Human Behavior Recognition Based on Axonometric Projections and PHOG Feature

Xiaoxia SHEN\textsuperscript{1,2}, Hua ZHANG\textsuperscript{1,2}, Zan GAO\textsuperscript{1,2,*}, Yanbin XUE\textsuperscript{1,2}, Guangping XU\textsuperscript{1,2}

1 Key Laboratory of Computer Vision and System, Tianjin University of Technology, Ministry of Education, Tianjin 300384, China
2 Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China

Abstract

Human behavior recognition has become a hot research topic in computer vision. In this paper, we propose an effective method to recognize human behaviors from sequences of depth maps, which provide additional body shape and motion information for behavior recognition. In our approach, we construct a novel difference motion history image, and propose axonometric projection to capture the target motion process, after that, pyramid histogram of orientated gradients is extracted for each view to describe the target motion. Large scale experimental results on challenging and public MSR-action3D behavior dataset demonstrate that the performances of our difference motion history image and our descriptors are satisfying, what’s more, our proposed axonometric projection approach further improves the performance, which significantly outperforms the state-of-the-art methods.

Keywords: Behavior Recognition; Difference Motion History Image; Axonometric Projection; Pyramid Feature

1 Introduction

Human behavior recognition is a widely studied area in computer vision. Its applications include surveillance systems, video analysis, robotics and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces. Behavior recognition including two parts: behavioral description and behavior classification. Currently, the focus of behavior recognition gradually concentrated on the behavior description. Researchers have explored different compact representations of human actions in the past few decades. Here we mainly divide them into 3 categories: 1) silhouette/contour/shape based representation; this is an effective representation to describe the shape of the body postures. Methods proposed in
the past for silhouette/contour/shape based behavior recognition can be divided into two major categories. One is to extract behavior descriptors from the sequences of silhouettes. Conventional classifiers are frequently used for recognition [1-4]. The other one is to extract features from each silhouette and model the dynamics of the action explicitly [5-7]. Although the above methods can get good performance, but they require background modeling, it is difficult to obtain relatively clean outlook, which will affect the subsequent performance of the algorithm. 2) joint/body parts based representation; this representation has also been popular for a few decades. Joints data are usually acquired using markers to the subjects and use multiple cameras to get the 3D positions. There are plenty of works focusing on action recognition using joints/body parts [8-10]. Especially, 3D joint/body parts are view point invariant and appearance invariant, in that the actions vary little from different actors. But in the past, extracting body parts/3D joints accurately is a difficult task, particularly under realistic imaging conditions. As such, low-level appearance features such as spatio-temporal interest points have also been popular recently. 3) spatio-temporal interest points representation; In these approaches, behaviors are represented as a collection of visual words, which is the codebook of spatio-temporal features. Schuldt et al. [11] integrated space-time interest point’s representation with SVM classification scheme. Dollar et al. [12] employed histogram of video cuboids for behavior representation. Wang et al. [13] represent the frames using the motion descriptor computed from optical flow vectors and represent actions as a bag of coded frames.

However, these researches of human behavior recognition mainly concentrate on video sequences captured by traditional RGB cameras, their performances often be affected by illumination changes, shadows, and environmental changes et al. On the other hand, the human motion is articulated, and capturing such highly articulated motion from monocular video sensors is a very difficult task. Those difficulties largely limit the performance of video-based human behavior recognition, as indicated in the studies in the past decade. The recent introduction of the Microsoft Kinect [14] (Fig. 1 shows several depth maps from MSR-Action3D Dataset [16]) may change the picture by providing 3D depth data of the scene, which largely eases the task of object segmentation. Kinect is not only able to obtain RGB information, but also the corresponding depth information which is very helpful for our task. Depth maps are able to provide additional body shape information to differentiate behaviors that have similar 2D projections from a single view. There are a few works on the recognition of human actions from depth data in the past three years. Li et al. [16] employed an action graph to model the dynamics of the actions and sample a bag of 3D points from the depth map to characterize a set of salient postures that correspond to the nodes in the action graph. 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 [17] proposed an EigenJoints-based action recognition system by using a NBNN classifier. However, the 3D positions of skeleton joints might be complete wrong if there are sever occlusions. Sung et al. [18] extract features from the skeleton data provided by Prime Sense from RGBD data from Kinect and use a supervised learning approach to infer activities from RGB and depth images from Kinect.

So far, human behavior recognition based on the depth information has made some achievements, but it still does not have robust descriptors, and their accuracies still cannot satisfactory. Thus, in this paper, human behavior recognition based on PHOG feature and axonometric projection is proposed. Specifically, for depth image sequences, we construct a novel difference motion history image to represent how the human is moving. Similar to MHI [1], the proposed difference MHI also stacks foreground motion regions to record where and how actions are per-
Fig. 1: Examples of the sequences of depth maps for behaviors: high arm wave, tennis serve, side kick and golf-swing

formed, however, MHI only keeps most nearest motions to capture the latest action, while our proposed difference MHI accumulates global activities through entire video sequences to represent the motion intensity. After that, axonometric projection to capture the target motion process is proposed, which can capture the global activities from front/top/left axonometric projections. Finally, pyramid histogram of orientated gradients is extracted for each axonometric projection to represent the target motion. Large scale experimental results on challenging and public MSR-action3D dataset show that our difference motion history image and our descriptor performance is much better than traditional motion history image, what is more, our axonometric projection approach further improves the performance, which significantly outperforms the state-of-the-art methods, and whose best performances on MSR-action3D of Cross Subject Test reach 88.8%.

The rest of this paper is organized as follows: In Section 2 details behavior representations using depth information. Section 3 introduces the Pyramid Feature. The experimental results are shown in Section 4. Finally, Section 5 concludes this paper.

2 Behavior Visual Representations

In this section, we will introduce two novel behavior representations to depict motion change process: depth difference motion history image (DDMHI) and axonometric projections.

2.1 Depth difference motion history image

To represent the change in depth for the motion in N frames, we find the difference between maximum and minimum of the nonzero values of depth across N frames for each pixel location. Fig. 2(b) illustrates the depth difference motion history image (DDMHI) for N frames of boxing behavior. Here, Eq. (3) represents the depth difference image for N frames.

\[
I_{\text{max}}(i, j) = \max\{D(i, j, t) : D(i, j, t) \neq 0, t \in [1...N]\}  
\]

\[
I_{\text{min}}(i, j) = \min\{D(i, j, t) : D(i, j, t) \neq 0, t \in [1...N]\}  
\]

\[
I_{\text{diff}} = I_{\text{max}} - I_{\text{min}}  
\]

Where \(D\) denotes depth map, \(I_{\text{max}}\) and \(I_{\text{min}}\) are maximum and minimum of the nonzero values of depth across \(N\) frames for each pixel location, \(I_{\text{diff}}\) denotes the depth difference motion history image.
2.2 Axonometric projection to capture motion history processing

In above 2.1 section, we have introduced DDMHI to capture human motion change processing, this image is obtained through the original depth image sequences, which are general description of the human behavior, and we treat them as a front axonometric projection (called FAP). Since an object can obtain different information from different perspectives, and this information plays a complementary role in behavior recognition, thus we propose an effective and efficient approach to depict human behavior, namely, axonometric projections (front axonometric projection, top axonometric projection and left axonometric projection). The following section will introduce these axonometric projections in detail.

A. Top Axonometric Projection (TAP)

First, we calculate the frame difference for two consecutive source images (Assuming the image size is m * n) from N frames depth video sequences to obtain a length of N-1 of frame difference video sequences (See Eq. (4)). Then, we accumulate the each pixel value by the column, thus, for each frame difference image, we can represent it by a 1*n row vector (as shown in Eq. (5)). Finally, we concatenate and accumulate all the row vectors obtained from all frame difference images to obtain an image whose size is (N-1)*n, and the image is named as top axonometric projection (TAP). For example, the first row vector in TAP is from the first frame difference image, and the second row vector in TAP is from the second frame difference image. The defined of TAP is given as following:

\[ B(i, j, t) = D(i, j, t + 1) - D(i, j, t), t \in [1...N - 1] \]  
\[ D_T(1, j, t) = \sum_j B(i, j, t), t \in [1...N - 1] \]

where, \( D \) is the depth map, \( B(i, j, t) \) denotes the frame difference image, \( D_T(1, j, t) \) denotes a row vector get from the frame difference image.

B. Left Axonometric Projection (LAP)

To obtain left axonometric projection, we also first calculate the frame difference for two consecutive source images to obtain a length of N-1 of frame difference video sequences (See Eq. (4)). Then, we accumulate the value of each pixel by the row, thus, for each frame difference image, we can represent it by an m*1 column vector (as shown in Eq. (6)). Finally, we concatenate and accumulate all the column vectors obtained from all difference images to obtain an image whose size is (m*(N-1)), and the image is named as left axonometric projection (LAP). The definition of LAP is given as following:

\[ D_L(i, 1, t) = \sum_i B(i, j, t), t \in [1...N - 1] \]
where $B(i, j, t)$ denotes the frame difference image and $D_L(i, 1, t)$ denotes a column vector obtained from the frame difference image. Front, Top and Left axonometric projections of two hand wave behavior are shown in Fig. 3. From them, we can understand that Top and Left axonometric projections also can provide additional shape information, which also are complementary for front axonometric projection.

![Fig. 3: Front, Top and Left axonometric projections of two hand wave behaviour on depth channel](image)

3 Pyramid Feature (PHOG)

Histogram of Orientated Gradients (HOG) [19] is an effective method to describe the body shape information. HOG can be well characterizes the target edges or gradient structure by extracting an edge or gradient distribution in a localized area, and further characterizes the human body shape. In fact, HOG considered the distribution of the image spatial location, but did not take into account the division of the image at different spatial scales on the classification performance. Based on this, we adopt the Pyramid Histogram of Orientated Gradients feature (PHOG [20]) to represent motion history images (DDMHI and axonometric projections). PHOG not only describes the global shape and the local details of human behavior, but also depicts the spatial information of human behavior, which is very helpful for behavior recognition. The representation is illustrated in Fig. 4, and the processing of extracting is shown as follows:

1) We extract edge contour of DDMHI using canny edge detection, which is used to describe shapes.

2) Pyramid segmentation: In our method, the DDMHI is divided into four layers, the first layer is the entire image; the second layer is that the entire image is divided into four sub-regions; while the third and fourth layer are that the previous sub-regions are further divided into four smaller sub-regions. Obviously, the size of sub-region in each layer is 1/4 of previous layer region.

3) After that, HOG vector is computed for each sub-region at each pyramid resolution level. Here, an image sub-region is quantized into $K$ bins and the orientation bins are evenly spaced over $0^\circ - 180^\circ$ (“unsigned” gradient) or $0^\circ - 360^\circ$ (“signed” gradient). Each bin in the histogram represents the number of edges that have orientations within a certain angular range.

4) The final PHOG descriptor for the image is a concatenation of all the HOG vectors. Our algorithm use four-tier structure, the orientation bins are evenly spaced over $0^\circ - 360^\circ$ and $K=20$ bin, thus, for each DDMHI, the dimension of PHOG is $(4^0 + 4^1 + 4^2 + 4^3) \times 20 = 1700.$
4 Experimental Evaluations on MSR Action3D Dataset

To illustrate the effectiveness and superiority of the axonometric projections and PHOG feature, we evaluate the proposed method on MSR Action3D dataset with SVM classifier (RBF kernel). In addition, we extensively compare our approach with the state-of-the-art methods under a variety of experimental settings.

4.1 Database creation and experimental setup

A. MSR Action3D Dataset

MSR-Action3D dataset is a behavior dataset of depth sequences captured by a depth camera-kinect. This dataset contains twenty behaviors: 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 behavior was performed by ten subjects for three times. The frame rate is 15 frames per second and resolution 640*480. Altogether, the dataset has 23797 frames of depth map for 402 behavior samples. Some examples of the depth sequences are shown in Fig. 1.

Those behaviors were chosen to cover various movements of arms, legs, torso and their combinations, and the subjects were advised to use their right arm or leg if an action is performed by a single arm or leg. Although the background of this dataset is clean, this dataset is challenging because many of the actions in the dataset are highly similar to each other.

B. Experimental Setup

In order to facilitate a fair comparison, we follow the same experimental settings as [16] to split 20 categories into three subsets as listed in Table 1. The AS1 and AS2 were intended to group behaviors with similar movement, while AS3 was intended to group complex behaviors together. In each subset, there are three different tests: Test One, Test Two, and Cross Subject Test. In Test One, 1/3 of the subset is used as training and the rest as testing; in Test Two, 2/3 of the subset is used as training and the rest as testing. Both of them are non-cross-subject tests. In Cross Subject Test, 1/2 of subjects are used as training and the rest ones used as testing.
Table 1: Three subsets of behaviors used in the experiments

<table>
<thead>
<tr>
<th>Action Set 1 (AS1)</th>
<th>Action Set 2 (AS1)</th>
<th>Action Set 3 (AS1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>horizontal arm wave</td>
<td>high arm wave</td>
<td>high throw</td>
</tr>
<tr>
<td>hammer</td>
<td>hand catch</td>
<td>forward throw</td>
</tr>
<tr>
<td>forward punch</td>
<td>draw x</td>
<td>side kick</td>
</tr>
<tr>
<td>high throw</td>
<td>draw tick</td>
<td>jogging</td>
</tr>
<tr>
<td>hand clap</td>
<td>draw circle</td>
<td>tennis swing</td>
</tr>
<tr>
<td>bend</td>
<td>two hand wave</td>
<td>tennis serve</td>
</tr>
<tr>
<td>tennis serve</td>
<td>forward kick</td>
<td>golf swing</td>
</tr>
<tr>
<td>pick up &amp; throw</td>
<td>side-boxing</td>
<td>pick up &amp; throw</td>
</tr>
</tbody>
</table>

4.2 The results of axonometric projections

A. Evaluations of APS_PHOG

We first evaluate the effect of axonometric projections and PHOG feature to recognition performances. We extract PHOG feature in front axonometric projection (DDMHI), top axonometric projection and left axonometric projection respectively, and concatenated those PHOG feature as the final feature for axonometric projections, namely, Axonometric Projections_PHOG (APS_PHOG) feature. We use the three-tier structure on TAP and LAP, and when K=20 bin the performance is best, so each TAP and LAP generates PHOG feature with the dimension of \((4^0 + 4^1 + 4^2) \times 20 = 420\). So, the dimension of final feature (APS_PHOG feature) for axonometric projections is \(1700 + 420 + 420 = 2540\). Table 2 shows the results of APS_PHOG compared to the DDMHI_PHOG feature.

From Table 2, we can see that the APS_PHOG feature significantly outperforms the DDMHI_PHOG feature in all testing cases (Test One, Test Two, and Cross Subject Test) on three subsets, and it obtain very high recognition accuracy. This indicates that we add TAP and LAP can get more useful information for human behaviour representation, and they did improve the accuracy of behaviour recognition. This further shows that our proposed Pyramid feature has strong robustness and stability.

Table 2: Performance comparison of APS_PHOG and DDMHI_PHOG feature

<table>
<thead>
<tr>
<th></th>
<th>Fetre</th>
<th>Test One</th>
<th>Test Two</th>
<th>Cross Subject Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1</td>
<td>DDMHI_PHOG</td>
<td>85.6</td>
<td>94.1</td>
<td>79.2</td>
</tr>
<tr>
<td></td>
<td>APS_PHOG</td>
<td>94.8</td>
<td>97.8</td>
<td>90.6</td>
</tr>
<tr>
<td>AS2</td>
<td>DDMHI_PHOG</td>
<td>86.4</td>
<td>92.0</td>
<td>73.7</td>
</tr>
<tr>
<td></td>
<td>APS_PHOG</td>
<td>95.2</td>
<td>98.8</td>
<td>81.4</td>
</tr>
<tr>
<td>AS3</td>
<td>DDMHI_PHOG</td>
<td>90.0</td>
<td>93.5</td>
<td>85.2</td>
</tr>
<tr>
<td></td>
<td>APS_PHOG</td>
<td>97.9</td>
<td>98</td>
<td>94.6</td>
</tr>
<tr>
<td>Overall</td>
<td>DDMHI_PHOG</td>
<td>87.3</td>
<td>93.2</td>
<td>79.4</td>
</tr>
<tr>
<td></td>
<td>APS_PHOG</td>
<td>96</td>
<td>98.2</td>
<td>88.8</td>
</tr>
</tbody>
</table>
B. Comparison with State-of-the-art

We compare our APS_PHOG approach with the state-of-the-art methods including 3D Silhouettes [13] and EigenJoints [17] on the MSR Action3D dataset in Table 3. The 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. Moreover, the average accuracy of our method is higher than both 3D Silhouettes and EigenJoints. In addition to recognition accuracy, our approach is much more compact than 3D Silhouettes. And the performance of EigenJoints will affected by skeleton tracking, it is because the skeleton tracker sometimes fails and the tracked joint positions are quite noisy. So, our method has a better stability.

Table 3: Recognition accuracies of our method compared to the state-of-the-art methods

<table>
<thead>
<tr>
<th>Feature</th>
<th>Test One</th>
<th>Test Two</th>
<th>Cross Subject Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1 3D Silhouettes [16]</td>
<td>89.5</td>
<td>93.4</td>
<td>72.9</td>
</tr>
<tr>
<td>EigenJoints [17]</td>
<td>94.5</td>
<td>97.3</td>
<td>74.5</td>
</tr>
<tr>
<td>Ours</td>
<td>94.8</td>
<td>97.8</td>
<td>90.6</td>
</tr>
<tr>
<td>AS2 3D Silhouettes [16]</td>
<td>89.0</td>
<td>92.9</td>
<td>71.9</td>
</tr>
<tr>
<td>EigenJoints [17]</td>
<td>95.4</td>
<td>98.7</td>
<td>76.1</td>
</tr>
<tr>
<td>Ours</td>
<td>95.2</td>
<td>98.8</td>
<td>81.4</td>
</tr>
<tr>
<td>AS3 3D Silhouettes [16]</td>
<td>96.3</td>
<td>96.3</td>
<td>79.2</td>
</tr>
<tr>
<td>EigenJoints [17]</td>
<td>97.3</td>
<td>97.3</td>
<td>96.4</td>
</tr>
<tr>
<td>Ours</td>
<td>97.9</td>
<td>98</td>
<td>94.6</td>
</tr>
<tr>
<td>AS4 3D Silhouettes [16]</td>
<td>91.6</td>
<td>94.2</td>
<td>74.7</td>
</tr>
<tr>
<td>EigenJoints [17]</td>
<td>95.7</td>
<td>97.9</td>
<td>82.3</td>
</tr>
<tr>
<td>Ours</td>
<td>96.0</td>
<td>98.2</td>
<td>88.8</td>
</tr>
</tbody>
</table>

5 Discussion and Future Work

In this paper, we have proposed novel behavior representations Axonometric Projections and Pyramid feature for human behavior recognition with depth cameras. The compact and discriminative behavior representation is able to capture the global activities from front/top/left axonometric projections. And the proposed APS_PHOG feature is discriminative enough to classify human behaviors with subtle differences. Experimental results on MSR Action3D dataset have clearly shown the promising performance of the proposed method and also significantly outperform the existing state-of-the-art methods. In the future, we will keep our investigations along several directions. First, we aim to exploit the effectiveness of the proposed technique for the understanding of more complex activities. Next, we will keep investigating on combining RGB information with depth information to provide more information and build more robust algorithms. Finally, we will extend our approach to other real-world applications.
References