3D Motion Trail Model based Pyramid Histograms of Oriented Gradient for Action Recognition

Bin Liang
Charles Sturt University
Email: bliang@csu.edu.au

Lihong Zheng
Charles Sturt University
Email: lzheng@csu.edu.au

***Abstract*—Human action recognition based on the depth maps is an important yet challenging task. In this paper, a new framework based on the 3D motion trail model (3DMTM) and Pyramid Histograms of Oriented Gradient (PHOG) is proposed to recognize human actions from sequences of depth maps. Specifically, a discriminative descriptor called 3DMTM-PHOG is proposed for depth-based human action recognition. The 3DMTM is generated through the entire depth video sequence to encode additional motion information from three projected orthogonal planes. By adding pyramid representation, Histograms of Oriented Gradient (HOG) descriptor is extended to PHOG which can well characterize local shapes at different spatial grid sizes for action recognition. PHOG is then computed from the 3DMTM as the 3DMTM-PHOG descriptor for the representation of an action. The proposed approach based on 3DMTM-PHOG descriptor is evaluated on MSR Action3D dataset captured by depth cameras. Experimental results show that the proposed approach outperforms the state-of-the-art methods and demonstrate the effectiveness and robustness of the proposed 3DMTM-PHOG descriptor.**

I. Introduction

Human action recognition has been a widely studied area in computer vision. It has many potential applications including human computer interaction, video surveillance, health care and content-based video retrieval. However, it is still a challenging task to recognize human actions accurately due to the large intra-class variability and inter-class similarity of actions, cluttered background, possible camera movements and illumination changes. There are two major issues for human action recognition. One is the sensor used for capturing the action information of human, and the other one is the representation of human actions that are dynamic and ambiguous [26].

In the past decades, research of human action recognition has concentrated on learning and recognizing human actions from video sequences captured by ordinary RGB cameras. The human actions are performed in 3D space, and capturing 3D human movements from RGB cameras is a very difficult task. Recently, the introduction of cost-effective depth cameras, *e.g.* the Microsoft Kinect, provides new possibilities to address difficult issues in human action recognition. Compared with ordinary RGB cameras, depth cameras can provide 3D depth data so that the infromation of actions can be more discriminative. Moreover, depth maps are insensitive to illumination changes. Besides, the motion ambiguities, such as the huge color and texture variability induced by clothing, hair, skin and background, could be bypassed. Depth information has long been regarded as an essential part of successful action recognition [8]. Thus, recent research work has been motivated to explore more efficient approaches based on depth maps. In addition, it is a hot topic that how to extract robust features from depth human action maps in an efficient way.

There is extensive literature in action recognition in a variety of research areas, including computer vision, machine learning, pattern recognition and signal processing[1, 23]. The spatio-temporal volume-based method is extensively adopted by comparing the similarity between two action volumes, and various detection and representation of spatio-temporal volumes have been proposed [2, 7, 9–11]. Other methods based on trajectory have also been widely explored for human activity recognition [18, 21]. In these approaches, human actions are interpreted by a set of body joints or other interest points. However, it is difficult to quickly and reliably extract or track body joints from ordinary RGB images. With the launch of depth cameras, the work of Shotton *et al.* [20] provides an efficient human motion capturing technology to accurately estimate the 3D skeleton joint positions from a single depth image. Thus, many methods using estimated 3D skeleton joint points [26, 28, 29] are popular for action recognition. However, this kind of methods has the limitation that it requires reliable skeleton data.

In this paper, we focus on recognizing human actions from the original depth data. An effective human action recognition method using 3DMTM-PHOG descriptor is presented in this paper. We propose Pyramid Histograms of Oriented Gradient (PHOG) descriptors extracted from the 3D Motion Trail Model (3DMTM) of actions. The 3DMTM [13] is employed to represent action motion information and static posture information in 3D space by projecting depth maps onto three orthogonal Cartesian planes. This model contains disparity information from the corresponding depth maps. Motivated by the success of HOG in human detection [6], we extend the HOG and propose a spatial pyramid representation to encode the 3DMTM for action recognition. In contrast to the PHOG descriptors proposed in [3] and [27], our proposed PHOG is directly extracted from the entire templates of 3DMTM, without extracting edges of objects or capturing the interesting regions, which is necessary in [3] and [27], respectively. Compared to the original depth data, the proposed 3DMTM-PHOG descriptor is more compact and more discriminative to encode human actions. Depth maps contain a great amount of data which could result in high computational costs. We then apply Principal Component Analysis (PCA) [19] on the 3DMTM-PHOG descriptor to reduce redundancy and increase the recognition speed. Support vector machine (SVM) [4] is employed to recognize multiple action categories in the

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.341

final stage. The experiments on MSR Action3D dataset [12] demonstrate that the proposed method is able to achieve better recognition accuracy than the state-of-the-art methods.

The remainder of this paper is organized as follows. Section II reviews three types of existing methods for action recognition. In section III, we provide a detailed procedure of our proposed method based on 3DMTM-PHOG. A variety of experimental results and discussions are presented in Section IV. At last, we give a conclusion of the paper and outline the future work in section V.

II. Related Work

In traditional video sequences captured by RGB cameras, as human actions are showing spatio-temporal patterns, action recognition mainly focuses on analyzing spatio-temporal volumes. The spatio-temporal interest points (STIPs) [10] are widely used in action recognition from videos. Besides, it is a common practice to use the distributions of the local features like HOG/HOF [11] or HOG3D [9] to represent the local spatio-temporal pattern. These local features are then combined to model different actions. Bobick and Davis [2] propose Motion History Image (MHI) and Motion Energy Image (MEI) for template matching. Tian *et al.* [22] employ Harris detector and local HOG descriptor on MHI to perform action recognition and detection. The core of these approaches is the detection and representation of spatio-temporal volumes.

With the release of depth sensors, research of action recognition based on depth information has been explored. Motivated by the joints estimation of Kinect and associated SDK, there have been many different approaches relying on joint points for action recognition. In [16], actions are modeled by dynamic temporal warping (DTW), which makes the 3D joint positions to a template, and action recognition can be done through a nearest-neighbor classification method. In [26], the joints of the skeleton are used as interest points. In this way, the shapes of the area surrounding the joint along with the joint location information are captured using a local occupancy pattern feature and a pairwise distance feature, respectively. Xia *et al.* [28] propose a compact representation of postures named HOJ3D using the 3D skeletal joint locations from Kinect depth maps. Then they train HMMs to classify the sequential postures into action types. Yang *et al.* [29] propose a type of features by adopting the differences of joints. EigenJoints are then obtained by PCA for classification. However, the 3D joint positions that are generated via skeleton tracking from the depth map sequences are generally more noisy. The performance of joint based methods heavily depends on a good estimation of skeleton information.

Furthermore, many other methods are proposed for action recognition based on the original depth data. Li *et al.* [12] propose a bag of 3D points model for action recognition. A set of representative 3D points from the original depth data is sampled to characterize the posture being performed in each frame. The 3D points are then retrieved in depth maps according to the contour points. Oreifej *et al.* [17] describe the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time (HON4D), depth, and spatial coordinates. However, these approaches generate a considerable amount of data which result in expensive computations in classification.

Different from these approaches, a novel descriptor (3DMTM-PHOG) is proposed in this paper to represent discriminative features of human actions, and we apply SVM to classify the proposed descriptors. In our framework, we employ 3DMTM to represent the action as a set of temporal templates. The 3DMTM is able to represent the motion information and static posture information of human actions in 3D space. Then 3DMTM-PHOG descriptor is proposed to represent the 3DMTM in different degree of details according to the selected pyramid levels. It requires no extraction of edges or interesting regions, which is necessary in some other methods. The proposed descriptor does not involve complicated computations (*e.g.*, bag of 3D points and HON4D), and it is much more robust to model 3D actions.

III. Proposed Method

The proposed framework for human action recognition from depth maps is demonstrated in Fig. 1. As shown, the framework consists of three components, 3D motion trial model (3DMTM) for action representation, feature extraction from 3DMTM using 3DMTM-PHOG descriptor and classification.

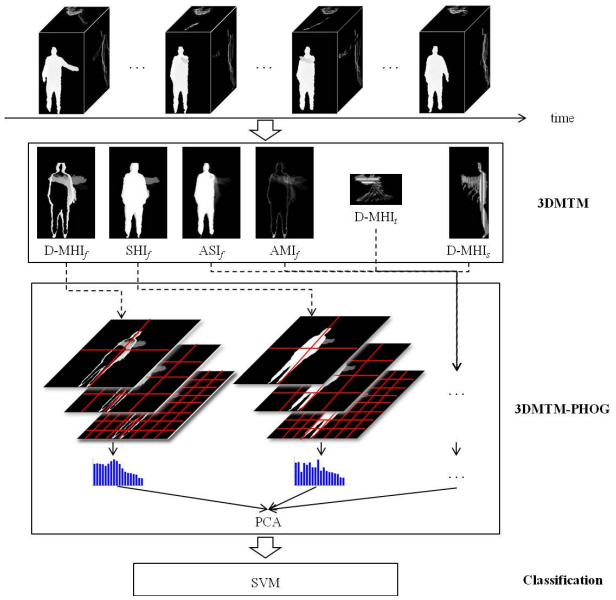


Fig. 1. The framework of the proposed method

A. 3D Motion Trial Model

MHI [2] presents the motion history by condensing the action sequence into a single gray scale image, preserving dominant motion history information. Although binary images or silhouette based images are able to represent a wide variety of body configurations, they could produce ambiguities in the presence of occlusions of body. Additionally, improper implementation of the update function, the MHI fails to cover most of the motion regions. Furthermore, the information of static posture history regions, repetitive movements and repetitive static postures is ignored in the MHI template [13].

In order to increase the robustness of action representation, we employ 3D Motion Trail Model (3DMTM) [13] to represent

human actions in 3D space. The 3DMTM employs four templates along the front view, *i.e.* depth motion history image (D-MHI$_f$), average motion image (AMI$_f$), static posture history image (SHI$_f$), and average static posture image (ASI$_f$). Additionally, another two D-MHIs (D-MHI$_t$ and D-MHI$_s$) along the top view and side view are also included in the 3DMTM.

The motion update function $\Psi_{\mathrm{M}}(x, y, t)$ and static posture update function $\Psi_{\mathrm{S}}(x, y, t)$ are defined to indicate the regions of motion and static posture with action performing. They are called for every frame analyzed in the action sequence:

$$\Psi_{\mathrm{M}}(x, y, t) = \begin{cases} 1 & \text{if } D_t > \varsigma_M, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$$\Psi_{\mathrm{S}}(x, y, t) = \begin{cases} 1 & \text{if } I_t - D_t > \varsigma_S, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where x, y represent pixel position and t is time. $I_t = (I_1, I_2, \ldots, I_T)$ is a depth map sequence and $D_t = (D_1, D_2, D_3, \ldots, D_T)$ is a difference image sequence indicating the absolute difference between two frames. In addition, these two update functions need thresholds ς_M and ς_S for motion and static information between consecutive frames.

Therefore, the depth motion history image (D-MHI) $\mathrm{H}_{\mathrm{M}}(x, y, t)$ can be obtained by using motion update function $\Psi_{\mathrm{M}}(x, y, t)$:

$$\mathrm{H}_{\mathrm{M}}(x, y, t) = \begin{cases} T & \text{if } \Psi_{\mathrm{M}}(x, y, t) = 1 \\ \mathrm{H}_{\mathrm{M}}(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{3}$$

Additionally, static posture history image (SHI) $\mathrm{H}_{\mathrm{S}}(x, y, t)$ can be generated utilizing the static posture update function $\Psi_{\mathrm{S}}(x, y, t)$ to compensate for static regions over the whole action sequence, which can be obtained in the similar way as D-MHI:

$$\mathrm{H}_{\mathrm{S}}(x, y, t) = \begin{cases} T & \text{if } \Psi_{\mathrm{S}}(x, y, t) = 1 \\ \mathrm{H}_{\mathrm{S}}(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{4}$$

Fig. 2 shows the D-MHI$_f$ and SHI$_f$ generated from the front view of one sample action (*horizontal arm wave*).

(a) D-MHI$_f$ (b) SHI$_f$

Fig. 2. D-MHI$_f$ and SHI$_f$ from front view of one sample action

In order to cover the information of repetitive movements and repetitive static postures over the whole action sequence, average motion image (AMI$_f$) and average static posture image (ASI$_f$) are employed. AMI$_f$ and ASI$_f$ are defined as follows:

$$\mathrm{A}_{\mathrm{M}} = \frac{1}{T} \sum_{t=1}^{T} \Psi_{\mathrm{M}}(x, y, t) \tag{5}$$

$$\mathrm{A}_{\mathrm{S}} = \frac{1}{T} \sum_{t=1}^{T} \Psi_{\mathrm{S}}(x, y, t) \tag{6}$$

Fig. 3 shows the AMI$_f$ and ASI$_f$ from the front view of one sample action.

(a) AMI$_f$ (b) ASI$_f$

Fig. 3. AMI$_f$ and ASI$_f$ from front view of one sample action

As human bodies and motions are performed in 3D space, the information loss in the depth channel could cause significant degradation of the representation and discriminating capability for human actions. In order to make use of the additional motion information from depth maps, each depth frame can be projected onto three orthogonal Cartesian planes, as shown in Fig. 4. Considering the information from the front view is dominant for the action and the projections onto top view and side view can be very coarse due to the resolution of the depth maps, only D-MHI$_t$ and D-MHI$_s$ are computed from the projections on top view and side view, respectively. Therefore, one action depth sequence can be modeled as six templates using the 3DMTM, as shown in Fig. 1.

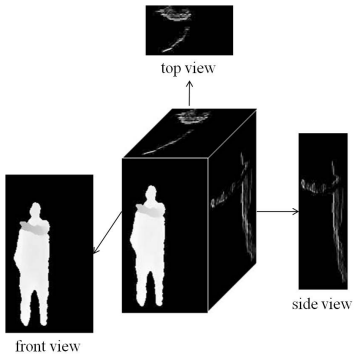


Fig. 4. Depth frame projections

B. The 3DMTM-PHOG Descriptor

Histograms of Oriented Gradients (HOG) has been experimentally proved to outperform other features to encode human figures in human detection [6]. In addition, edges and interesting regions can effectively encode object shapes and areas, and they have been widely used for action representation. Bosch *et al.* [3] extend the HOG and propose a spatial pyramid representation of object edges based on HOG to encode object shapes. In contrast to [3], the edges of human subjects are not extracted in the PHOG proposed in [27]. Instead, the gradients on the whole interesting regions are used to accumulate a histogram. However, edges or interesting regions are usually difficult to segment or extract in practice. To this end, we propose a novel descriptor based on 3DMTM and PHOG to characterize local shapes at different spatial scales for action recognition. The proposed 3DMTM-PHOG descriptor is extracted from the calculation of gradients in a dense grid of the 3DMTM to encode human actions representation. It is

directly performed on the six templates from the 3DMTM, which requires no edge or interesting regions extraction.

In HOG, the magnitude $m(x, y)$ and orientation $\theta(x, y)$ of the gradient on a pixel (x, y) are calculated as:

$$m(x,y) = \sqrt{g_x(x,y)^2 + g_y(x,y)^2} \tag{7}$$

$$\theta(x,y) = \arctan \frac{g_x(x,y)}{g_y(x,y)} \tag{8}$$

where $g_x(x, y)$ and $g_y(x, y)$ are image gradients along the x and y directions, respectively. In 3DMTM-PHOG, each template is divided into small spatial grids in a pyramid way at different pyramid levels. The pyramid at level l has $2^l \times 2^l$ grids, as shown in Fig. 5. Each gradient orientation is quantized into B bins. Gradients over all the pixels within a grid are accumulated to form a local B bins 1-D histogram. Therefore, each template from 3DMTM at level l is represented by a $B \times 2^l \times 2^l$ dimension vector. Since there are six templates in 3DMTM, we concatenate the six PHOG vectors as the 3DMTM-PHOG descriptor. The obtained feature vector, $V \in \mathbb{R}^d$ ($d = 6 \times B \times \sum_{l=1}^{L} (2^l \times 2^l)$), is the 3DMTM-PHOG descriptor of the 3DMTM. In our experiment, for $B = 9$ bins and $L = 3$ levels, the descriptor will be a 4536-dimension vector.

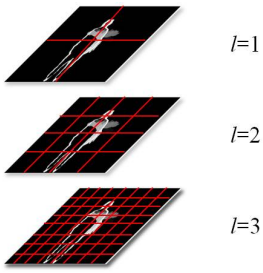


Fig. 5. 3DMTM-PHOG performed on D-MHI$_f$

The dimension of the 3DMTM-PHOG descriptor (*e.g.*, $d = 4536$ when $B = 9$, $L = 3$) is relatively high, which is not effective for classification. Many dimension reduction approaches have been proposed to solve this kind of problem. We employ the widely used principal component analysis (PCA) [19] due to its simplicity and effectiveness. Let $\Gamma \in \mathbb{R}^{p \times d}$ denote the first p principal components learned from the 3DMTM-PHOG descriptors of the training actions. We project the 3DMTM-PHOG descriptor V to the linear subspace spanned by the principal components Γ:

$$Y = \Gamma^T (V - \bar{V}) \tag{9}$$

where $\bar{V} \in \mathbb{R}^d$ is the mean 3DMTM-PHOG descriptor of all training actions, and $Y \in \mathbb{R}^p$ is the final 3DMTM-PHOG descriptor of the action. Moreover, after projecting the 3DMTM-PHOG descriptor to its principal subspaces, the speed of the action recognition can be increased without loss of accuracy.

C. Classification

As for classification, support vector machine (SVM) [4] is adopted for the final stage to classify the actions. A well-known SVM library LIBSVM [5] is used to train 3DMTM-PHOG and test the performance. Given that the RBF kernel can non-linearly map samples into a higher dimensional space, it is used to handle our case. The optimal parameters of the RBF kernel are obtained by 5-fold cross-validation procedure over the training actions.

IV. Experiments

The proposed method has been evaluated on the MSR Action3D dataset [12]. We compare the state-of-the-art methods to our approach. In all experiments, we select three levels for 3DMTM-PHOG descriptor with 95% principal components for PCA. The experimental results show that the 3DMTM-PHOG descriptor represents the human actions very well in terms of showing higher recognition accuracies.

A. MSR Action3D Dataset

The MSR Action3D dataset [12] is an action dataset of depth sequences captured by a depth sensor similar to the Kinect device. It contains 567 depth map sequences. There are 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*. Each action is performed by 10 subjects for 2 or 3 times. Some frames of the action sequences are shown in Fig. 6. The background in this dataset is preprocessed to clear the discontinuities induced from undefined depth regions. Nevertheless, this dataset is still challenging because many of the actions are highly similar to each other.

Fig. 6. Sample frames from the MSR Action3D dataset

B. Comparisons on Three Subsets

In order to evaluate the performance of the proposed method, our experiments are first conducted using different number of training samples. We follow the same experimental settings as [12] to divide the 20 actions into three subsets, each having 8 actions as listed in TABLE I. All the subsets (AS1, AS2 and AS3) are deliberately constructed so that similar actions are included within the same subset. For each subset,

Method (%)	T1				T2				CST			
	AS1	AS2	AS3	Overall	AS1	AS2	AS3	Overall	AS1	AS2	AS3	Overall
Bag of 3D Points [12]	89.5	89.0	96.3	91.6	93.4	92.9	96.3	94.2	72.9	71.9	79.2	74.7
Histograms of 3D Joints [28]	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 [29]	94.7	95.4	97.3	95.8	97.3	98.7	97.3	97.8	74.5	76.1	96.4	82.3
3DMTM-PHOG	**97.3**	**97.4**	**98.7**	**97.8**	**100.0**	**100.0**	**100.0**	**100.0**	**93.4**	**82.3**	**96.4**	**90.7**

TABLE II. PERFORMANCE EVALUATION OF PROPOSED METHOD ON THREE SUBSETS

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)	Forward kick (FK)	Golf swing (GW)
Pickup & throw (PT)	Side boxing (SB)	Pickup & throw (PT)

TABLE I. THE THREE ACTION SUBSETS

Method	Accuracy (%)
Recurrent Neural Network [15]	42.5
Dynamic Temporal Warping [16]	54.0
Hidden Markov Model [14]	63.0
Bag of 3D Points [12]	74.7
Histogram of 3D Joints [28]	79.0
Eigenjoints [29]	82.3
STOP Feature [24]	84.8
Random Occupancy Pattern [25]	86.2
Actionlet Ensemble [26]	88.2
3DMTM-PHOG	**90.7**

TABLE III. EVALUATION OF METHODS ON THE CROSS SUBJECT TEST

there are three different tests, *i.e.* Test One (T1), Test Two (T2), and Cross Subject Test (CST). In Test One, 1/3 of the samples are used as training and the rest as testing; in Test Two, 2/3 samples are used as training and the rest as testing; in the Cross Subject Test, subject 1, 3, 5, 7, 9 are used for training and 2, 4, 6, 8, 10 are used for testing. Since different subjects have their own styles to perform actions, there are large variations among training and testing actions in the Cross Subject Test.

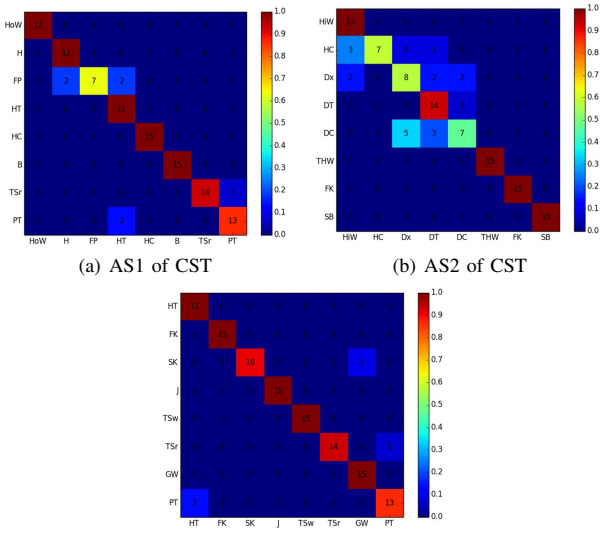
We compare our proposed method with other methods on the three subsets, and the overall accuracies are also provided for each test. As shown in TABLE II, the performance of **3DMTM-PHOG** is superior to other methods in terms of accuracies on all tests. The Bag of 3D Points [12] is a method proposed for action recognition based on the original depth data, while Histograms of 3D Joints [28] and Eigenjoints [29] rely on the estimation of joints positions. The results reflect the robustness of our proposed method, and demonstrate the 3DMTM-PHOG can represent distinctive features of human actions. Especially, our method obtains 100% recognition accuracy on Test Two, and outperforms other methods by 8%∼16% on the Cross Subject Test.

C. Comparisons on Cross Subject Test

We then compare the proposed method with other methods which have been already widely used for action recognition (*i.e.* Recurrent Neural Network [15], Dynamic Temporal Warping [16], and Hidden Markov Model [14]) on Cross Subject Test. The Cross Subject Test is more challenging because of the considerable variations in actions performed by different subjects. Cross subjects generate much larger intra-class variance than non-cross subjects.

TABLE III shows the experimental results by various methods. Our proposed method achieves the highest recognition accuracy of 90.7% on Cross Subject Test. Note that the Acionlet Ensemble [26] requires a feature selection process from 3D joint features and a multiple kernal learning process based on the SVM classifier to achieve the accuracy of 88.2%, whereas our proposed method is based on the original depth data without relying on the estimation of 3D joints positions. Especially, considering the large intra-class variations in this dataset, the proposed framework is quite robust.

Furthermore, confusion matrices of our method on Cross Subject Test are shown in Fig. 7. Actions with high similarity could produce relative low accuracies. In AS1, most actions are confused with *forward punch (FP)*, especially for *Hammer (H)* and *High throw (HT)*. In AS2, *Draw x (Dx)*, *Draw tick (DT)*, and *Draw circle (DC)* are confused between each other, as they have highly similar movements. Since actions in AS3 have significant differences, the recognition results are better than the other subsets.



(a) AS1 of CST (b) AS2 of CST

(c) AS3 of CST

Fig. 7. Confusion matrices on Cross Subject Test

V. CONCLUSION

This paper presents a new effective framework to perform human action recognition on depth sequences. The framework is based on the proposed 3DMTM-PHOG descriptor, which can better represent the human actions in a compact and discriminative way. The 3DMTM is able to capture the motion information and static posture information from front/top/side/

views. To encode the feature from the templates of 3DMTM, we propose to represent each template using PHOG descriptor in different degree of details according to the selected pyramid levels. The proposed 3DMTM-PHOG descriptor requires no edge or interesting regions extraction, which is usually necessary in some other methods. We adopt PCA to project the descriptor onto its principal subspaces to reduce the redundancy and increase the speed of recognition. The experimental results on MSR Action3D dataset demonstrate the effectiveness and robustness of the proposed 3DMTM-PHOG descriptor. The proposed method obtains 100% recognition accuracy on Test Two, and on the most challenging Cross Subject Test it obtains 90% recognition accuracy, which significantly outperforms the existing methods. Our future work will focus on multimodal information for action recognition, *i.e.* combining joint positions and original depth data, to improve recognition accuracy in the Cross Subject Test.

REFERENCES

[1] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[2] 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, pp. 257–267, 2001.

[3] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.

[4] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[5] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

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

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[8] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3d object dataset: Putting the kinect to work," in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 141–165.

[9] A. Klser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *In BMVC08*.

[10] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[12] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[13] B. Liang and L. Zheng, "Three dimensional motion trail model for gesture recognition," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, Dec 2013, pp. 684–691.

[14] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 359–372.

[15] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1033–1040.

[16] M. Müller and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2006, pp. 137–146.

[17] O. Oreifej, Z. Liu, and W. Redmond, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[18] C. Rao and M. Shah, "View-invariance in action recognition," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2. IEEE, 2001, pp. II–316.

[19] J. Shlens, "A tutorial on principal component analysis," *Systems Neurobiology Laboratory, University of California at San Diego*, 2005.

[20] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[21] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2004–2011.

[22] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 3, pp. 313–323, 2012.

[23] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[24] 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," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2012, pp. 252–259.

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

[26] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[27] J. Wang, P. Liu, M. F. She, A. Kouzani, and S. Nahavandi, "Human action recognition based on pyramid histogram of oriented gradients," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2449–2454.

[28] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[29] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 14–19.