DTOP: Dense Trajectories on Planes for Action Recognition from Depth Sequences

Author1's name

Adress 1

Author2's name

Adress 2

Abstract

Dense trajectory-based approaches on 2D video have been demonstrated state-of-the-art at action recognition since it can capture most discriminative motions. However, there are not many studies related to exploiting the discriminative motions in depth video. In this work, we extend the approach on depth video and show its effectiveness for action recognition. We extract dense trajectories from 2D videos transformed from depth video and apply trajectory-aligned descriptors to calculate motion features. To obtain the 2D transformed videos, we build views, which can capture the discriminative motions similar to observing actions from different directions. We evaluate this approach on framework of action recognition using the benchmark MSR Action 3D, MSR Gesture 3D and 3D Action Pairs datasets. Evaluation results show that our proposed approach is effective for action recognition on depth video and outperforms the state-of-the-art approaches.

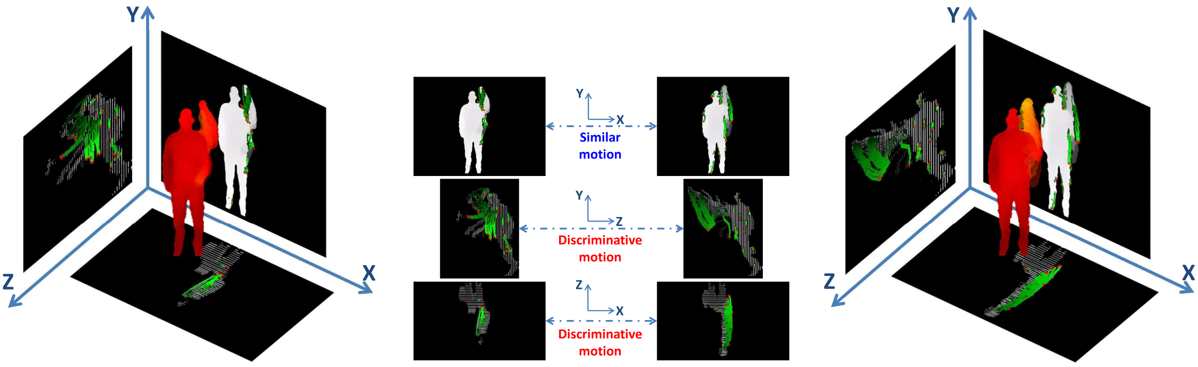
Keywords: `Dense trajectories, action recognition, depth map, projection`

1. Introduction

Action recognition in videos has been one of the active research fields in computer vision [1, 2] due to its wide applications in areas like surveillance, video retrieval, human-computer interaction and smart environments. Due to the diversity and complexity of actions, as well as complicated environment (e.g background clutter and illumination variation), action recognition is still a challenging problem. Recent approaches can be divided into three major categories: silhouette-based [3–6], salient point-based [7–12] and trajectory-based [13–15]. All approaches, basically, try to capture motion information that appears in videos, since motion is crucial information for presenting actions. Based on work of H.Wang et al. [16], dense trajectory-based approach has been demonstrated that it is the state-of-the-art approach for action recognition [17–19].

With relative works, most studies mainly investigate on video sequences captured by traditional 2D cameras. Although, there are many improvements on the approach for action recognition in domain of 2D videos [20, 21], the mentioned challenges are still difficult to handle. With the development of new RGB-D cameras, e.g. Kinect camera, capturing color images as well as depth maps has become feasible in real time. The depth maps can enrich information for cues, such as body shape and motion information. In addition, depth information is less sensitive to the challenges RGB information usually deals with. Due to these advantages, recent research trend concentrates on exploiting depth maps for action recognition [22–29]. However, with our best knowledge, none success with combining dense trajectories, the state-of-the-art approach on 2D video, and depth video. In this paper, we investigate to exploit the dense trajectory-based approach on depth video.

The key idea of the dense trajectory-based approach is to capture most discriminative trajectories in video. Therefore, in order to effectively exploit this approach on depth video, it is necessary to extract the trajectories in depth video. To do that, a straightforward method is to consider depth value as intensity value and adapts



(a) Action *Hammer* (b) Action *Forward Punch*

Figure 1: Illustration of our dense trajectory-based approach. The original sequence of depth maps is projected onto three planes, corresponding to three views: *front*, *side* and *top*, to form 2D videos. After that, the dense trajectory motion features are calculated for each 2D video.

extracting dense trajectories on 2D transformed videos. Unfortunately, the method will lead to inherent cases of the trajectory-based approaches, it is confused to identify actions contain similar motions. For example, *forward punch* and *hammer* may be confused actions, if we view them from front, since they contain similar movements respectively: “lift arm up” and “stretch out”. Obviously, it is difficult to distinguish such actions with data contains less discriminative information as depth data. This is major reason to require additional information for effectively recognizing actions.

The basis idea to deal with such cases is to observe actions from various directions. Information achieved from the view directions can provide clearer cues to discriminate such actions. To collect such information from depth video, a simple way is to project depth maps onto view planes, see figure 1. The projections are easily obtained by the mentioned advantages of depth data. Data projected on the planes is then gathered to generate corresponding 2D videos. Dense trajectory-based motion features are then calculated on 2D videos to generate a final feature representation for depth video.

To evaluate the effectiveness of our method, we conduct experiments on MSR Action 3D dataset, MSR Gesture 3D dataset and 3D Action Pairs dataset. Experimental results show that our proposed method beats the state-of-the-art methods in constrain of only using depth data. The results also present our contributions: (1) we propose an effective method to exploit trajectories in depth video, (2) we perform comprehensive experiments on the challenging benchmark dataset and indicate that our proposed method is the best when compared with the state-of-the-art depth-based methods.

After a brief review of the related work in Section 2, the proposed method is described in Section 3. Sections 4 and 5 present the experimental settings and results. In section 6 we provide some concerned discussions. The summaries of our work are given in Section 7.

2. Related Works

In terms of action recognition in 2D video, there are three popular approaches used in several action recognition systems, including silhouette-based, salient point-based and trajectory-based. The silhouette-based approach, as described in [3–6], is powerful since it encodes a great deal of information in a sequence of images. However, it is sensitive to different viewpoints, noise and occlusions. Besides, it depends on the accuracy of localization, background subtraction or tracking for exactly extracting region of interest. An other approach based on salient points generates a compact video representation and accepts background clutter, occlusions and scale changes. The effectiveness of this approach is also showed in several works [7–12]. However, in case of recognizing complicated motions, the salient point-based approach deals with several challenges, due to the lack of relationship of salient points. In recent studies [13–15], the trajectory-based approach captures moving patterns in video, thereby it provides additional information to recognize motions more exactly.

For depth video, most recent methods exploit depth information into two major directions. The first one is to adapt 2D techniques-based methods for depth data.

The second one is to use depth value as its mean.

For the first direction, Yang.X et al. [26] propose the Depth Motion Maps (DMM) to accumulate global activities in depth video sequences. The DMM are generated by stacking motion energy of depth maps projected onto three orthogonal Cartesian planes. And the Histogram of Oriented Gradients (HOG) [30] are computed from the DMM to represent an action video. Another approach proposed by Xia.L and Aggarwal.J.K [28] presents a filtering method to extract spatio-temporal interest points from depth videos (DSTIPs). In this approach, they extend a work of Dollar et al. [8] to adapt for depth data. Firstly, 2D and 1D filters (e.g. Gaussian and Gabor filters) are applied respectively on to the spatial dimensions and temporal dimension in depth video. A correction function then is used to suppress points as depth noises. Finally, points with the largest responses by this filtering method will be selected as the DSTIPs for each video. Besides, a depth cuboid similarity feature (DCSF) is proposed to describe a 3D cuboid around the DSTIPs with supporting size to be adaptable to the depth.

For the second direction, [22] used a bag of 3D points to characterize a set of salient postures. The 3D points are extracted on the contours of the planar projections of the 3D depth map. And then, about 1% 3D points are sampled to calculate feature. Unlike [22], works [23, 24, 27] use occupancy patterns to represent features in action videos.

Vieira et al. [24] proposed a new feature descriptor, called Space-Time Occupancy Patterns (STOP). This descriptor is formed by sparse cells divided by the sequence of depth maps in a 4D space-time grid. The values of the sparse cells are determined by points inside to be on the silhouettes or moving parts of the body. Wang et al. [27] presented semi-local features called Random Occupancy Pattern (ROP) features from randomly sampled 4D sub-volumes with different sizes and different locations. The random sampling is performed under a weighted scheme to effectively explore the large dense sampling space. Besides, authors also apply a sparse coding approach to robustly encode these features. The work by Wang et al. [23] designed a feature to describe the local "depth appearance" for eah joint, named Local Occupancy Patterns (LOP). The LOP features are computed based on 3D point cloud around a particular joint. Moreover, they concatenate the LOP features with skeleton information-based features and apply Short Fourier Transform to obtain the Fourier Temporal Pyramid features at each joint. The Fourier features are utilized in a novel actionlet ensemble model to represent each action video.

Recently, Oreifej and Liu [29] presented a new descriptor for depth maps, named Histogram of Oriented 4D Surface Normals (HON4D). To construct the HON4D, firstly, the 4D normal vectors are computed from the depth sequence. At the next step, the 4D normal vectors is distributed into spatio-temporal cells. To quantize the 4D normal vectors, the 4D space is quantized by using vertices of a regular polychoron. The quantization, then, is refined by additional projectors to make the 4D normal vectors in each cell denser and more discriminative. Afterwards, the HON4D features in cells are concatenated to represent a depth action video.

Inspired by results of Shotton et al. [31] and Xia.L et al. [32], the work by Yang et al. [25, 33] developed skeleton-based methods from sequence of depth maps. [25] proposed an EigenJoints-based action recognition system using a Naive-Bayes-Nearest-Neighbor classifier. The system is able to capture the characteristics of posture, motion and offset information of frames. In addition, non-quantization of descriptors and *Video-to-Class* distance computation in this work are showed effective for action recognition. In work of J.Luo [33], a new discriminative dictionary learning algorithm (DL-GSGC) was proposed to incorporate both group sparsity and geometry constraints. Besides, to keep temporal information, a temporal pyramid matching method was used on each sequence of depth maps.

Different from the previous approaches, we use a dense trajectory-based approach for action recognition. We do not require to segment human body like [22, 26]. As well as, skeleton extraction as in [23, 25] is not also required in our work. We investigate the benefit of generating 2D transformed videos from depth data, as mentioned in [22, 26]. Moreover, we leverage the effectiveness of trajectory feature to represent an action video. In our best knowledge, no work has previously proposed to adapt the dense trajectory-based approach for human action recognition in depth video. We conduct evaluations on recognition accuracy in depth video using dense trajectories proposed by Wang et al. [16].

3. Proposed Method

This paper presents an effective method for action recognition on depth video by adapting the dense trajectory-based motion feature. First, we provide a brief review of the dense trajectory-based feature proposed by Wang.H et al. [16]. Related parts, such as: dense sampling, tracking and feature descriptors are also referred to. Second, we present how our proposed method can provide much discriminative motion information from depth video. Finally, our general framework on depth video is mentioned at the end of this section.

3.1. Dense trajectories

Trajectories provide a compact representation of motion information in video. Trajectories from intensity videos can be used for multimedia event detection (MED), video mining, action classification and so on. Trajectory extraction much depends on both processes: sampling and tracking. Some concerned methods, such as [13, 14] used KLT tracker [34], or [15] matched SIFT descriptors between consecutive frames to obtain feature trajectories. Recently, the dense trajectory-based motion feature proposed by [16] has achieved the state-of-the-art performances on MED systems, such as, segment-based system [17] on TRECVID MED 2010, 2011, or AXES [18], and BBNVISER [19] on TRECVID MED 2012.

In order to obtain trajectories, there are two important steps: sampling and tracking. [16] propose sampling on a dense grid with a step size of 5 pixels. The sampling is performed at multiple scales with a factor of $1/\sqrt{2}$. Then, tracking is the next step to form trajectories. At each scale, in frame t, each point $P_t = (x_t, y_t)$ is tracked to point $P_{t+1} = (x_{t+1}, y_{t+1})$ in next frame t+1 by:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)}, \qquad (1)$$

where $\omega = (u_t, v_t)$ denotes the dense optical flow field, M is the kernel of median filtering, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of P_t. The algorithm of [35] is adopted to compute the dense optical flow. And to avoid a drifting problem, a suitable value of trajectory length is set to 15 frames. Besides, trajectories with sudden changes are removed.

After extracting trajectories, two kinds of descriptors: a trajectory shape descriptor and a trajectory-aligned descriptor can be adopted. In our experiments, we only use

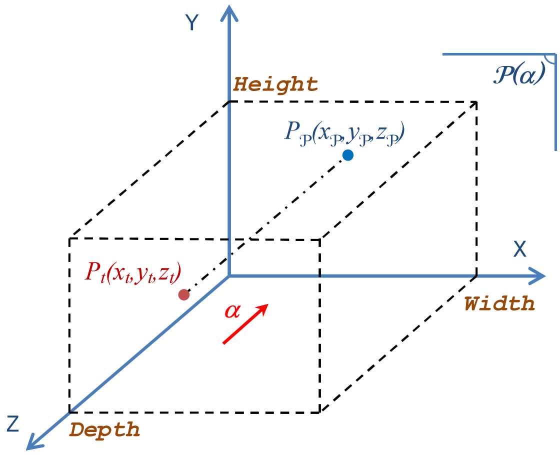


Figure 2: An illustration of the projection. Point $P_{\mathcal{P}}$ is the projection of point P_t along a view direction α onto a view plane $\mathcal{P}(\alpha)$.

trajectory-aligned descriptors including the HOG [30], the Histogram of Optical Flow (HOF) [9], and the Motion Boundary Histogram (MBH) [36]. HOG captures local appearance information, while HOF and MBH encode local motion pattern. The descriptors are computed within a space-time volume ($N \times N$ spatial pixels and L temporal frames) around the trajectory. This volume is divided into a 3D grid (spatially $n_\sigma \times n_\sigma$ grid and temporally n_τ segments). The default settings of these parameters are $N = 32$ pixels, $L = 15$ frames, $n_\sigma = 2$, and $n_\tau = 3$.

According to the authors [9, 16, 37, 38], all the three descriptors have shown the effectiveness for action recognition. The experimental settings for these descriptors are based on an empirical study showed in [16]. We also conduct our experiment on all the three descriptors when compared to the depth-based state-of-the-art methods.

3.2. Proposed Method for Dense Trajectory-based Approach

Our proposed method to adapt the dense trajectory-based approach for human action recognition on depth video is as follow. At first, intensity videos are formed from the sequence of depth maps, as illustrated in figure 1. At this step, to obtain an intensity video from a view direction α, corresponding to a view plane $\mathcal{P}(\alpha)$: $ax + by + cx + d = 0$, in each depth map t, each point $P_t(x_t, y_t, z_t)$ is projected to $P_{\mathcal{P}}(x_{\mathcal{P}}, y_{\mathcal{P}}, z_{\mathcal{P}})$ on the view plane $\mathcal{P}(\alpha)$, see in figure 2, by:

$$P_t(x_t, y_t, z_t) \xrightarrow{\mathcal{P}(\alpha)} P_{\mathcal{P}}(x_{\mathcal{P}}, y_{\mathcal{P}}, z_{\mathcal{P}}) \tag{2}$$

where,

$$x_{\mathcal{P}} = x_t - \frac{ax_t + by_t + cz_t + d}{a^2 + b^2 + c^2} a \tag{3}$$

$$y_{\mathcal{P}} = y_t - \frac{ax_t + by_t + cz_t + d}{a^2 + b^2 + c^2} b \tag{4}$$

$$z_{\mathcal{P}} = z_t - \frac{ax_t + by_t + cz_t + d}{a^2 + b^2 + c^2} c \tag{5}$$

And the intensity value v at the projected point $P_{\mathcal{P}}$ is computed by:

$$v(P_{\mathcal{P}}) = \frac{ax_t + by_t + cz_t + d}{\sqrt{a^2 + b^2 + c^2}} \tag{6}$$

So, given a set of 3D points $\mathcal{S}(t) = \{(x_t, y_t, z_t)|(x_t, y_t, z_t) \in t\}$, we have a projection $\mathcal{S}_\alpha(t) = \{(x_{\mathcal{P}}, y_{\mathcal{P}}, z_{\mathcal{P}})|(x_{\mathcal{P}}, y_{\mathcal{P}}, z_{\mathcal{P}}) \in \mathcal{P}(\alpha)\}$. Therefore, a set of the projections obtained from a given sequence of M depth maps under a view direction α is an expected intensity video $\mathcal{R}(\alpha) = \{\mathcal{S}_\alpha(t)|t = \overline{1..M}\}$. Each intensity video obtained from the corresponding projection onto the sequence of depth maps can be regarded as a 2D transformed video of action in depth video.

In particular, we choose three representations to represents for three view directions: front, side, and top in 3D space, corresponding to three view planes, respectively: Oxy, Oyz and Ozx. With these view directions, the corresponding projections are respectively:

$$\mathcal{S}_{\text{front}}(t) = \{(x_t, y_t, 0)|(x_t, y_t, 0) \in \mathcal{P} : z = 0\} \tag{7}$$

$$\mathcal{S}_{\text{side}}(t) = \{(0, y_t, z_t)|(0, y_t, z_t) \in \mathcal{P} : x = 0\} \tag{8}$$

$$\mathcal{S}_{\text{top}}(t) = \{(x_t, 0, z_t)|(x_t, 0, z_t) \in \mathcal{P} : y = 0\} \tag{9}$$

And the corresponding intensity values in the three projections are, respectively:

$$v(P_{\text{front}}) = z_t \tag{10}$$

$$v(P_{\text{side}}) = x_t \tag{11}$$

$$v(P_{\text{top}}) = y_t \tag{12}$$

3.3. Our framework overview

In this section, we provide a brief introduction about our framework for action recognition task. The first step is to transform projection results from sequences of depth maps into corresponding 2D videos. Transforming depth video into the 2D videos is necessary due to dimensional gap when we adapt 2D techniques for 3D data. Afterwards, the dense trajectories [16] are extracted from the 2D transformed videos. With this approach, we do not care the challenges from human body segmentation as well as skeleton extraction. Trajectory-aligned descriptors are computed then. At the next step, with each 2D transformed video $\mathcal{R}(\alpha_i)$, $i = \overline{1..N}$, corresponding feature representation $F(\alpha_i) = (b^1_{\alpha_i}, b^2_{\alpha_i}, ..., b^K_{\alpha_i})$ is quantized from a set of raw trajectory features using a bag-of-words (BoW) model with K visual words. For quantization, the hard-assignment technique is used to compute histograms of the visual words on the 2D transformed videos. An *early fusion* scheme which integrates unimodal features before learning, then, is used to generate feature representation $\mathcal{F} = (F(\alpha_1), ..., F(\alpha_N))$ for action in the sequence of depth maps. After the final feature representations are generated, we adopt the popular Support Vector Machine (SVM) for classification. In practice, we use the precomputed-kernel technique with the histogram intersection kernel for the classification step. Besides, we perform the one-vs-all strategy for multi-class classification.

Our proposed trajectory-based approach is compared with the state-of-the-art methods in human action recognition using depth data. Actually, our approach does not care skeleton extraction, which is used as an important factor in some works, such as [23, 25]. In fact, extracting skeleton exactly is still an completely unsolved problem, due to the challenges, such as cluttered background, hardware quality, camera motion, so on.

4. Experimental Settings

In this section, we provide a brief description about our experimental framework in step by step. Figure 3 shows the framework for the dense trajectory-based approach, including the following steps:

Step 1. Each depth sequence is projected onto three planes corresponding to three views: front, side and

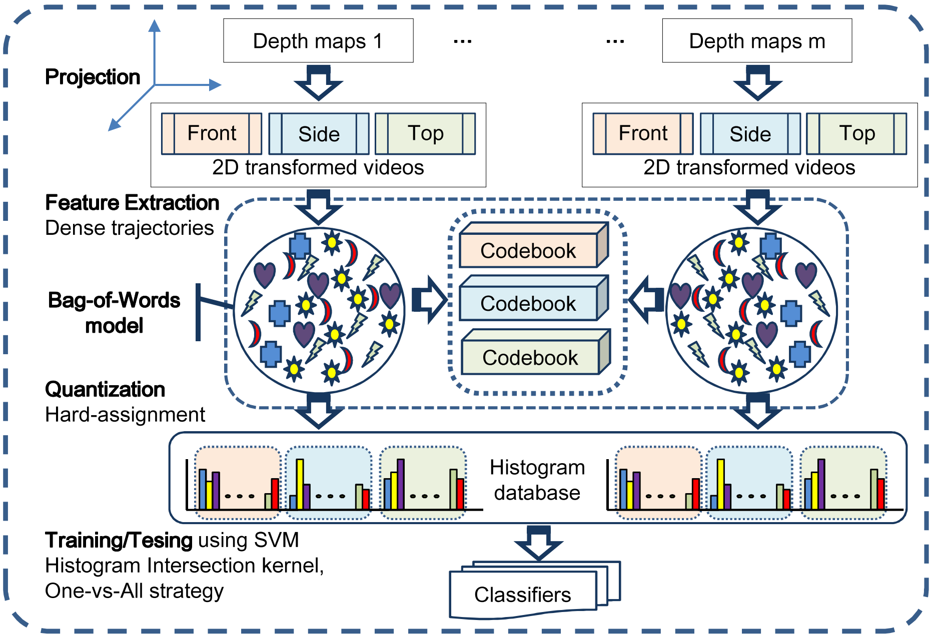


Figure 3: Our Framework Overview

top. Results obtained from the projections are three 2D transformed videos.

Step 2. We use the application available online[1] to extract dense trajectories and calculate aligned-descriptors (i.e. MBH, HOG and HOF) for each 2D transformed video. Experimental results reported in section 5 attach to the MBH descriptor. The HOG, HOF descriptors will be mentioned in the section 6.

Step 3. To represent 2D transformed videos, the BoW model is applied. At this step, we create three codebooks with 2000 visual codewords for each using K-mean algorithm to cluster.

Step 4. In order to quantize a large number of dense trajectory motion features extracted at *Step 2*, we apply the hard-assignment strategy. Feature representations for 2D transformed videos quantized at this step will be fused under the early scheme to form feature representations of corresponding actions. The feature representations for actions are, then, separated into two histogram databases for training and testing.

Step 5. In order to classify actions, in our implementation, we use the libSVM library published online by author[2]. We adopt the format requirements of the library to synchronize the annotation and the data. We apply histogram intersection kernel to compute matching matrices before we do training and testing with SVM. For testing, the one-vs-all strategy is used. Predicted value of each action is defined as the maximum score obtained from all the classifiers. This score shows that a human action is confused with another or not.

Action Subset 1 (AS1)	**Action Subset 2 (AS2)**	**Action Subset 3 (AS3)**
horizontal arm wave	high arm wave	high throw
hammer	hand catch	forward kick
forward punch	draw x	side kick
high throw	draw tick	jogging
hand clap	draw circle	tennis swing
bend	two hand wave	tennis serve
tennis serve	side-boxing	golf swing
pick up & throw	forward kick	pick up & throw

Table 1: The three action subsets used in the experiments

[1]http://lear.inrialpes.fr/~wang/dense_trajectories

[2]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

5. Experiments

This section presents the experimental results for applying our proposed approach on MSR Action 3D dataset, MSR Gesture 3D dataset, and 3D Action Pairs dataset. All experimental results are reported under the settings mentioned in section 4. In comparison with the state-of-the-art methods, our reported result is calculated on fusing feature representations from the combinations of three views: front, side and top. In addition, an evaluation related to selecting compensation information from the three views will be also mentioned. All the results are compared in terms of recognition accuracy. The best performance is highlighted in bold.

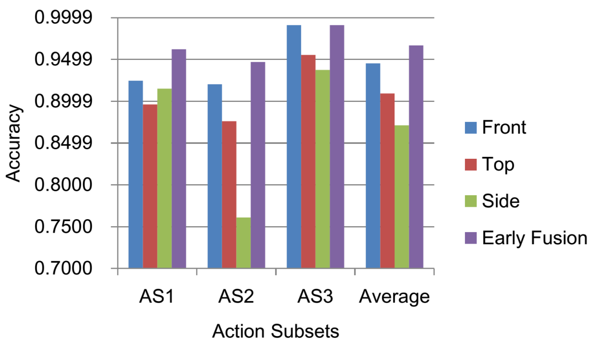


Figure 4: Comparison of recognition accuracy by using the early fusion scheme from different views.

	a02	a03	a05	a06	a10	a13	a18	a20
a02	0.83	0	0.17	0	0	0	0	0
a03	0	0.92	0.08	0	0	0	0	0
a05	0	0.36	0.64	0	0	0	0	0
a06	0	0	0	1.0	0	0	0	0
a10	0	0	0	0	1.0	0	0	0
a13	0	0	0	0	0	1.0	0	0
a18	0	0	0	0	0	0	1.0	0
a20	0	0	0	0	0	0.07	0	0.93

(a) Action Subset 1

	a01	a04	a07	a08	a09	a11	a12	a14
a01	1.0	0	0	0	0	0	0	0
a04	0.08	0.84	0.08	0	0	0	0	0
a07	0	0	0.79	0.07	0.07	0	0.07	0
a08	0	0	0	1.0	0	0	0	0
a09	0	0	0	0.13	0.87	0	0	0
a11	0	0	0	0	0	1.0	0	0
a12	0	0.13	0	0	0	0	0.87	0
a14	0	0	0	0	0	0	0	1.0

(b) Action Subset 2

	a06	a14	a15	a16	a17	a18	a19	a20
a06	1.0	0	0	0	0	0	0	0
a14	0	1.0	0	0	0	0	0	0
a15	0	0	1.0	0	0	0	0	0
a16	0	0	0	1.0	0	0	0	0
a17	0	0	0	0	1.0	0	0	0
a18	0	0	0	0	0	0.93	0.07	0
a19	0	0	0	0	0	0	1.0	0
a20	0	0	0	0	0	0	0	1.0

(c) Action Subset 3

Table 2: The confusion tables for three action subsets from MSR Action 3D dataset. Notice that action names are identified by indices of actions in figure 5.

5.1. *MSR Action 3D Dataset*

5.1.1. *A Brief Introduction*

This dataset [22] contains 20 actions, as showed in figure 5. Actions are performed by ten subjects for two or three times in the context of game console interaction. In total, there are 567 sequences of depth maps. The depth maps are shot at frame rate of 15 fps. The size of the depth map is 640×480, we resize into 320×240 to ensure processing efficiency.

In order to conduct a fair comparison, we use the same experimental settings as [22, 23, 25, 26, 28, 29]. In the settings, the dataset is divided into three action subsets. Each subset has 8 actions (Table 1). The two subsets AS1 and AS2 present that grouped actions have similar movements. The subset AS3 groups complex actions together. For instance, action *hammer* seems to be confused with action *forward punch* in AS1 or similar movements between action *hand catch* and action *side boxing* in AS2. As for each subset, we select half of the subjects as training and the rest as testing (i.e. cross subject test).

5.1.2. *Recognize Actions from Single-View*

In this part, we evaluate the dense trajectory-based approach for action recognition under observing actions from single-view. A straightforward view is front view. In order to obtain action presentation on front view from depth video, a simple way is to consider depth value as intensity value. Table 2 shows three confusion matrices corresponding to evaluations on three action subsets from MSR Action 3D dataset. Consider results reported in table 2, we found that two subsets AS1, AS2 contain many confused actions. For example, *hammer* (a03) and

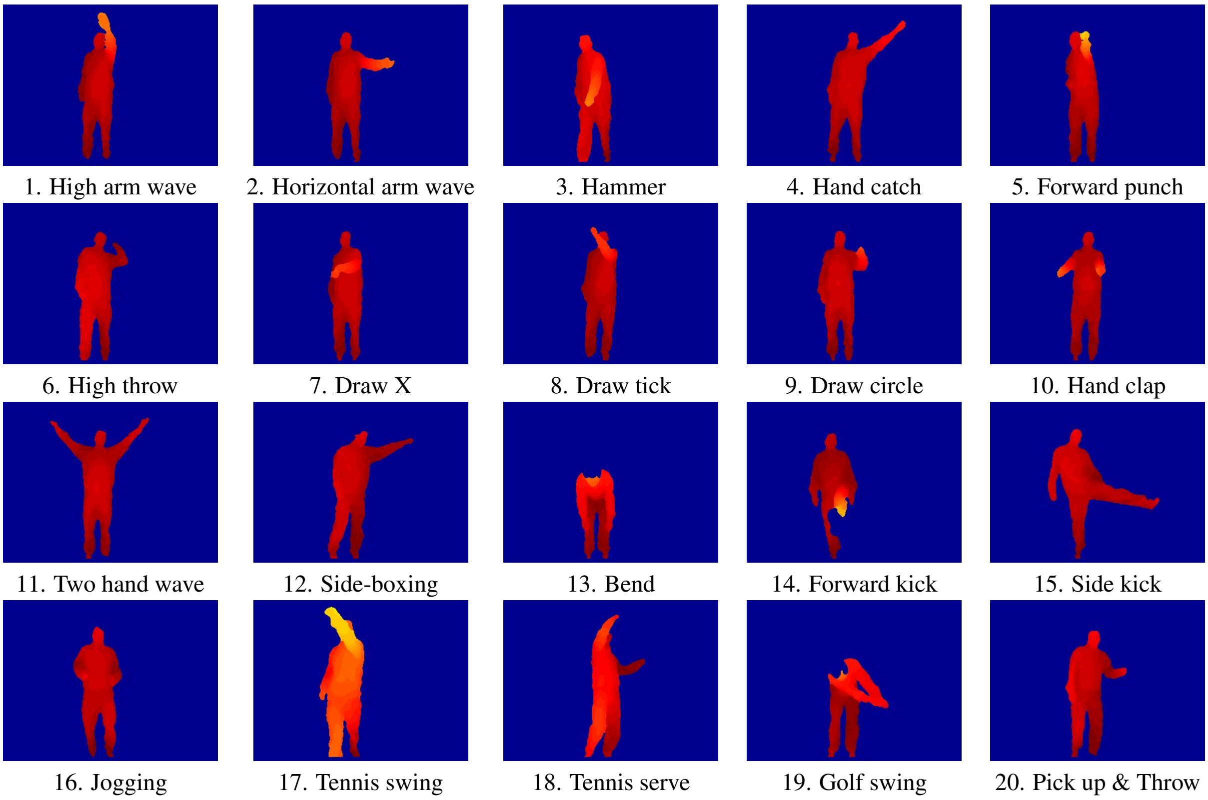


Figure 5: Example frames for twenty actions from MSR Action 3D dataset [22].

forward punch (a05) in AS1, or *side-boxing* (a12) and *hand catch* (a04) in AS2. When analyzing such actions, we found that the main cause is due to similar movements of actions in the same view direction. That is reason why we need compensate motion information from other views (e.g. side view and top view).

5.1.3. Compensate Motion Information from Other Representations

To compensate more discriminative motion information, we conducted experiments based on compensating information from other views, such as side and top, for the action representation from front view. We report the experimental results on three action subsets and the average of the three subsets. Figure 4 shows a comparison between the 2D representations from front, side and top and their fusion representation. Expectedly, the average recognition accuracy of the fusion, which is 96.67% accuracy, is better than the average recognition accuracy of the individual representations on the three action subsets. Obviously, our proposed approach shows the effectiveness of leveraging depth information to capture much more discriminative motion information.

5.1.4. The Role of Views

Figure 4 shows the role of views to our approach. Experimental results confirm that action representations from front achieve the best performances. Obviously, the front view is an indispensable component to merge information. For the rest, we perform experiments on view combinations with front view.

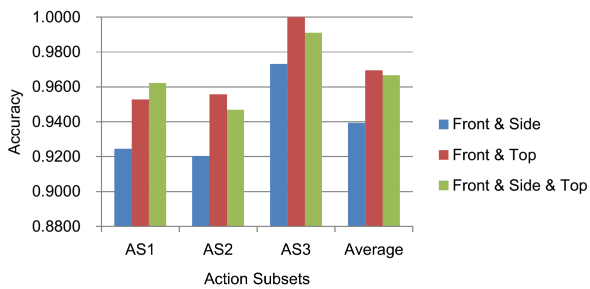


Figure 6: Comparison of recognition accuracy on combinations of intensity representations on MSR Action 3D dataset.

In order to conduct the experiments, we create additional combinations: front and side, front and top. Figure 6 shows the performance of the view combinations. Interestingly, the achieved performance (96.95%) from the combination of front and top beats the performance based on combining all the three views (96.67%) as well as the combination of front and side (93.94%), in terms of average accuracy. In addition, based on experimental results, as described in figure 6, compensating information indicates two interesting points. Firstly, compensating information from various views can cause unexpected risks, due to erroneous information from certain views. In this case, merging information from side view into the combination of front and top views causes to decrease the performance of the recognition system. Secondly, the experimental results have provided a good choice to decrease computational cost but still ensures a convincing performance. This issue can lead to looking for optimal solution of combining views. This is a promising challenge to overcome and build an effective and efficient recognition system.

Method	**Accuracy (%)**
W.Li et al. [22]	74.70
A.W.Vieira et al. [24]	84.80
X.Yang et al. [25]	82.33
J.Wang et al. [27]	86.50
J.Wang et al. [23]	88.20
X.Yang et al. [26]	91.63
O.Oreifej & Z.Liu [29]	88.89
L.Xia & J.Aggarwal [28]	89.30
J.Luo et al. [33]	96.70
Ours (FRONT + SIDE)	93.94
Ours (FRONT + TOP)	**96.95**
Ours (FRONT + SIDE + TOP)	96.67

Table 3: The performance of our approach on MSR Action 3D dataset. Notice that experimental results reported in this table is based on combinations of three views: front, side and top. Besides, we also use MBH descriptor only to calculate trajectory features.

5.1.5. Comparison with the state-of-the-art

Table 3 shows evaluation results of our proposed approach and the state-of-the-art approaches in terms of average accuracy on three action subsets from MSR Action 3D dataset (seeing table 1). The compared approaches are based on various feature representations, such as silhouette features [22, 26], skeletal joint features like [23, 25], local occupancy patterns [24, 27], normal orientation features [29] and cuboid similarity features [28]. Under the same setting (i.e cross subject test), the result table indicates that our approach beats all of them.

Method	**Accuracy (%)**
X.Yang et al. [26]	89.2
J.Wang et al. [27]	88.5
O.Oreifej & Z.Liu [29]	92.45
Ours (FRONT+SIDE)	93.22
Ours (FRONT+TOP)	92.66
Ours (FRONT+SIDE+TOP)	**94.35**

Table 4: The performance of our approach on MSR Gesture 3D dataset, compared to previous approaches

5.2. MSR Gesture 3D Dataset

The Gesture3D dataset [27] is a hand gesture dataset of depth sequences captured by a depth camera. This dataset contains a set of gesture defined by American Sign Language (ASL). In the dataset, there are 12 gestures as described in figure 7. There are ten subjects, each performs each gesture two or three times. In total, the dataset contains 333 depth sequences. The main challenge in the dataset is self-occlusion issues. We follow the experimental settings in [27] to do evaluation on our approach. We obtain the accuracies described in table 4, where our approach outperforms all previous approaches.

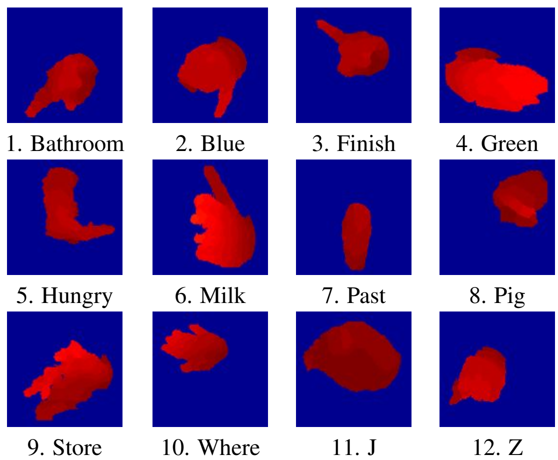
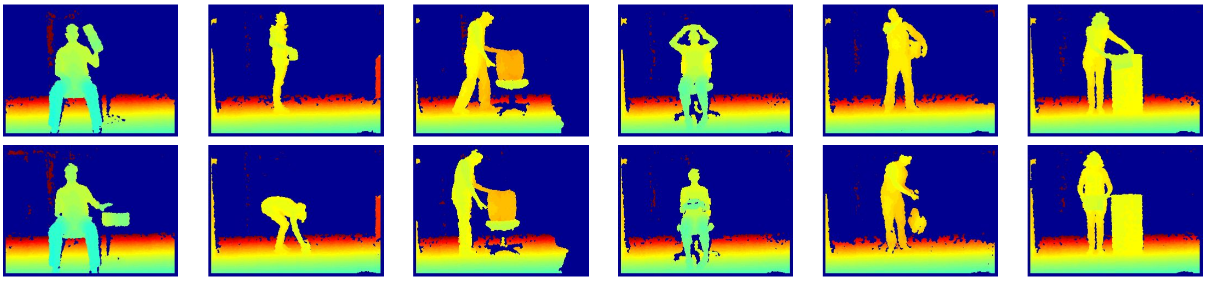


Figure 7: Example frames for hand gestures from MSR Gesture 3D dataset [27].



Pick up/Put down a box | Lift/Place a box | Push/Pull a chair | Wear/Take off a hat | Put on/Take off a backpack | Stick/Remove a poster

Figure 8: Example frames for six pairs from 3D Action Pairs dataset [29]. Each column shows two frames from a pair of actions.

5.3. 3D Action Pairs Dataset

The 3D Action Pairs dataset [29] is a new type of action dataset. The dataset contains pairs of actions, such that within each pair the motion and the shape cues are similar, but their correlations vary. It is useful to evaluate how well the approaches capture the prominent cues jointly in depth sequences. There are six pairs of actions, see figure 8. Each action is performed three times by ten subjects, where the first five subjects are used for testing, and the rest for training.

Method	Accuracy (%)
O.Oreifej & Z.Liu [29]	96.67
Ours (FRONT+SIDE)	92.22
Ours (FRONT+TOP)	**99.44**
Ours (FRONT+SIDE+TOP)	92.78

Table 5: The performance of our approach on 3D Action Pairs dataset, compared to the state-of-the-art approach.

	a01	a02	a03	a04	a05	a06	a07	a08	a09	a10	a11	a12
a01	1.0	0	0	0	0	0	0	0	0	0	0	0
a02	0	1.0	0	0	0	0	0	0	0	0	0	0
a03	0	0	0.80	0	0	0	0	0	0.133	0.067	0	0
a04	0	0	0	0.933	0	0	0	0	0	0.067	0	0
a05	0	0	0	0	1.0	0	0	0	0	0	0	0
a06	0	0	0	0	0	1.0	0	0	0	0	0	0
a07	0	0	0	0	0	0	1.0	0	0	0	0	0
a08	0	0	0	0	0	0	0	1.0	0	0	0	0
a09	0	0	0	0	0	0	0	0	1.0	0	0	0
a10	0	0	0	0	0	0	0	0	0	1.0	0	0
a11	0	0	0	0	0	0	0	0	0	0	1.0	0
a12	0	0	0	0	0	0	0	0	0	0	0.133	0.867

(a) HON4D

	a01	a02	a03	a04	a05	a06	a07	a08	a09	a10	a11	a12
a01	0.933	0	0	0	0	0	0	0.067	0	0	0	0
a02	0	1.0	0	0	0	0	0	0	0	0	0	0
a03	0	0	1.0	0	0	0	0	0	0	0	0	0
a04	0	0	0	1.0	0	0	0	0	0	0	0	0
a05	0	0	0	0	1.0	0	0	0	0	0	0	0
a06	0	0	0	0	0	1.0	0	0	0	0	0	0
a07	0	0	0	0	0	0	1.0	0	0	0	0	0
a08	0	0	0	0	0	0	0	1.0	0	0	0	0
a09	0	0	0	0	0	0	0	0	1.0	0	0	0
a10	0	0	0	0	0	0	0	0	0	1.0	0	0
a11	0	0	0	0	0	0	0	0	0	0	1.0	0
a12	0	0	0	0	0	0	0	0	0	0	0	1.0

(b) Ours (FRONT+TOP)

Table 6: The confusion tables for 3D Action Pairs dataset.

We compare our performance in this dataset with the HON4D approach [29], which is the state-of-the-art performance until current time. We summarize results in table 5, and demonstrate the confusion tables in table 6. It is clear that our approach significantly outperforms the state-of-the-art approach for suffering from confusion appeared within action pairs.

6. Discussions

6.1. The Impact of Our Method on Descriptors

For intensity data, according to [16] MBH is the best feature descriptor for dense trajectories. Therefore, in previous experiments, we only use MBH descriptor to represent motion information. Due to the difference between depth data and intensity data, how our approach has influenced other trajectory-aligned descriptors (i.e. HOG, HOF). In this part, we conduct similar experiments on these descriptors to answer this issue.

We report the average recognition accuracies on the three descriptors and on the combinations of the three views: front, side, and top. Table 7 shows interesting results. Firstly, experimental results verify that the MBH descriptor is still the best trajectory-aligned descriptor in comparison with the HOG, HOF descriptors on the experimental datasets. Secondly, although the HOG, HOF descriptors are not the best, their performance is comparable to the state-of-the-art approaches, as men-

	MSR Action 3D			MSR Gesture 3D			3D Action Pairs		
Combination	**MBH**	**HOG**	**HOF**	**MBH**	**HOG**	**HOF**	**MBH**	**HOG**	**HOF**
FRONT+SIDE	93.94	93.01	91.82	93.22	**92.09**	89.83	92.22	82.78	92.22
FRONT+TOP	**96.95**	92.14	**92.70**	92.66	90.40	88.14	**99.44**	**90.00**	**93.89**
FRONT+SIDE+TOP	96.67	**94.53**	92.42	**94.35**	91.53	**92.09**	92.78	88.89	91.67

Table 7: The performance of descriptors (MBH, HOG, and HOF) on MSR Action 3D dataset, MSR Gesture 3D datset, and 3D Action Pairs dataset.

tioned in section 5. In addition, lower-cost descriptors like HOG, HOF have more benefits for decreasing computational cost in processes, such as feature extraction and video representation (using the BoW model). These advantages provide a promising way for building effective and efficient systems.

6.2. MSR Daily Activity 3D Dataset

The MSR Daily Activity 3D dataset is proposed by [23], which includes 16 daily activities, as described in figure 9. In this dataset, background objects and subjects appear at different distances to the camera. In order to evaluate our approach to this dataset, we follow the experimental settings mentioned in [23]. In this experiment, we conduct the dense trajectory-based approach on view combinations only use MBH descriptor to describe motion feature. In comparison with the state-of-the-art approach, we evaluate recognition accuracy on depth data only. Constraint means that skeleton information-based approaches are not considered. Since, in real environment, skeleton extraction must suffer from complex backgrounds, skeleton information captured from depth camera is often unstable. Table 8 shows a comparison of recognition accuracy with the state-of-the-art approaches on MSR Daily Activity 3D dataset. In [28], since they modified this dataset to do evaluation, it is not fair to compare. Therefore, to ensure the fair comparison, we follow a framework similar to [28] and evaluate on the original MSR Daily Activity 3D dataset. In condition of only using depth data, [23, 28, 29] report a unexpected performance.

Although our approach outperforms the previous approaches, it is clear that the effectiveness is not well. It is important to note that why in condition of only using depth data, most of methods are failed. When considering failed samples, such as *playing a game*, *writing on a paper*, and *using a laptop*, we found that most of them are confused with action *still*. For *playing a game*, main motion focuses on motion of fingers, it is very difficult to discriminate from depth noise. For *writing on a paper* and *using a laptop*, hand gestures are major motions to present motion information. But it is not fortunately, most of the movements are hidden by interactive objects (i.e. book, laptop). That is one reason to explain for the failure. The second one is performing similar movements with different objects, such as *talking on the phone* and *drinking water*. In these cases, objects are small and textureless, so, it is not easy to identify them. Therefore, if only depending on depth data, it is very challenging to recognize these actions exactly. Due to these reasons, in order to improve the performance of recognition systems in terms of interaction, adding more information related to motion, such as interactive objects, is necessary.

Method	**Accuracy**
LOP [23]	42.5
HON4D [29]	52
DSTIP&DCSF [28]	56.88
Ours (FRONT+SIDE)	**65.63**
Ours (FRONT+TOP)	59.38
Ours (FRONT+SIDE+TOP)	**65.63**

Table 8: The performance of our approach on MSR Daily Activity 3D Dataset. Notice that results are reported in terms of only using depth data.

7. Conclusions

We proposed the Pseudo-3D Trajectories, a 2D trajectory-based approach, for human action recognition using depth data in this work. We evaluated our approach by using the dense trajectory motion feature on the challenging datasets. More interestingly, our proposed trajectory-based approach only applied for one representation beats all the recent state-of-the-art approaches in terms of depth data. Besides, in order to deal with confused actions due to similar movements, compensating information from other representations is proposed. Therefore, the effectiveness of our approach

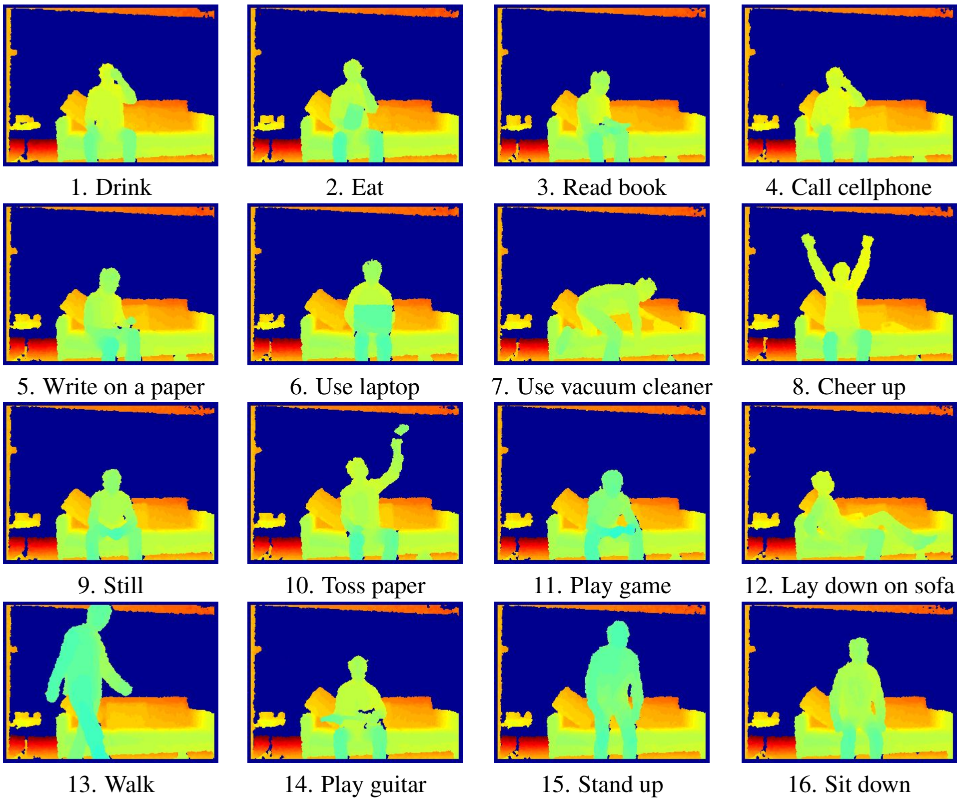


Figure 9: Example frames for sixteen activities from MSR Daily Activity 3D dataset [23].

on depth datasets like MSR Action 3D, MSR Gesture 3D and 3D Action Pairs is confirmed.

A trajectory-based approach with compensating information from separate representations shows promising results. This opens a general approach to leverage intensity-based techniques for depth data. This also suggests the importance of trajectory-based motion information on human action recognition using depth data. Therefore, exploiting depth-based motion trajectories can be beneficial for action recognition systems using depth cameras. This is also an interesting idea for our future work.

References

[1] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2847–2854.

[2] R. Poppe, A survey on vision-based human action recognition, Image and vision computing 28 (6) (2010) 976–990.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, Vol. 2, IEEE, 2005, pp. 1395–1402.

[4] Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8.

[5] S. N. Vitaladevuni, V. Kellokumpu, L. S. Davis, Action recognition using ballistic dynamics, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[6] A. Yilmaz, M. Shah, Actions sketch: A novel action representation, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 984–989.

[7] I. Laptev, On space-time interest points, International Journal of Computer Vision 64 (2-3) (2005) 107–123.

[8] P. Dollár, V. Rabaud, G. Cottrell, 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.

[9] I. Laptev, M. Marszalek, C. Schmid, 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.

[10] M. Bregonzio, S. Gong, T. Xiang, Recognising action as clouds of space-time interest points, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1948–1955.

[11] A. Klser, M. MarszaÅek, C. Schmid, A spatio-temporal descrip-

tor based on 3d-gradients, in: Proceedings of the British Machine Vision Conference (BMVC08), Leeds, United Kingdom, September 2008, 2008, pp. 995–1004.

[12] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 650–663.

[13] P. Matikainen, M. Hebert, R. Sukthankar, Trajectons: Action recognition through the motion analysis of tracked features, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 514–521.

[14] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 104–111.

[15] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, 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.

[16] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action Recognition by Dense Trajectories, in: IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, United States, 2011, pp. 3169–3176.
URL `http://hal.inria.fr/inria-00583818/en`

[17] S. Phan, T. D. Ngo, V. Lam, S. Tran, D.-D. Le, D. A. Duong, S. Satoh, Multimedia event detection using segment-based approach for motion feature, Journal of Signal Processing Systems 74 (1) (2014) 19–31.

[18] D. Oneata, M. Douze, J. Revaud, S. Jochen, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, et al., Axes at trecvid 2012: Kis, ins, and med, in: TRECVID workshop, 2012.

[19] P. Natarajan, P. Natarajan, S. Wu, X. Zhuang, A. Vazquez-reina, S. N. Vitaladevuni, C. Andersen, R. Prasad, G. Ye, D. Liu, et al., Bbn viser trecvid 2012 multimedia event detection and multimedia event recounting systems.

[20] M. Jain, H. Jégou, P. Bouthemy, Better exploiting motion for better action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 2555–2562.

[21] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 3551–3558.

[22] W. Li, Z. Zhang, 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.

[23] J. Wang, Z. Liu, Y. Wu, 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.

[24] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, 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] X. Yang, 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.

[26] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the 20th ACM international conference on Multimedia, ACM, 2012, pp. 1057–1060.

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

[28] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 2834–2841.

[29] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 716–723.

[30] N. Dalal, 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.

[31] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Communications of the ACM 56 (1) (2013) 116–124.

[32] L. Xia, C.-C. Chen, J. Aggarwal, Human detection using depth information by kinect, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, IEEE, 2011, pp. 15–22.

[33] J. Luo, W. Wang, H. Qi, Group sparsity and geometry constrained dictionary learning for action recognition from depth maps, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 1809–1816.

[34] B. D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision., in: IJCAI, Vol. 81, 1981, pp. 674–679.

[35] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Image Analysis, Springer, 2003, pp. 363–370.

[36] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 428–441.

[37] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al., Evaluation of local spatio-temporal features for action recognition, in: BMVC 2009-British Machine Vision Conference, 2009.

[38] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1996–2003.

[39] C. G. Snoek, M. Worring, A. W. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, 2005, pp. 399–402.