Pseudo-3D Trajectories: An Effective Approach for Motion Representation in Depth Data

Chien-Quang LE

The Graduate University for Advanced Studies

Duy-Dinh LE

National Institute of Informatics, The Graduate University for Advanced Studies

Shin'ichi SATOH

National Institute of Informatics, The University of Tokyo

Abstract

Leveraging the motion information of trajectories shows the effectiveness to the human action recognition in intensity videos. However, the issue is that this approach direction is effective or not when represents motions in depth video is not still answered. In this paper, we will deal with this issue by conducting experiments based on intensity trajectory features to present motion information from one depth video representation. Beside, in order to ensure including depth information, we propose a method based on compensating motion information from other representations. Evaluated on the benchmark datasets, our method significantly outperforms the state-of-the-art depth-based methods.

Keywords: `Trajectory, action recognition, depth data, feature representation`

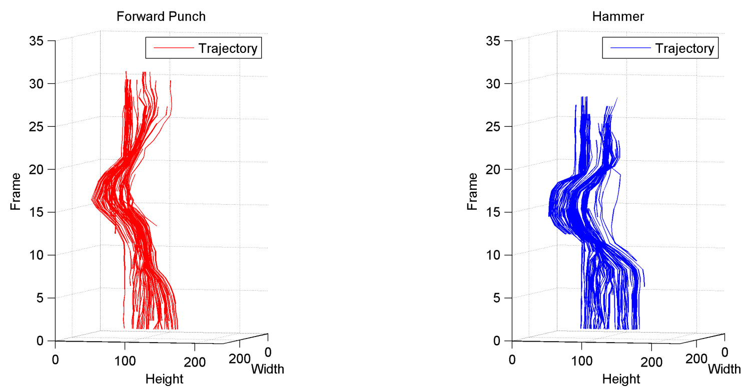
---

1. Introduction

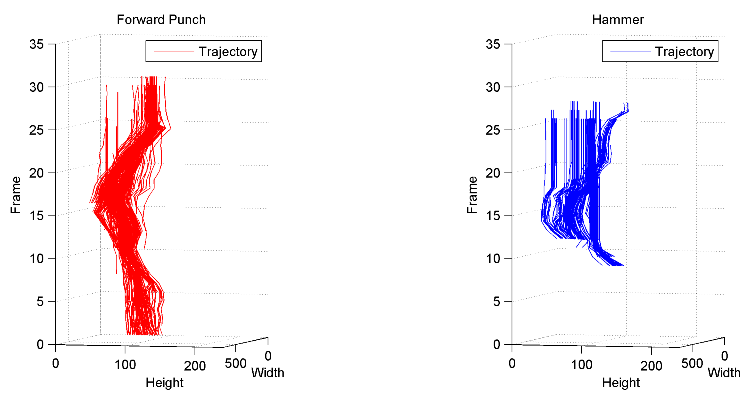
Recently, with the development of RGB-D cameras such as Kinect, depth data pioneers many potential research directions for human action recognition. Compared with conventional intensity images, depth maps support more several advantages. For example, depth maps provide shape information, which can be clearer than intensity images. Moreover, the depth data is less affected by illumination variations. However, an issue is that intensity-based methods are effective or not on depth data, which has not been much interested.

For action recognition, in order to effectively adapt intensity-based methods for depth data, we need satisfy two major factors. Firstly, a robust feature representation is extremely important to exactly capture motion information. Secondly, to ensure that a motion contains full information in depth video, merging depth information into feature representation is an indispensable requirement. However, the recent proposed methods do not combine two the factors completely. Works [1, 2] consider depth value as intensity value and adapt the intensity-based techniques. Although, they can achieve reasonable results, they deal with many limitations. [1] can leverage depth information from the projections of depth maps. But its feature representation based on global motion such as HOG easily cause confusion by similar postures. [2] can ensure depth information in feature descriptor computation. But this approach does not guarantee the reliability when extracting local points, due by textureless data and depth noise. Beside, methods in [3, 4] only focus on exploiting depth information without leveraging the effectiveness of the intensity-based features. Therefore, we propose an effective method that can satisfy both mentioned factors.

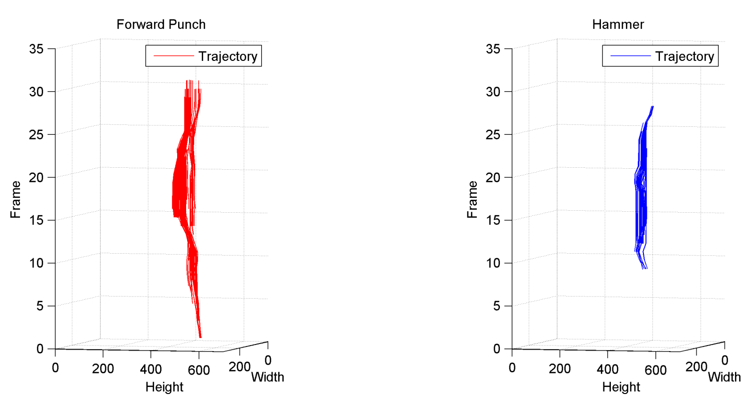
In this paper, we use a feature representation based on dense trajectories proposed by [5], due to the effectiveness of this approach in many problems, including activity recognition and multimedia event detection (MED). The trajectories obtained by tracking densely sampled points using optical flow fields.



(a) From front view



(b) From side view



(c) From top view

Figure 1: An illustrative comparison between trajectories' shape of actions *Forward Punch* and *Hammer*.

After extracting the trajectories, trajectory-aligned descriptors will be adopted. Then, features computed from the descriptors will be used to represent motion information in video.

However, the lack of depth information in feature representation can cause several confused cases, as shown in figure 1(a). Thus, to ensure that depth information is not ignored, a basic idea is to combine the motion information from various views. The view-based representations can be achieved by projecting depth maps onto the corresponding planes. The projections are easily obtained by the mentioned advantages of depth data.

We conduct experiments on MSR Action 3D dataset and MSR Daily Activity 3D 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 adaptive method for depth video representation by using intensity-based features. (2) We perform comprehensive experiments on the challenging benchmark datasets and indicate that our 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

2.1. Trajectories Extraction

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 [6, 7] used KLT tracker [8], or [9] matched SIFT descriptors between consecutive frames to obtain feature trajectories. Recently, the dense trajectories feature proposed by [5] has achieved state-of-the-art performances on MED systems, such as, segment-based system [10] on TRECVID MED 2010, 2011, or AXES [11], and BBNVISER [12] on TRECVID MED 2012.

Although, depth data has been studied ago several decades, the trajectory extraction in depth videos is not still paid attention to. This is obviously a significant deficiency for motion-based recognition systems using depth data.

2.2. Feature Representation from Depth Videos

In terms of human action recognition in depth video, most recent methods exploit depth information into two major directions. The first one is adapting intensity 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. [1] propose the Depth Motion Maps (DMM) to accumulate global activities in depth video sequences. And the Histogram of Oriented Gradients (HOG) are computed from the DMM to represent an action video. Another approach bases on spatio-temporal interest points proposed by Xia.L and Aggarwal.J.K [2]. In this approach, they extend a work of Dollar et al. [13] to adapt for depth data.

For the second direction, [14] uses 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. [15, 16, 3] use occupancy patterns to represent feature in action video. Another approach proposed by Oeifej et al. [4] leverages the distribution of surface normal orientation in the 4D space of time, depth and spatial coordinates to build a feature histogram. Inspired by results of Shotton et al. [17] and Xia.L et al. [18], works [19, 3] propose new types of features based on skeleton information.

Different from other approaches, we use a trajectory-based approach for action recognition. We do not care to segment human body like [14, 1]. We only investigate the benefit of generating intensity representations from depth data, as mentioned in [14, 1]. Moreover, we leverage the effectiveness of trajectory feature to represent an action video. In our best knowledge, no method has previously proposed adapting trajectory-based approach for human action recognition in depth video. We conduct evaluations on recognition accuracy using dense trajectories motion feature proposed by Wang et al. [5].

3. Proposed Method

This paper presents a effective depth video representation by adapting intensity trajectories-based motion features. First, we provide a brief review of the dense trajectories-based feature proposed by Heng Wang et al. [5]. Related parts, such as: dense sampling, tracking and feature descriptors are also referred to. Our trajectories-based approach for depth data is mentioned at the end of this section.

3.1. Dense trajectories

In order to obtain trajectories, there are two important steps: sampling and tracking. [5] 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:

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

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 [20] 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. Beside, 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.

Trajectory Shape Descriptor. This descriptor describes the shape of a trajectory in the simplest way. Given a trajectory of length L, its shape is concatenated by a sequence of displacement vectors $S = (\Delta P_t, ..., \Delta P_{t+L-1})$, where $\Delta P_t = P_{t+1} - P_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. In order to make the descriptor invariant to scale changes, the final result is then achieved by normalizing the shape vector by the overall magnitude of the displacement vectors:

$$\bar{S} = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{k=t}^{t+L-1} \|\Delta P_k\|}, \tag{2}$$

Trajectory-aligned Descriptor. The descriptors are much more complex than the trajectory shape descriptor. They 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 $n_\tau = 3$.

In order to capture the local motion and appearance around a trajectory, three kinds of descriptors have been employed: the Histogram of Oriented Gradient (HOG) [21], the Histogram of Optical Flow (HOF) [22], and the Motion Boundary Histogram (MBH) [23]. For HOG, orientation information is quantized into 8-bin histogram. HOF is 9-bin histogram. Since the feature of a trajectory is calculated and concatenated from sub-volumes of a 3D volume, the final representation has 96 dimensions for HOG and 108 dimensions for HOF. MBH descriptor computes derivatives on both horizontal and vertical components of optical flow $I_\omega = (I_x.I_y)$. Similar to HOG descriptor, the orientation information is quantized into 8-bin histogram. Since the motion information is combined along two directions, the final representation is $96 \times 2 = 192$-bin histogram. By presenting gradient of optical flow, MBH descriptor is able to suppress global motion information and only keep local relative changes in pixels.

According to the authors [22, 5, 24, 25], 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 [5]. We also conduct our experiment on all the three descriptors when compared to the depth-based state-of-the-art methods.

3.2. Pseudo-3D trajectory-based Approach for Motion Feature in Depth Data

Our proposed trajectory-based approach for human action recognition in depth data is as follow. At first, intensity representations are formed from the sequence of depth maps, as illustrated in Figure 2. In particular, we choose three representations to represents for 3 view directions: front, side, and top in 3D space. Forming the representations is necessary due to dimensional gap when we adapt 2D techniques for 3D data. After that, the dense trajectories are extracted from the intensity representations. And the feature descriptors are also computed in this step. At the next step, with each intensity representation, corresponding feature representation is quantized from raw trajectory features by apply a bag-of-words (BoW) model. An *early fusion* scheme is used to generate the final feature representation for action in the sequence of depth maps (Fig. 3).

Figure 2: Illustration on Our Proposed Method

Figure 3: Our Framework Overview

In order to generate intensity representations from the sequence of depth maps, we use the approach proposed in [14]. This technique is also used in [1]. Basically, this method projects depth maps onto three orthogonal planes in Casterian space to obtain corresponding intensity representations. However, motion representation for human action in the previous approaches is accumulated from global motion information. Therefore, these approaches must deal with the challenges from human segmentation problem in more complicated datasets. In contrast to the previous ones, we pay attention to capture local motion information for representing human actions. With the approach, we do not care the challenges for segmenting human body. To effectively use local motion information, we leverage the effectiveness of trajectory-based representation. In practice, we adopt the dense trajectory-based approach proposed in [5]. Thus, motion information in depth data can be reproduced by complementary motion information in different intensity representations.

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 [3, 19]. In fact, extracting skeleton exactly is still an unsolved problem, due to the challenges, such as cluttered background, hardware quality, camera motion, so on.

4. Experimental Settings

4.1. Dataset

We test our method on MSR Action 3D dataset. This dataset contains 20 actions, as showed in Table 1. 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.

ID	Action Name	ID	Action Name
1	high arm wave	11	two hand wave
2	horizontal arm wave	12	side-boxing
3	hammer	13	bend
4	hand catch	14	forward kick
5	forward punch	15	side kick
6	high throw	16	jogging
7	draw x	17	tennis swing
8	draw tick	18	tennis serve
9	draw circle	19	golf swing
10	hand clap	20	pick up & throw

Table 1: 20 actions in MSR Action 3D dataset

In order to conduct a fair comparison, we use the same experimental settings as [14, 19, 1, 3, 2, 4]. In the settings, the dataset is divided into three action subsets. Each subset has 8 actions (Table 2). 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).

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 2: The three action subsets used in the experiments

4.2. Evaluation Method

Figure 3 shows our evaluation framework for the trajectory-based features. We perform experiments using the proposed approach and compare with the state-of-the-art methods on depth data. We use the application available online[1] to extract dense trajectories and aligned-features. To quantize a large number of features obtained by densely sampling, the BoW model is applied. At first, in each intensity representation, we randomly get about 80,000 extracted trajectories for clustering with K-mean algorithm. Then, a codebook of 2000 visual codewords is formed for each. After that, the hard-assignment technique is used to compute histograms of the visual words on the corresponding intensity representations.

Once all the BoW histograms are generated, we adopt the late-fusion scheme with the popular Support Vector Machine (SVM) for classification. In practice, we use the precomputed-kernel technique with the histogram intersection

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

measurement for the classification step. In our implementation, we use the libSVM library published online by author[2] and perform the one-vs-all strategy for multi-class classification. We adopt the format requirements of the library to synchronize the annotation and the data. For testing, the BoW histograms of corresponding intensity representations are concatenated to generate the final feature representation. The predicted value is defined as the maximum score obtained from all the classifiers. This score shows that a human action is confused with another or not.

5. Experimental Results

This section presents the experimental results from applying our proposed approach on MSR Action 3D dataset. We also report the results on the main intensity representation (i.e. front projection). Beside, an evaluation related to selecting compensation information from other representations will be also mentioned. All the results are compared in terms of the accuracy. The best performance is highlighted in bold.

5.1. Recognize Actions from An Intensity Representation

Table 3 - lists the results from our trajectory-based approach on front representation. Interestingly, the result table indicates that this approach beats all the state-of-the-art methods based on silhouette features [14, 1], skeletal joint features [19, 3], local occupancy patterns [16, 15], normal orientation features [4] and cuboid similarity features [2]. However, the results also show that there is significant difference of the performance among the used feature descriptors.

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

Method	Accuracy (%)
Bag of 3D Points [14]	74.70
STOP [15]	84.80
EigenJoints [19]	82.33
Random Occupancy Patterns [16]	86.50
Local Occupancy Patterns [3]	88.20
Depth Motion Maps-based HOG [1]	91.63
Histogram of Oriented 4D Normals [4]	88.89
Depth Cuboid Similarity Feature [2]	89.30
Ours	**94.53**

Table 3: Results on front representation using MBH descriptor.

Action Subsets		
AS1	**AS2**	**AS3**
92.45	92.04	99.11

Table 4: Results on three action subsets.

Consider the results on action subsets reported in Table 4, we found that two subsets AS1, AS2 contain many confused actions. For example, action-pair *hammer* and *forward punch* in AS1, or *side-boxing* and *hand catch* in AS2, as showed in Table 5. When analyzing confused actions, we found that the main cause is due to similar movements. And, since depth data is textureless, it makes recognition more difficult. That is a reason why we need compensate information from other intensity representations.

	a02	**a03**	**a05**	**a06**	**a10**	**a13**	**a18**	**a20**
a02	**0.833**	0	**0.167**	0	0	0	0	0
a03	0	**0.917**	**0.083**	0	0	0	0	0
a05	0	**0.364**	**0.636**	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	**0.067**	0	**0.933**

(a) Action Subset 1

	a01	**a04**	**a07**	**a08**	**a09**	**a11**	**a12**	**a14**
a01	**1.0**	0	0	0	0	0	0	0
a04	**0.083**	**0.833**	**0.083**	0	0	0	0	0
a07	0	0	**0.786**	**0.071**	**0.071**	0	**0.071**	0
a08	0	0	0	**1.0**		0	0	0
a09	0	0	0	**0.133**	**0.867**	0	0	0
a11	0	0	0	0	0	**1.0**	0	0
a12	0	**0.133**	0	0	0	0	**0.867**	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	**0.933**	**0.067**	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 5: Confusion matrices on three subsets.

5.2. Compensate Motion Information from Other Representations

In this part, we conduct experiments based on compensating information from all the rest representations for the front representation. Figure 4 reports a better view in comparing the performance of fusion with separate representations. Expectedly, the average fusion performance, which is 96.67% accuracy, is better than all the separate ones on each representation. Obviously, our proposed approach outperforms the mentioned state-of-the-art methods.

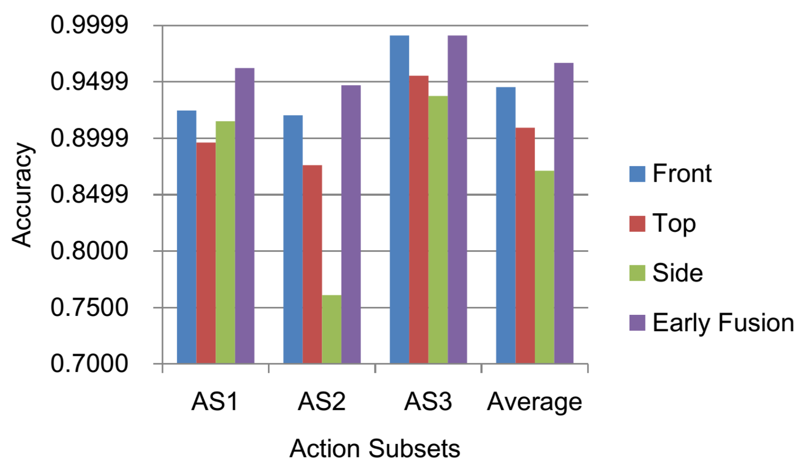


Figure 4: Results from using the early fusion scheme on representations

Beside, based on experimental results in figure 4, compensating information indicates two interesting points. The first one confirms that recognition result from front representation is better than the others (i.e. side and top). The second one shows that compensated information from other representations for front representation supports final predictions effectively. Thus, our proposed approach can be applied for any intensity-based techniques, in general.

6. Discussions

6.1. The Impact of Our Method on Descriptors

For intensity data, according to [5] 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 section, we conduct similar experiments on these descriptors to answer this issue.

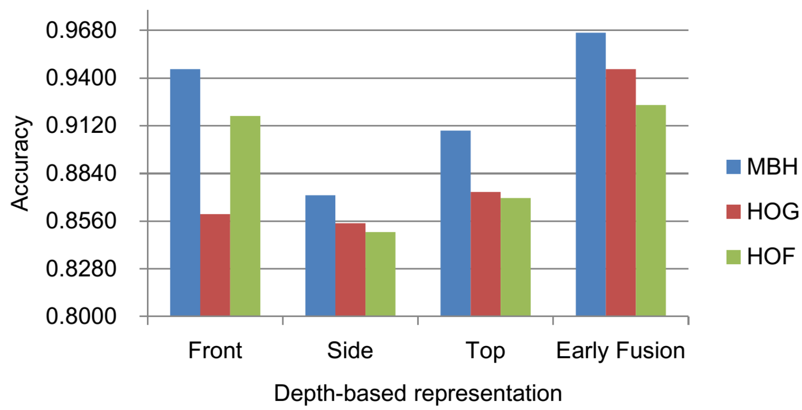


Figure 5: Results on trajectory-aligned descriptors

Figure 5 shows interesting results. Although, recognition results on descriptors HOG, HOF are not good for each intensity representation, the final results after fusing have been significantly improved. The results indicate that the performances of HOG and HOF, respectively 94.53% and 92.42%, also outperform the state-of-the-art methods, as mentioned in Table 3. 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. Evaluate the Role of Intensity Representations

In this section, we consider the role of representations to our proposed method. Figure 4 confirms that front representation achieves the best result. Obviously, it is an indispensable component to merge information. For the rest, we perform experiments on representation combinations with front representation. Experimental results are reported in Figure 6.

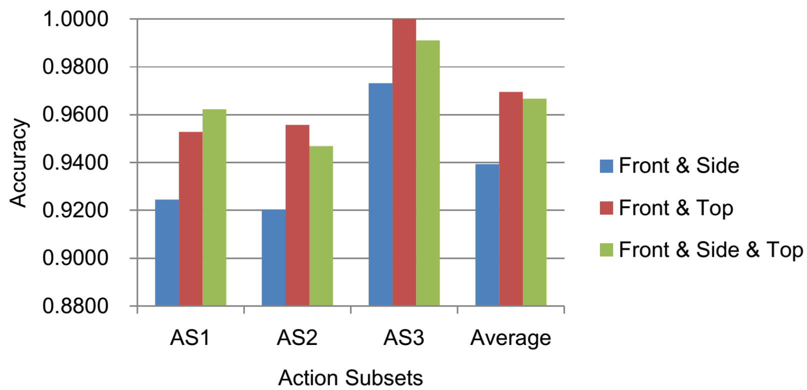


Figure 6: Results on combinations of representations

In order to conduct the experiments, we create combinations: front and side, front and top. Figure 6 indicates that the combination of front and top is better than the combination of front and side. More interestingly, the achieved performance, which is 96.95% accuracy, from the combination of front and top beats the performance based on combining all the representations, in terms of average. Actually, the discovery provides a good choice to decrease computational cost but still ensures a convincing performance.

6.3. MSR Daily Activity 3D Dataset

The MSR Daily Activity 3D dataset is proposed by [3], which bao gm 16 daily activities (Fig. 7) such as talking on the phone, reading a book, playing game,

... etc. In this dataset, background objects and subjects appear at different distances to the camera. Table 6 shows a comparison between the state-of-the-art methods on MSR Daily Activity 3D dataset. In this experiment, we conduct our trajectory-based approach only on front representation. In condition of only using depth data, [3, 4, 2] report a unexpected performance. In [2], they modified this dataset to do evaluation. It is not fair to compare. Therefore, to ensure a fair comparison, we follow a framework similar to [2] and evaluate on original MSR Daily Activity 3D dataset.

Although our method outperforms all the state-of-the-art methods, it is not our aim. 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 action focus on motion of fingers, it is very difficult to discriminate from depth noise. For *writing on a paper* and *using a latop*, hand gestures are major actions 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 is small and textureless. Thus it is very difficult to identify these actions exactly if only depending on depth data.

Method	**Accuracy**
LOP [3]	42.5
HON4D [4]	52
DSTIP&DCSF [2]	56.88
Ours	**62.5**

Table 6: Performance of Methods on MSR Daily Activity 3D Dataset. Notice that results are reported in terms of only using depth data.

(a) Reading book

(b) Drinking water

(c) Talking on a phone

(d) Playing game

(e) Writing on a paper

(f) Using a laptop

Figure 7: Sample actions on MSR Daily Activity 3D dataset

7. Conclusions

We proposed using the trajectory-based approach for human action recognition using depth data in this work. We evaluated our approach by using the dense trajectories motion feature on MSR Action 3D 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. Beside, in order to deal with confused actions due to similar movements, compensating information from other representations is proposed. Therefore, the effectiveness of our approach on depth datasets like MSR 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-trajectory-based motion information for human action can be beneficial for an action recognition system. This is also an interesting idea for our future work.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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`

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.