Action Recogntion with Depth Maps Using HOG Descriptors of Multi-view Motion Appearance and History

DoHyung Kim, Woo-han Yun, Ho-Sub Yoon, Jaehong Kim
Intelligent Cognitive Technology Research Department
Electronics and Telecommunications Research Institute
Daejeon, Korea.
{dhkim008, yochin, yoonhs, jhkim504}@etri.re.kr

***Abstract*—The goal of this work is to recognize human actions only using depth maps without additional joints information. As a practical solution, we present a novel volumetric representation of global shape of depth motion, Depth Motion Appearance (DMA). The proposed framework also extracts dynamic information of the body movements called Depth Motion History (DMH), an extended version of motion history image. In the framework, a huge amount of data of an action video is summarized into concise action representation maps observed from multi-view. A histogram of oriented gradients then describes local appearances and shapes of the DMAs and DMHs, which results in more compact and discriminative action representation. The presented method has been compared with the state-of-the-art approaches on a public dataset. The experimental result demonstrates that our approach achieves a better and more stable performance with a relatively smaller feature maps and lower complexity.**

Keywords-Action recognition; Depth maps; Depth motion appearance; Depth motion history; Histogram of oriented gradients.

I. INTRODUCTION

Despite numerous research efforts and advances in the last decade, traditional human action recognition with the sequence of 2D color images is still a challenging problem. Human actions are in essential continuous evolution of dynamic motion of three-dimensional body parts and articulated joints. In addition, same action can be performed in various ways of body movements by each individual and two different actions having a similar trajectory of motion make it more difficult to distinguish correctly. So, the absence of depth information could lead to significant degradation of discriminating capability of an action recognizer and consequently limit its performance.

In recent years, the technology of action recognition has entered a new phase with the release of the low-cost depth cameras like Microsoft Kinect [1]. These depth cameras provide 3D depth data as well as color image sequences in real time, which makes it possible to explore the fundamental solution for traditional problems in human action classification. Recent studies taking advantage of 3D information have been showing advanced results compared to the traditional 2D video-based researches [2][3][5].

As it is well known, the human actions could be modeled by the motion of a set of three-dimensional articulated joints [4]. So, if we can obtain 3D positions of key joints in real time with reasonable accuracy, action recognition can be successfully accomplished. However, estimating 3D joint positions is still a challenging task. Although some consumer depth cameras provide body joints information, the estimated joint positions are coarse and sometimes have significant errors particularly when body parts are self-occluded like two hands crossing. Moreover, most depth sensors only provide a sequence of depth maps. For these practical reasons, the work presented here has focused on recognizing human actions only using depth maps without additional information of the joints of the skeleton.

The main contributions of this work include two aspects.

First, we propose the Depth Motion Appearance as a new way of describing the global 3D shape of a body movement. It is a 3D depth map which represents a region of forward depth motion stacked through all of the depth images of an action. Our method can be differentiated from the prior depth map-based studies. The work by Li et al. [5] only uses 2D projects of key poses instead of direct utilization of the 3D information, which could essentially lead to sub-optimal feature representations. While our method makes full use of 3D information of all depth maps in the sequence, which results in the improved discriminating power. Xiaodong et al. [6] generate a binary map of motion energy by computing and thresholding the difference between consecutive depth maps. But, their method crucially does not consider dynamic information of the body movements. On the contrary, our framework effectively combines the appearance feature with the temporal feature extracted by an extended framework of motion history image [7].

Second, the proposed approach yields the best accuracy when compared with many previous state-of-the-art action recognition methods based on 3D silhouettes or joints. Moreover, the result is achieved with relatively small feature sets. An entire sequence of depth maps can be encoded just to a 4096 dimensional HOG (Histogram of Oriented Gradients) descriptor [8]. This fact indicates that our action representation method is highly discriminative as well as computationally efficient.

This paper is organized as follows. In Section 2, the overview of the proposed framework is described. The detailed description of the proposed features is given in Section 3. An evaluation model and the experimental results

 ISBN: 978-1-61208-353-7

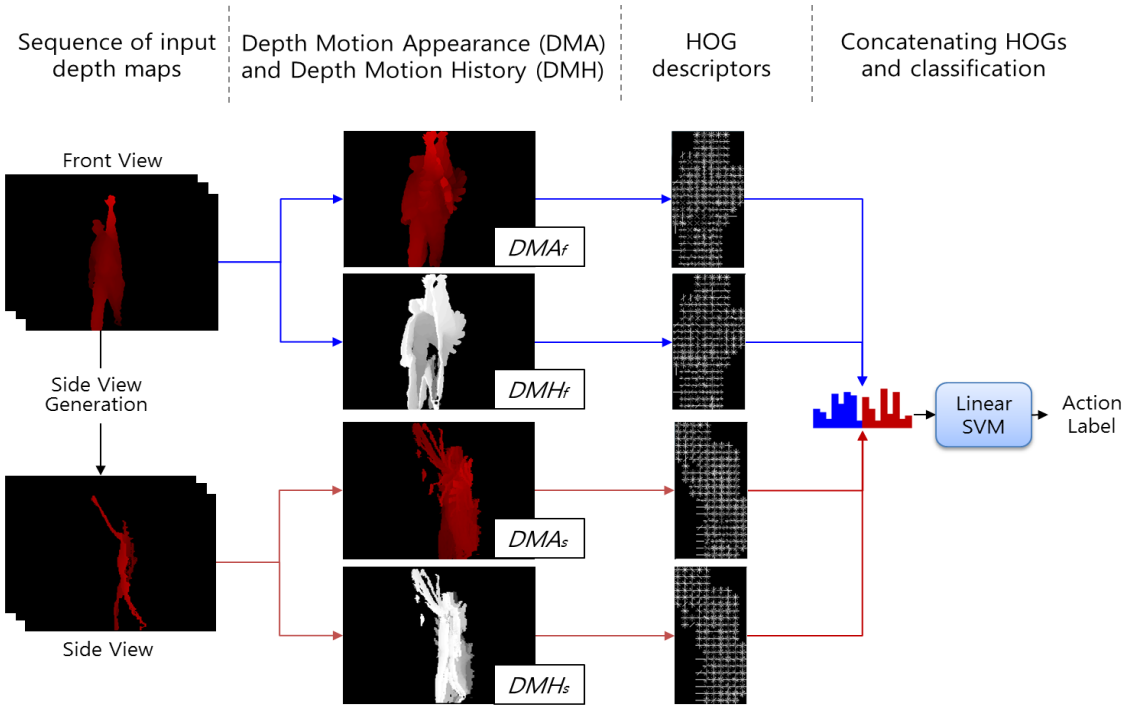


Figure 1. Overview of the feature extraction and action classification framework proposed in this paper.

are shown in Section 4. Finally, Section 5 concludes this paper.

II. Framework Overview

The proposed framework of the feature extraction and action classification is shown in figure 1. When the sequence of front-viewed depth maps is fed into the framework, it first generates a side-viewed depth map from the input depth map in order to acquire additional evidences. The framework then accumulates global activities through entire sequence of the depth images from each view and creates action representation maps called Depth Motion Appearance (DMA) and Depth Motion History (DMH). The DMA is an accumulated form of 3D depth information. It has no temporal information about the sequence of the motion, which can be complemented by the DMH that includes dynamic information of the entire motion region. In total four representation maps, two maps from each view, are generated for one action video. The system then calculates Histogram of Oriented Gradients (HOG) feature descriptors [7] for size-normalized DMAs and DMHs. The descriptors are concatenated into one single HOG descriptor which is fed into a linear Support Vector Machine (SVM) [9]. The linear SVM classifies the HOG descriptor and finally yields the action label of the query sequence.

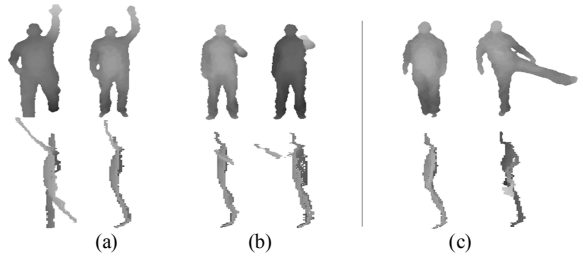


Figure 2. Original front-viewed depth maps (top row) and newly created side-viewed depth maps (bottom row): (a) and (b) depth maps with similar frontal shape but discriminable profile shape, (c) the opposite case of (a) and (b).

III. Action Representation

A. View generation

The side-viewed depth map provides an additional body shape and motion information different from that extracted from the frontal depth image. As shown in Figure 2, similar actions which are difficult to be distinguished from the front view might be easily discriminable in a lateral view and the opposite is true as well. Therefore, taking advantage of observations from various views can be an efficient and effective approach for 3D action classification. In order to capture full body actions, actors are commonly located at a long distance from depth sensors, which leads to a low depth resolution for the body region. So, interpolation methods are basically needed to estimate and produce new depth points when creating side-viewed depth images.

B. Depth Motion Appearance

The DMA is a volumetric representation of depth motion which describes the overall shape and appearance of a body movement forming an action. As for each view, we can

 ISBN: 978-1-61208-353-7

(a) (b)

Figure 3. DMAs generated from different human actions: (a) sequences of input depth maps, (b) DMAs (side boxing, two hand wave, and tennis swing from top to bottom).

obtain the DMA by accumulating all depth maps of an action video from start to end.

$$DMA_v(i,j,t) = \begin{cases} D_v(i,j,t) & ,if\ DMA_v(i,j,t-1) = 0 \\ min(D_v(i,j,t), DMA_v(i,j,t-1)), & else. \end{cases} \quad s.t.\ D_v(i,j,t) > 0 \tag{1}$$

where v denotes the view, $D_v(i,j,t)$ is a depth value at a pixel position i, j of the t th input depth map under the view v, and $DMA_v(i,j,t)$ is a depth value at a pixel position i, j of the DMA_v generated from t input depth maps. The depth values of foreground region of an input depth map are only calculated for creating the DMA.

Figure 3 shows several action sequences and their DMAs respectively. For each action, the DMA represents its own distinctive appearance of body movement, which means it can be a strong feature for action classification. In addition, the DMA has an advantage in practical terms because it does not require any threshold values at all.

C. Depth Motion History

Although the DMA is a good method to represent appearance of a body movement, it does not include temporal information at all. Human actions are in essential continuous evolution of dynamic motion of body parts and articulated joints. Therefore, the absence of dynamic information on a sequence of movements can be a tremendous loss for an action recognizer. For extraction of temporal features, we present a method called the DMH, which is an extended form of Motion History Image (MHI) [8]. Traditional MHI can only cover the motion history occurred on the 2D image plane. With the depth information we can now encode the history of the motion along the depth changing directions.

$$DMH_v(i,j,t) = \begin{cases} \tau, & if\ |D_v(i,j,t) - D_v(i,j,t-1)| > \delta \\ max(DMH_v(i,j,t-1) - 1, 0), & else. \end{cases} \tag{2}$$

where $DMH_v(i,j,t)$ denotes a history value of depth motion at a pixel position i, j of the DMH_v created from t input depth map under the view v. τ is a time window for history and δ is a threshold value for depth difference between consecutive depth maps. The generated DMH is a two-dimensional image template where pixel intensity is a function of the recency of depth motion in a sequence.

D. Histogram of Oriented Gradients

The presented action representation method summarizes a great amount of depth data of the entire video into just four maps. We exploit the HOG method to describe local appearance and shape of the DMAs and DMHs. The HOG technique figures out the distribution of intensity gradients or edge directions in localized portions of an image [7]. Since the descriptor operates on localized cells, the method upholds invariance to geometric and photometric transformations.

For all the DMAs and DMHs, foreground regions are cropped and then normalized to a fixed size. Despite the same action, it can be variously performed by different actors. The size normalization can reduce intra-class variations including a human body type, a motion scale, and a distance between an actor and a sensor. We then achieve HOG descriptors by dividing each map into 8×16 non-overlapping cells and for each cell compiling a histogram of 8 gradient directions for the pixels within the cell. The local histograms are contrast-normalized using L2-norm measure. Each feature map is described as a HOG descriptor with the dimension of 8×16×8=1024 and we finally obtain a 4096 dimensional HOG descriptor from the entire action video. The HOG descriptor is fed into a multi-class linear SVM classifier that is implemented by using an open source library, LIBSVM [9].

IV. Experimental Results

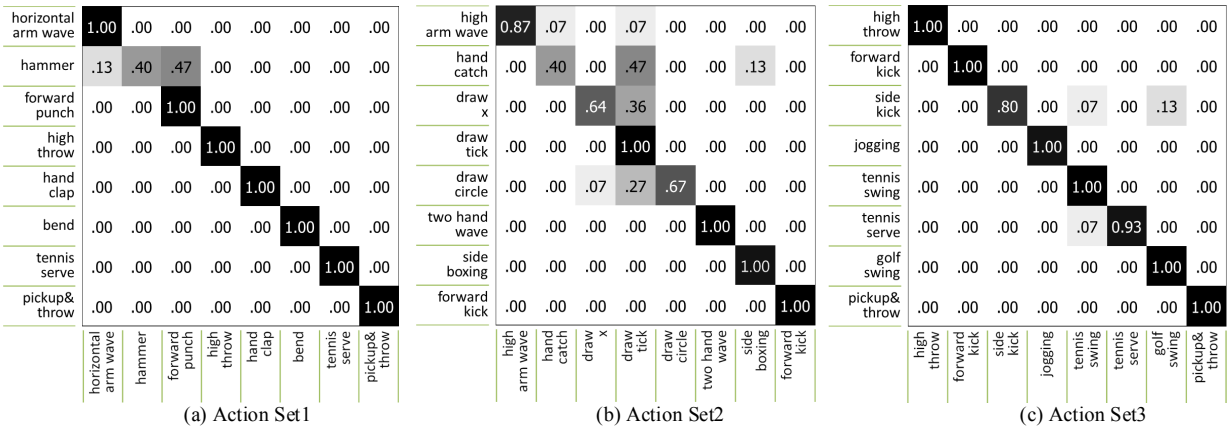
A. MSR Action3D dataset

The MSR Action3D dataset [5][10] is a public dataset on which a large number of methods have been experimented. The dataset provides sequences of depth maps captured by a depth sensor similar to the Kinect device. It contains 20 actions: *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*. The actions were chosen in the context of using the actions to interact with game consoles. They reasonably capture the various movements of arms, legs, torso and their combinations. In total, 567 depth map sequences are available. The resolution of the depth maps is 320×240. The dataset also provides the 3D joint positions extracted by the skeleton tracker [11]. Although the background of the dataset is clean, this dataset is still challenging due to the small inter-class variations among actions. Some actions of the dataset are shown in figure 3.

B. Evaluation of the proposed method

We evaluate our method with cross subject test setting [5][19], where the samples of the first five subjects are used

 ISBN: 978-1-61208-353-7



(a) Action Set1 (b) Action Set2 (c) Action Set3

Figure 4. Confusion matrices of the proposed method under the cross subject test setting on the MSR Action3D dataset

in training and the rest of the samples for testing. The cross subject test is more challenging and closer to the real world situation because the subjects used for training are different from those used for testing, which results in the considerable variations in the same action.

Table 1 shows the result of a comparative analysis of the proposed feature descriptors on each view and their combinations. Both the DMA and DMH show the competitive accuracy of 79.50% and 85.95%, respectively, just for the front view. It basically proves that our feature descriptors are appropriate to discriminate 3D human actions. We also achieved significant improvement on the recognition accuracy through combination of the observations from multiple views, 89.61% for the DMA and 87.64% for the DMH. This result means that reproducing new evidence from diverse views is an effective and practical approach to increase the discriminating power. We could finally obtain the outstanding recognition rate of 90.45% by combining the HOG descriptors of the multi-view DMA and DMH.

TABLE I. COMPARISON OF RECOGNITIONS RATES (%) FOR THE PROPOSED FEATURE DESCRIPTORS ON EACH VIEW AND THEIR COMBINATIONS ON THE MSR ACTION3D DATASET.

Feature Descriptors	Front view	Side view	Multi-view
DMA+HOG	79.50	69.66	89.61
DMH+HOG	85.95	70.78	87.64
DMA+DMH+HOG	85.95	71.07	**90.45**

The confusion matrices of the proposed method are illustrated in Figure 4. The recognition rates on Action Set1, Action Set2, and Action Set3 under the cross subject test setting were 92.37%, 82.35%, and 95.63%, respectively. The accuracy on Action Set2 containing many similar actions is relatively lower than those on the other two sets. The accuracies for hammer in Action Set1 and hand catch in Action Set2 are quite low compared to the other actions. This is because the way of performing these two actions varies depending on the subjects. In Action Set2, we observed that draw x, draw tick, and draw circle are mutually confused because they all have very similar trajectories of hand motion. For actions in Action Set3 in which body movements are quite different from one another, our method works very well.

C. *Comparison with the state-of-the-art methods*

We compared our approach with several previous methods. In terms of used primitives, previous 3D action recognition solutions could be categorized as 1) skeleton-based approaches that model the pose of the human body using motion of a set of 3D articulated joints [12][13][14], 2) depth map-based approaches that represent actions with volumetric and temporal features extracted from the entire depth maps in a sequence [6][15][16][17][18], and 3) hybrid solutions which combine information extracted from both the joints of the skeleton and the depth maps [19][20].

TABLE II. PERFORMANCE OF THE PROPOSED METHOD ON THE MSR ACTION3D DATASET COMPARED WITH THE PREVIOUS STATE-OF-THE-ART RESULTS.

Methods	Accuracy (%)
HOJ3D [12]	78.97
EigenJoint [13]	82.33
STOP [16]	78.20
DMM+HOG [6]	85.52
Random Occupancy Patterns [17]	86.50
Actionlet Ensemble [19]	88.20
HON4D+D_{disc} [18]	88.89
JAS+MaxMin+ HOG^2 [20]	94.84
DMA+DMH+HOG (*ours*)	**90.45**

As shown in Table 2, the proposed method clearly outperforms many well-known state-of-the-art approaches utilizing diverse primitives. It is also observed that the accuracy of our method is lower than that of one hybrid method [20] that exploits both joints and depth map information. Here, it is important to note that the goal of this work is to classify actions only using raw depth maps without additional joints information. Considering cost-effectiveness and extensibility, we believe our method has highly competitive performance.

 ISBN: 978-1-61208-353-7

V. Conclusion And Future Work

In this paper, we presented a practical and effective solution to three-dimensional human action recognition especially only using a sequence of depth maps. The method extracted a compact and discriminative HOG descriptor of the Depth Motion Appearances and Depth Motion Histories from multi-view. The experimental results on the public dataset showed that the proposed approach significantly outperformed the previous action classification methods.

As future work, we plan to investigate other descriptors based on both depth and skeleton information to manage the problem of human-object interaction and develop a dynamic classifier to reduce inter-class variations.

Acknowledgment

This work was supported by the IT R&D program. [10041610, The development of the recognition technology for user identity, behavior and location that has a performance approaching recognition rates of 99% on 30 people by using perception sensor network in the real environment]

References

[1] Z. Zhang, "Microsoft kinect sensor and its effect," Multimedia, IEEE, vol. 19, no. 2, 2012, pp. 4-10.

[2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," Computer Vision and Image Understanding, vol. 115, no. 2, 2011, pp. 224-241.

[3] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," Pattern Recognition Letters, 2013, pp. 1995-2006.

[4] V. M. Zatsiorsky, Kinematics of Human Motion. Human Kinetics Publishers, 1997.

[5] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on. IEEE, 2010, pp. 9-14.

[6] Y. Xiaodong, C. Zhang, and Y. Tian. "Recognizing actions using depth motion maps-based histograms of oriented gradients," Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 1057-1060.

[7] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 3, 2001, pp. 257-267.

[8] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, 2005, pp. 886-893.

[9] C. Chang and C. Lin, " LlBSVM: a library for support vector machines," ACM Tansactions on Intelligent Systems and Technology, vol. 2, no. 3:27, 2011.

[10] Microsoft Research. MSR Action Recognition Datasets, http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/default.htm

[11] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," Communications of the ACM, vol. 56, no. 1, 2013, pp. 116-124.

[12] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on. IEEE, 2012, pp. 20-27.

[13] Y. Xiaodong, and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on. IEEE, 2012, pp. 14-19.

[14] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," In Proc. Of ACM SIGGRAPH/Eurographics Symp. on Computer Animation, 2011, pp. 147–156.

[15] E. Frigerio, M. Marcon, and S. Tubaro, "Improving action classification with volumetric data using 3D morphological operators," Acoustics, Speech and Signal Processing, 2013 IEEE International Conference on, 2013, pp. 1849-1853.

[16] A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, and M.F Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," Progress in Pattern Recognition, Image Analysis, Computer Vision, and Application. 2012, pp. 252-259.

[17] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," Computer Vision–ECCV 2012, 2012, pp. 872-885.

[18] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," Computer Vision and Pattern Recognition, 2013, pp. 716-723.

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," Computer Vision and Pattern Recognition, 2012, pp. 1290-1297.

[20] E. Ohn-Bar and M. M. Trivedi, "Joint Angles Similiarities and HOG2 for Action Recognition," Computer Vision and Pattern Recognition Workshops, 2013 IEEE Computer Society Conference on. IEEE, 2013, pp. 465-470.

 ISBN: 978-1-61208-353-7