints. But our method largely outperforms EigenJoints for cross-subject tests. The significant improvement of our method in cross-subjects tests is probably because the normalization process in computing depth motion maps helps to reduce variations of different subjects as well as the robust action representation of DMM-HOG. In addition to recognition accuracy, out approach is much more compact than 3D Silhouettes. Moreover, our method achieves the state-of-the-art results using a short sub-sequence (35 frames), while 3D Silhouettes relied on the entire video sequences.

**Table 2. Recognition accuracies (%) of our method compared to the state-of-the-art methods on MSR Action3D dataset.**

| | 3D Silhouettes [7] | EigenJoints [14] | Ours |
|---|---|---|---|
| AS1One | 89.5 | 94.7 | **97.3** |
| AS2One | 89.0 | **95.4** | 92.2 |
| AS3One | 96.3 | 97.3 | **98.0** |
| AS1Two | 93.4 | 97.3 | **98.7** |
| AS2Two | 92.9 | **98.7** | 94.7 |
| AS3Two | 96.3 | 97.3 | **98.7** |
| AS1CrSub | 72.9 | 74.5 | **96.2** |
| AS2CrSub | 71.9 | 76.1 | **84.1** |
| AS3CrSub | 79.2 | **96.4** | 94.6 |

## 5. CONCLUSION

In this paper, we have proposed an effective action recognition method by using DMM-HOG descriptors. The compact and discriminative action representation is able to capture the global activities from front/side/top views. The experimental results on MSR Action3D dataset demonstrate that our approach significantly outperforms the existing state-of-the-art methods. In addition, we observe that in our framework a short sub-sequence of 30-35 frames is sufficient to perform action recognition with reasonably accurate results. The future work will focus on combining skeleton joints and depth maps to recognize actions, and incorporating more subjects to improve recognition in the cross subject test.

## 6. ACKNOWLEDGEMENT

This work was supported in part by NSF grant IIS-0957016 and Microsoft Research.

## 7. REFERENCES

[1] Bobick, A. and Davis, J. 2001. The Recognition of Human Movement Using Temporal Templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

[2] Dalal, N. and Triggs, B. 2005. Histograms of Oriented Gradients for Human Detection. *CVPR*.

[3] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. 2005. Behavior Recognition via Sparse Spatio-Temporal Features. *VS-PETS*.

[4] Gunes, H. and Piccardi, M. 2009. Automatic Temporal Segment Detection and Affect Recognition from Face and Body Display. *IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics*.

[5] Laptev, I. 2005. On Space-Time Interest Points. *International Journal of Computer Vision*.

[6] Laptev, I. Marszalek, M., Schmid, C., and Rozefeld, B. 2008. Learning Realistic Human Actions from Movies. *CVPR*.

[7] Li, W., Zhang, Z., and Liu, Z. 2010. Action Recognition based on A Bag of 3D Points. *IEEE Workshop on CVPR for Human Communicative Behavior Analysis*.

[8] Masood, S., Ellis, C., Tappen, M., Laviola, J., and R. Sukthankar. 2011. *IEEE Workshop on ICCV for Human Computer Interaction: Real Time Vision Aspects of Natural User Interfaces*.

[9] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. 2011. Real-Time Pose Recognition in Parts from Single Depth Images. *CVPR*.

[10] Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., and Li, J. 2009. Hierarchical Spatio-Temporal Context Modeling for Action Recognition. *CVPR*.

[11] Tian, Y., Cao, L., Liu, Z., and Zhang, Z. 2011. Hierarchical Filtered Motion for Action Recognition in Crowded Videos. *IEEE Trans. on Systems, Man, and Cybernetics – Part C: Applications and Reviews*.

[12] Wang, J., Liu, Z., Wu, Y., and Yuan, J. 2012. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. *CVPR*.

[13] Wang, J., Liu, Z., and Wu, Y. 2012. Robust 3D Action Recognition with Random Occupancy Patterns. *ECCV*.

[14] Yang, X. and Tian, Y. 2012. EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. *IEEE Workshop on CVPR for Human Activity Understanding from 3D Data*.

[15] Zhu, G., Yang, M., Yu, K., Xu, W., and Gong, Y. 2009. Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor. *ACM Multimedia*.