Spatio-Temporal Pyramid Model Based on Depth Maps for Action Recognition

Haining Xu1, Enqing Chen1, Chengwu Liang1, Lin Qi1, Ling Guan2
1Department of Information Engineering, Zhengzhou University, Zhengzhou, China
1 hxu@gs.zzu.edu.cn
2Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada
2 lguan@ee.ryerson.ca

Abstract—This paper presents a novel human action recognition method by using depth maps. Each depth frame in a depth video sequence is projected onto three orthogonal Cartesian planes. Under each projection view, we divide the entire depth maps into several sub-actions. The absolute difference between two consecutive projected maps is accumulated through a depth video (several sub-actions) sequence to form a Depth Motion Map (DMM) to describe the dynamic feature of an action. Also the difference within the threshold between two consecutive projected maps is calculated through the entire depth video to form another kind of Depth Static Map (DSM) to describe the static feature. Collectively, we call them Temporal Pyramid of Depth Model (TPDM). Then Spatial Pyramid Histograms of Oriented Gradient (SPHOG) is computed from the TPDM for the representation of an action. For classification, we apply support vector machine (SVM) to classify the proposed descriptors based on MSR Action3D dataset. Experimental results demonstrate the effectiveness of our proposed method.

Keywords- human action recognition; depth Maps; TPDM; Depth Model; Depth Static Map; SPHOG.

I. INTRODUCTION

In computer vision and pattern recognition, human action recognition is an active branch. There are many potential applications in human-machine interactivity, including video analysis, surveillance systems, robotics and so on. Although researches in action recognition have advanced rapidly in recent years, it mainly concentrates on video sequences captured by RGB cameras [1] Among the different types of features for representation, spatio-temporal interest points [2] and silhouettes are commonly used [3]. Since human actions are always performed in 3D space, it is a very difficult task to capture 3D human actions by using RGB cameras. With imaging techniques advance, the introduction of cost-effective depth cameras, e.g. the Microsoft Kinect, facilitates a variety of visual recognition tasks including action recognition. Compared with RGB images, depth images provide 3D depth data, which is insensitive to changes in lighting conditions and ignores the color variability induced by clothing, skin, hair and background. Thus, recent research work has been to explore more effective methods based on depth maps. How to extract robust features from depth maps in an effective way becomes a hot topic. 3D skeleton joint points and the original depth maps are two major categories. In [4] Shotton et al. offer an effective capturing technology from a single depth map to accurately estimate the 3D skeleton joint positions. Many action recognition methods using estimated 3D skeleton joint points [5, 6] are popular in recent years. However, it needs reliable skeleton data while the estimation can't guarantee the skeleton joint points are always right.

Different from 3D skeleton joint points, the original depth maps not only provide additional body shape but also offer motion information which generate similar projections from different views to distinguish actions. All of these make original depth maps outperform 3D skeleton joint points. Fig.1 illustrates the depth maps for actions of High Throw, Pickup Throw and Draw X on MSR Action3Ddataset [6].

![Fig.1. The sampled sequences of depth maps for actions of (a) High Throw, (b) Pickup Throw, and (c) Draw X](image_url)

In this paper, we focus on recognizing human actions by using the original depth map sequences. A new descriptor TPDM-SPHOG is presented and proved to be an effective way to solve the human action recognition problem. X. Yang et al. [7] proposed Depth Motion Map (DMM) of the entire video (one action) to described the dynamic feature of an action. In our method, we further subdivide a depth video into a set of sub-actions by using an adaptive temporal pyramid.

* Corresponding author Email: hxu@gs.zzu.edu.cn
The DMM on the sub-actions can capture the spatial layout and temporal order in a global way which is more discriminative than the DMM on the entire video. Besides, in order to make full use of the depth sequences, a novel descriptor is proposed. We call it Depth Static model (DSM), which can describe the static feature of an action. We call all of them Temporal Pyramid of Depth Model (TPDM): 4 sub-actions DMM and 1 DSM on each orthogonal Cartesian plane. Inspired by the success of HOG in human detection [8], we employ a spatial pyramid representation based on HOG [9] to encode TPDM for action recognition. In [10], J. Wang et al. extracted edges of objects and captured the interesting regions, and then extracted SPHOG from the template for action recognition. In order to reduce the amount of calculation, we directly extract SPHOG from the entire templates of TPDM. The proposed TPDM-SPHOG descriptor is more discriminative than the original depth maps, but it contains abundant data which could result in high computational costs. So Principal Component Analysis (PCA) is used to reduce redundancy of the TPDM-SPHOG descriptor and speed up the process of action recognition. Support vector machine (SVM) [11] was used to recognize action at last. The experiments on MSR Action3D dataset [6] demonstrate that the TPDM-SPHOG descriptor is more robust and achieves better recognition accuracy than the state-of-the-art methods.

The rest of the paper is organized as follows. In Sect. 2, related work is presented. In Sect. 3, the details of generating TPDM-SPHOG feature descriptors are offered. A variety of experimental results and discussions are presented in Sect. 4. Conclusions are given in Sect. 5.

II. RELATED WORK

In the past few decades, researchers have explored different methods on human action recognition and obtained great success. Methods of space–time such as spatio-temporal features, trajectories, and space-time volumes have been widely used for human action recognition from traditional RGB sequences. In [12], Schuldt, C. et al. used spatio-temporal interest points as descriptor, and then an SVM classifier was applied to the descriptor. In [13], Y. Wang et al. computed optical flow vectors as the motion descriptor to represent the frames and describe actions as a bag of words. Dollar et al. [14] used histogram of video cuboids for action representation. A. Bobick et al. [15] employed motion-history images (MHI) and motion energy images (MEI) as motion templates and used the motion templates as temporal and spatial characteristics model for human action recognition.

However, all of the above features are from traditional RGB maps. In recent years, with the release of depth sensors, action recognition based on depth information has become more and more concerned by researchers. Shotton et al. [4] offered an effective capturing technology from a single depth map to accurately estimate the 3D skeleton joint positions. L. Xia et al. [5] proposed a view invariant posture representation of action recognition, which computed 3D joint locations (HOI3D) within an improved spherical coordinate system. Then LDA was used and clustered into k posture visual words. A discrete hidden Markov model was employed on these visual words for action recognition at last. J. Wang et al. [10] used the location information of the joint and captured the shapes of the area surrounding the joint, then using them as pattern feature for action recognition.

Furthermore, the most outstanding achievements were obtained on the original depth data. In [6] Li et al. used a bag of 3D points as model feature for action recognition. They sampled a group of representative 3D points from the original depth data in each frame and employed them as the posture. Then the 3D points were regained in depth maps according to the outline points. After that, a bag of 3D points model for action recognition. J. Wang et al. [16] used a weighted sampling scheme extracted random occupancy pattern (ROP) from original depth maps as the features. Then they employed a sparse coding approach to encode ROP features and used it as the last features. However, the disadvantage of their methods is with great amount of computational complexity.

III. PROPOSED METHOD

In this part, our proposed method will be introduced in details. The framework of TPDM-SPHOG for human action recognition from depth maps is demonstrated in Fig.2.

![Fig.2](image)

Fig.2. (a) is the time index and associate Motion energy used to build the adaptive temporal pyramid. The temporal segments are obtained by repeatedly and evenly subdividing the normalized motion energy vector instead of the time axis; (b) is the framework of TPDM on front view (Tennis Swing); (c) is the framework of TPDM-SPHOG (Tennis Swing)
On the temporal dimension, the entire depth maps are divided into 4 sub-actions under three orthogonal Cartesian planes. On each plane, we compute associated motion energy on the 4 sub-actions which are then stacked to obtain four DMMs. Also we stack static energy to obtain one DSM on the entire depth video. Collectively, we call them Temporal Pyramid of Depth Model (TPDM). On the spatial dimension, Spatial Pyramid Histograms of Oriented Gradient (SPHOG) is computed from the TPDM for the representation of an action.

A. Temporal Pyramid of Depth Model

The temporal pyramid was first introduced by Laptev et al. [17]. They took a rough temporal order which evenly subdivided the depth sequences by the frame index to form the temporal pyramid. O. Oreifej et al. [9] and J. Wang et al. [18] used it in depth sequences to incorporate cues from the temporal context. When performing the same action, the motion speed or frequency could have varied from person to person. Therefore it is inflexible to handle this variance by evenly subdividing a video along the time axis. In order to solve this problem, we adopt an adaptive temporal pyramid based on the motion energy, which was introduced by X. Yang et al. [19]. Different from [19], we only use 2 levels pyramid rather than 3 levels. Also there are 3 sub-actions on the first level and 2 special sub-actions (They are all obtained on the entire depth video) on the second level instead of 4 sub-actions on the first level, 2 sub-actions on the second level and 1 sub-action on the third level. The reason for the changes is 3 sub-actions could describe the entire action more sufficiently. Also it reduces computational complexity.

Depth Motion Map (DMM) was first introduced by X. Yang et al. [19]. They obtained DMM by calculating the difference between two consecutive projected maps which beyond the threshold. It can efficiently describe the dynamic feature of an action. So the same approach in [19] is adopted in our work while the procedure to obtain DMM is modified. For a given sequence, firstly we obtain the projected maps $I'_i$ ($i \in \{1,2,3\}$) by project the $i$th frame $I_i$ onto three orthogonal planes (front view, side view and top view). As observed in equation (1), the absolute difference between two consecutive maps is to generate a binary map without threshold, where $DMM'_i$ is the motion energy of the first $i$ frames. We compute the motion energy by accumulating summations of non-zero elements of binary maps to obtain the total energy of the entire depth video of an action.

$$DMM'_i = \sum_{j=1}^{i-1} |I'_j - I'_{j+1}|$$  \hspace{1cm} (1)

We can obtain $DMM'_n$ ($n$ is the total number of frames of an action) which is the depth motion map on the entire depth video. The proposed adaptive temporal pyramid is built on this measurement as shown in the left top of Fig.2 (a). We evenly subdivide the normalized motion energy vector into a set of segments, whose corresponding frame indices are used to partition a video. We use a 2-level temporal pyramid : {T1-T2, T2-T3, T3-T4}, {T1-T4}.

Now we have evenly subdivided the normalized motion energy vector into a set of segments, with equation (1) we can obtain 3 DMMs on the first level and 1 DMM on the second level at each view. As shown in Fig. 2 (b), in the front view of $DMM$s, there are three maps on the first level and one map on left of the second level. The situation of the top view and side view is the same with the front view. Therefore we have twelve $DMM$ for each action in total.

On the other side, $DMM$ only represents motion information in the depth sequences. In order to make full use of the depth sequences, a novel descriptor is proposed. We call it Depth Static model (DSM). DSM is calculated by obtaining the difference within the threshold between two consecutive projected maps as:

$$DSM'_i = \sum_{j=1}^{i-1} |I'_j - I'_{j+1}| \leq \Phi$$  \hspace{1cm} (2)

Where $\Phi$ is the threshold. Note that $DSM$ is obtained on the entire depth sequences. As we can see in Fig. 2 (b), the last map is the $DSM$ whose energy largely concentrated on the motionless part. Therefore, $DSM$ effectively captures the characteristics of motionless part in a distinguishable way, making the descriptor more discriminative. Now we obtained 4 $DMM$ and 1 $DSM$ on each view, so there are $(4+1)\times3=15$ templates for each action. These 15 templates (12 DMMs and 3 DSMs) form the TPDM.

B. The TPDM-SPHOG Descriptor

Histogram of Oriented Gradients (HOG) is a kind of characteristics for target recognition. With HOG, interesting regions and edges can efficiently encode object areas and shapes. The proposed TPDM-SPHOG descriptor is obtained by computing gradients in a dense grid of the TPDM, which is employed to encode human actions. It is directly implemented on the 15 templates from the TPDM and not required to extract interesting regions or edge.

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

$$m(x,y) = \sqrt{g_x(x,y)^2 + g_y(x,y)^2}$$  \hspace{1cm} (3)

$$\theta(x,y) = \arctan \frac{g_y(x,y)}{g_x(x,y)}$$  \hspace{1cm} (4)

Where $g_x(x,y)$ and $g_y(x,y)$ are image gradients along the $x$ and $y$ direction. As shown in Fig.2 (c), firstly we place each template (12 $DMM$s and 3 $DSM$s in one action) at $L$ levels (In our experiment, $L=3$). Then each template is divided into small spatial grids at different levels. When $L=1$, this level has $2 \times 2$ spatial grids. When $L=2$, it has $4 \times 4$ spatial grids. When $L=3$, the number of spatial grids is $12 \times 9$ . Then Spatial Pyramid Histograms of Oriented Gradient are computed on these templates. In preparation for computing the TPDM-SPHOG descriptors, a local histogram of B bins is formed and the values of the bins are accumulated. Firstly the gradient orientation of each pixel within a grid is quantized into one of B bins. Then the corresponding bin in the local histogram is incremented by the value of the gradient. Since there are 15 templates in one action, we concatenate the
15 SPHOG vectors as the TPDM-SPHOG descriptor, a feature vector H gather all the TPDM-SPHOG together. The dimension of H is $\text{Dim} - H = (2 \times 2 + 4 \times 4 + 12 \times 9) \times 15 \times B$, Since the Dim-H of TPDM-SPHOG is too large, we reduce the dimension using PCA. Totally, compared with the reference [19], our method has improved the features for human action recognition in the following three aspects: First, the image resizing is applied to $\text{DMMs}$ but not to each projected depth map. Second, the HOG is computed on TPDM (4 $\text{DMMs}$ and 1 $\text{DSM}$) instead of 1 $\text{DMM}$ at each view. Third, the HOG in our experiment is 3-level SPHOG rather than only one level HOG.

C. Classification

As for classification, support vector machine (SVM) [11] is employed for the last stage to classify the actions. SVM has been extensively used in computer vision to achieve the state-of-the-art performances in image and video classifications, a well-known SVM library LIBSVM [20, 21] which is used to train TPDM-SPHOG and test the performance.

IV. Experiment and Discussion

In this section, we apply our method to the public domain Microsoft Research (MSR) Action3D dataset [6] with the depth map sequences. Besides, we compare our method with the existing depth-based methods. The methods designed for color-based videos are not included in our comparisons because they have been widely shown to be unsuited for depth maps [22-24].

A. Experimental Setup

MSR Action3Ddataset [6] has 567 depth map sequences captured by an RGBD camera. This dataset includes 10 subjects, 20 actions, where each subject perform each action 2 or 3 times. These 20 actions are listed in Table I. In order to evaluate the performance of the existing methods we employ the same experimental settings as [6], dividing the 20 actions into three subsets, where each subset having 8 actions. Within all the subsets (AS1, AS2 and AS3) the similar actions are deliberately constructed together. As for each subset, there are three different tests: Test One, Test Two, and Cross Subject Test. Since each subject perform each action 2 or 3 times, the first action of the subset is used as training and the last 1 or 2 as testing in Test One; in Test Two, the first two actions of the subset is used as training and the rest as testing; in Cross Subject Test, we take subject 1, 3, 5, 7, 9 for training and 2, 4, 6, 8, 10 for testing. The Cross Subject Test is challenging because different subjects have their own styles to perform actions, so there are large variations among training and testing actions.

In part III, we discuss the TDPM, but the $\text{DMM}$ and $\text{DSM}$ on the MSR Action3Ddataset are not obtained by stacking energy across an entire depth video sequence but remove the beginning M (in our experiment, M=3) and last M frames. For the first M and last M frames, the subjects were just at a stand-still position and had small body movements. The small body movements will result in a stand-still body shape with large pixel values along the edges, which may produce a large amount of recognition errors. Thus the beginning M and last M frames may lead to the redundancy of information rather than contribute to the motion characteristics. Then we remove the all zero rows and columns in the template. Besides, different subjects have different figures and different action video sequences have different sizes. The template size may different from each other. So, we resize all templates at the same size under the same projection view in order to reduce the intra-class variability. The size on MSR Action3Ddataset after resize is $100 \times 75$.

<table>
<thead>
<tr>
<th>Action Set 1 (AS1)</th>
<th>Action Set 2 (AS2)</th>
<th>Action Set 3 (AS3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal Wave (2)</td>
<td>High Wave (1)</td>
<td>High Throw (6)</td>
</tr>
<tr>
<td>Hammer (3)</td>
<td>Hand Catch (4)</td>
<td>Forward Kick (14)</td>
</tr>
<tr>
<td>Forward Punch (5)</td>
<td>Draw X (7)</td>
<td>Side Kick (15)</td>
</tr>
<tr>
<td>High Throw (6)</td>
<td>Draw Tick (8)</td>
<td>Jogging (16)</td>
</tr>
<tr>
<td>Hand Clap (10)</td>
<td>Draw Circle (9)</td>
<td>Tennis Swing (17)</td>
</tr>
<tr>
<td>Bend (13)</td>
<td>Hands Wave (11)</td>
<td>Tennis Serve (18)</td>
</tr>
<tr>
<td>Tennis Serve (18)</td>
<td>Forward Kick (14)</td>
<td>Golf Swing (19)</td>
</tr>
<tr>
<td>Pickup Throw (20)</td>
<td>Side Boxing (12)</td>
<td>Pickup Throw (20)</td>
</tr>
</tbody>
</table>

**TABLE I**

**ACTION SUBSETS AND TESTS USED IN OUR EXPERIMENTS**

B. Evaluations of TPDM-SPHOG and Comparisons with The State-of-the-arts

In order to balance the computing speed and recognition accuracy, we take B=9, L=3 in our experiment. Since the dimension of TPDM-SPHOG is too large, PCA was used to reduce the dimension. Fig.3 shows the recognition accuracies of TPDM-SPHOG with different dimensions under various test sets. We can see that recognition accuracy changes little as the dimension changes. Taking into account both recognition accuracy and computational cost, we choose 108 as the dimensionality for TPDM-SPHOG in all of our experiments.

Under B=9, L=3 and 108 dimensions, we compare our method with other methods on three subsets, and the overall accuracies are also provided for each test, as shown in Table II. We can see that the accuracies of the experiments (Test
One, Test Two, Cross Subject Test) on MSR Action3D dataset are all over 95% and the performance of our method is superior to other methods in terms of accuracies in most of the tests. The HOI3D [5] and EigenJoints [25] are methods proposed for action recognition relying on the estimation of joints positions, while the Bag of 3D Points [6], STOP [26] and Depth Motion Maps [7] are methods proposed for action recognition based on the original depth maps. All these method are tested at the same group situation on MSR Action3Ddataset [6]. Besides, we mainly compare the recognition accuracy of our method with [7]. As shown in Fig.4, the recognition accuracy of our method is much higher than the method in [7] when using only SPHOG to encode DMM of 12 sub-actions (DSM is not used). It shows that the proposal of the S-DMM significantly enhances the performance of recognition. After joined the DSM, the ascension of recognition accuracy is even more obvious, clearly showing the effectiveness of our method.

C. Comparisons on Cross Subject Test

When performing the same action, different subjects have various sizes and speeds, which may generate much larger intra-class variance on Cross Subject Test. That is why the Cross Subject Test is more challenging than non-cross subjects Test (Test One and Test Two). Table III shows Performance evaluation of existing methods and our method on Cross Subject Test. Our method performs pretty well and achieves the highest recognition accuracy of 96.1% (average). Especially, 99.1% on AS1 have an almost perfect accuracy which is much higher than the existing methods. The 92.9% on AS2 is significantly outperforms any other methods. Besides, the computational cost of our proposed method is much lower than other methods for our descriptor is based on the original depth data without relying on the estimation of 3D joints positions.

To study the results further, we show the confusion matrix of AS2 on Cross Subject Test in Fig. 3 which is the only one displaying troublesome numbers in Table II. In Fig. 5, Draw X (7) and Draw Circle (9) are mutually confused as they contain highly similar poses. It is hard to distinguish even though we directly observe these two actions with our eyes.
V. CONCLUSIONS

In this paper, we propose a low-cost computational TPDM-SPHOG descriptor and use the support vector machine (SVM) to recognize actions. On the temporal dimension, Temporal Pyramid Depth Model (TPDM) was proposed to describe the action. The entire depth maps are divided into several sub-actions under three orthogonal Cartesian planes. Then we obtain four DMM to describe the dynamic feature of an action and one DSM to describe the static feature at each view of an action sequence. On the spatial dimension, Spatial Pyramid Histograms of Oriented Gradient (SPHOG) is computed from the TPDM for the representation of an action. The recognitions rate of 96.1% on Cross Subject Test outperforms the existing methods being compared. Our future work will focus on combining joint positions and original depth data for recognition, to further improve recognition accuracy in 3D human action recognition.

VI. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (61201251, 61331021, 61210005).

REFERENCES