PAPER

Human Action Recognition from Depth Videos Using Pool of Multiple Projections with Greedy Selection

Chien-Quang LE^{†a)}, Sang PHAN^{††b)}, Thanh Duc NGO^{†††c)}, Nonmembers, Duy-Dinh LE^{††d)}, Shin'ichi SATOH^{†††e)}, Members, and Duc Anh DUONG^{†††f)}, Nonmember

SUMMARY Depth-based action recognition has been attracting the attention of researchers because of the advantages of depth cameras over standard RGB cameras. One of these advantages is that depth data can provide richer information from multiple projections. In particular, multiple projections can be used to extract discriminative motion patterns that would not be discernible from one fixed projection. However, high computational costs have meant that recent studies have exploited only a small number of projections, such as front, side, and top. Thus, a large number of projections, which may be useful for discriminating actions, are discarded. In this paper, we propose an efficient method to exploit pools of multiple projections for recognizing actions in depth videos. First, we project 3D data onto multiple 2D-planes from different viewpoints sampled on a geodesic dome to obtain a large number of projections. Then, we train and test action classifiers independently for each projection. To reduce the computational cost, we propose a greedy method to select a small yet robust combination of projections. The idea is that best complementary projections will be considered first when searching for optimal combination. We conducted extensive experiments to verify the effectiveness of our method on three challenging benchmarks: MSR Action 3D, MSR Gesture 3D, and 3D Action Pairs. The experimental results show that our method outperforms other state-of-the-art methods while using a small number of projections.

key words: action recognition, depth sequences, multiple projections, greedy method

1. Introduction

Human action recognition is one of the essential tasks in computer vision because of its many potential applications such as surveillance, content-based video search, sport video analysis, and human-computer interaction [1], [2]. With recent advances in imaging technology, it has become possible to capture depth information in real time. Depth cameras have several advantages over conventional cameras. For example, depth images are much easier to segment and are insensitive to changes in lighting conditions. In this

paper, we study the problem of human action recognition from depth video sequences.

One straightforward approach is to treat the depth information as pixel intensities for adapting state-of-the-art 2D features to depth videos. Unfortunately, this method exposes the inherent limitations of the 2D feature-based approaches. Moreover, it results in confused motion patterns that cannot be distinguished by observing 2D videos from one view. The actions of Hammer and Forward Punch shown in Fig. 1 are such examples. Another possible approach is to project depth data onto specific 2D-planes, such as the front, side and top [3], [4]. These methods aggregate discriminative information from different viewpoints to generate more discriminative representations of actions. However, this approach discards a large number of projections which may contain useful information for discriminating actions.

We handle this problem in a novel way by exploiting pools of multiple projections to obtain more discriminative information. The motivation of this technique is that two different actions may be confused in the same view but can be differentiated in some other view, as shown in Fig. 4. To do that, we project depth data onto 2D-planes via different viewpoints sampled on a geodesic dome (see Figs. 2 and 3). The projected values in these planes are turned into pixel intensities. As a result, we obtain a much larger number of projections. Using these projections, we can generate a set of 2D action clips for each depth video. It is worth noting that a subset of the 2D action clips comes from special cases of the two aforementioned approaches, such as front, side and top projections. Once the set has been obtained, we can employ state-of-the-art 2D techniques for action recognition.

We have to deal with the following challenges in order to exploit multiple projections in action recognition. The first one is that the extraction is too time consuming, since it requires a huge number of 2D action clips to be processed. Secondly, using all possible viewpoints to recognize actions may not be necessarily optimal. As shown in Figs. 1 and 4, there are actions that cannot be distinguished from certain viewpoints. These viewpoints can compromise the recognition performance when combined with other discriminative viewpoints. To handle these challenges, we propose a greedy method, called multiple projection-based greedy selection (MPGS), to select an optimal subset of viewpoints. MPGS iteratively selects a set of viewpoints based on the

Manuscript received October 16, 2015.

Manuscript revised March 12, 2016.

Manuscript published April 25, 2016.

[†]The author is with the Graduate University for Advanced Studies (SOKENDAI), Kanagawa-ken, 240–0115 Japan.

^{††}The authors are with the National Institute of Informatics (NII), Tokyo, 101–8430 Japan.

^{†††}The authors are with the University of Information Technology, Ho Chi Minh City, Vietnam.

a) E-mail: lqchien@nii.ac.jp

b) E-mail: plsang@nii.ac.jp

c) E-mail: thanhnd@uit.edu.vn

d) E-mail: leddy@nii.ac.jp

e) E-mail: satoh@nii.ac.jp

f) E-mail: ducda@uit.edu.vn

DOI: 10.1587/transinf.2015EDP7430

Fig. 1 Illustration of our ad hoc projection-based approach. The original sequence of depth maps is projected onto three 2D-planes, corresponding to three views: *front* (XY), *side* (ZY) and *top* (XZ), to form 2D videos. After that, dense trajectory motion features are extracted from the 2D videos. Two actions of *Forward Punch* and *Hammer* may be confused if we view them from the front because they both contain “lift arm up” and “stretch out” movements that are indistinguishable when viewed from the front. However, if we properly use the depth information in the motion pattern analysis, it is possible to extract discriminative motion patterns which are inaccessible from one fixed view (i.e., from the front) but discernible from different views (i.e., from the side or top).

discrimination ability of their combination. In each iteration, only one best viewpoint is added to the current combination. This method converges at a given number of viewpoints or no selected viewpoint. The key idea is that the best complementary viewpoints will be considered first when searching for the optimal combination.

We conducted comprehensive experiments on three challenging benchmarks: the MSR Action 3D dataset, MSR Gesture 3D dataset, and 3D Action Pairs dataset. The results show that our method not only outperforms the baseline methods that do not use selected-projection-based features but also beats other state-of-the-art methods in action recognition using depth video sequences.

The main contribution of this paper is to propose a new method to obtain a small yet robust combination of projections for multiple projection-based action recognition from depth videos. The combination is optimal in situations in which different actions are almost indistinguishable from the same viewpoint. In addition, we propose to use dense trajectory-based features to recognize 2D action clips generated from depth videos. This proposal is motivated by the success of dense trajectory-based features in 2D action recognition [5]–[10]. To the best of our knowledge, no previous multiple projection-based study has exploited these features for recognizing actions in depth videos.

The work presented in this paper is an extension of our earlier work [11], with the following additions:

- We compared our method with a number of baselines, including the single viewpoint-based method, specific viewpoint-based method, and all sampled viewpoint-based method.
- We also compared it with other state-of-the-art methods.
- Experimental results are reported for the MSR Gesture

3D dataset and 3D Action Pairs dataset, which are challenging benchmarks.

The rest of the paper is organized as follows. Related work is briefly discussed in Sect. 2. Section 3 describes the proposed method. The experimental results are presented in Sect. 4. Finally, Sect. 6 concludes the paper.

2. Related Work

The recent approaches to recognition of actions in depth sequences can be divided into two main categories. The first category includes approaches originally proposed for conventional RGB videos that have been extended to deal with depth sequences. The second category consists of approaches that consider the nature of depth data.

Regarding the first category, [4] proposed Depth Motion Maps (DMMs) to capture global activities in depth sequences. DMMs are generated by stacking the motion energies of depth maps projected on three orthogonal Cartesian planes. A Histogram of Oriented Gradients (HOG) [12] is then computed from the DMMs to represent an action. Another approach, proposed by [13], uses filtering to extract spatio-temporal interest points from depth videos (DSTIPs). This can be considered as an extension of the work of [14] to depth data. First, 2D and 1D filters (e.g., Gaussian and Gabor filters) are respectively applied to the spatial dimensions and temporal dimension of the depth video. A correction function is used to suppress noisy depth points. The points with the largest responses resulting from this filtering are selected as the DSTIPs for each video. In addition, a depth cuboid similarity feature (DCSF) is used to describe a 3D cuboid around each DSTIP, in which the apparent size varies with the depth. The study [13] illustrated the effectiveness of using 2D techniques for depth data.

Regarding the second category, [3] propose to use a bag of 3D points to characterize a set of salient postures. The 3D points are extracted from the contours of planar projections of the 3D depth map. About 1% of the 3D points are sampled in order to calculate a feature. Unlike [3], the methods described in [15]–[17] to represent features in action sequences.

The authors of [15] propose a new feature descriptor, called Space-Time Occupancy Patterns (STOP). This descriptor is formed from the sparse cells of a 4D space-time grid dividing up a sequence of depth maps. The points inside the sparse cells are typically on the silhouette or on the moving parts of an object. [16] created semi-local features, called Random Occupancy Pattern (ROP) features, from randomly sampled 4D sub-volumes of different sizes and at different locations. The random sampling is performed according to a weighted scheme in order to effectively explore the large dense sampling space. The authors also used sparse coding to robustly encode these features. [17] designed features, named Local Occupancy Patterns (LOPs), to describe the local “depth appearance” of each joint of the body. LOP features are computed on the basis of a 3D point cloud around a particular joint. In addition, they concatenate the LOP features with skeleton information-based features and apply a Short Fourier Transform to obtain Fourier Temporal Pyramid features at each joint. The features are utilized in a novel actionlet ensemble model to represent each action in the video.

Recently, [18] presented a new descriptor for depth maps, named Histogram of Oriented 4D Surface Normals (HON4D). To construct the HON4D, 4D normal vectors are first computed from the depth sequence. Next, these 4D normal vectors are distributed into spatio-temporal cells. To quantize the 4D normal vectors, the 4D space is quantized by using the vertices of a regular polychoron. The quantization is refined by additional projectors to make the 4D normal vectors in each cell denser and more discriminative. Afterwards, the HON4D features in the cells are concatenated to represent an action in the depth video. A number of studies [3], [15]–[18] have shown the effectiveness of approaches that directly exploit 3D data for action recognition. However, several constraints, such as segmentation of the human body in [3], or the location of the body in [15]–[18], have somewhat limited their applicability.

Inspired by the results of J. Shotton et al. [19] and L. Xia et al. [20], some studies [21], [22] have developed skeleton-based methods for depth map sequences. X. Yang et al. [21] 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 of the frames. In addition, non-quantized descriptors and *Video-to-Class* distance computations have proved to be effective for action recognition. J. Luo et al. [22] proposed a new discriminative dictionary learning algorithm (DL-GSGC) to better represent 3D joint features. This algorithm incorporates both group sparsity and geometric constraints. In addition, to keep temporal

information, a temporal pyramid matching method can be used on each sequence of depth maps. Most of the skeleton information-based approaches have state-of-the-art performance on benchmark datasets. However, their dependence on skeleton information is a detriment when such information is not available or incorrect.

Different from the previous studies [3], [4], we do not require the human body to be segmented. Moreover, unlike methods in [17], [21], no skeleton has to be extracted. Additionally, we investigate the benefit of using multiple projections from depth videos. In contrast to earlier findings [3], [4], our method provides an optimal combination of projections to improve recognition accuracy significantly. Finally, we leverage the trajectory feature to represent actions in video. To the best of our knowledge, no previous study has attempted to adapt the dense trajectory-based approach to action recognition in depth video. We evaluated the recognition accuracy of our method on depth video using the dense trajectories proposed by H. Wang et al. [5].

3. Proposed Method

3.1 3D Action Decomposition

Our aim is to exploit discriminative motion patterns of an action from different viewpoints. The discriminative motion patterns are used to accurately recognize confused actions. We collect these patterns as follows. First, we decompose each 3D action into a set of 2D actions. A simple and effective way is to map each 3D action to given viewpoints. Each viewpoint is defined by a pair consisting of a polar angle θ and azimuthal angle ϕ , as illustrated in Fig. 2. We define $(\phi, \theta) = (\pi/2, \pi/2)$ as the frontal view. In particular, viewpoints are calculated in a step of $\pi/6$ (see Fig. 3). There are 22 viewpoints in total.

Next, we use the dense trajectory-based approach [5], [6] to extract discriminative motion patterns. We provide a brief description of this approach in Sect. 3.2.

Fig. 2 The mapping model is used to decompose 3D actions. Each viewpoint P is formed by a pair of azimuthal angle ϕ and polar angle θ .

Fig. 3 Viewpoint distribution.

3.2 Feature Extraction

Dense trajectories [5] have been demonstrated to be effective in action recognition. Our motivation for using dense trajectories is to capture discriminative motion patterns from 2D actions decomposed from 3D actions. To extract trajectories from videos, Wang et al. [5], [6] proposed sampling on a dense grid with a step size of 5 pixels. The sampling is performed at multiple scales with a factor of $1/\sqrt{2}$. Displacement information from a dense optical flow field is used to track the dense points. 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 the next frame $t + 1$ by using:

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

where $\omega = (u_t, v_t)$ denotes the dense optical flow field, M is the kernel used in median filtering, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of P_t . The algorithm presented in [23] uses dense optical flow. To avoid drifting, it sets a suitable trajectory length of 15 frames. It also removes trajectories with sudden changes.

Once the trajectories have been extracted, two kinds of descriptor, i.e., a trajectory shape descriptor and a trajectory-aligned descriptor, can be used. 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 up into a 3D grid (spatially into $n_\sigma \times n_\sigma$ grid and temporally into n_τ segments). The default settings of these parameters are $N = 32$ pixels, $L = 15$ frames, $n_\sigma = 2$, and $n_\tau = 3$. In our experiments, we only used the Motion Boundary Histogram (MBH), a trajectory-aligned descriptor, because of its effectiveness [5], [6].

3.3 Viewpoint Selection

As mentioned in Sect. 3.1, the sampling viewpoints method provides rich information while also dealing with the following challenges. The first is the high computational cost due to a huge number of augmented action clips. Second, action representations created from all available viewpoints

Fig. 4 Illustration of two actions *high arm wave* and *two hand wave*. We can easily discriminate the two actions from viewpoint 13, but not from viewpoint 4.

may contain confusing information, as shown in Fig. 4. Therefore, it is critical to collect a compact set of viewpoints.

The viewpoint selection method can be divided into two stages: data organization and greedy selection from the pool of viewpoints. In the first stage, we divide actions into a development dataset for training and a test dataset for testing. In addition, we further divide the development dataset into a training dataset and validation dataset. Both datasets are used in the second stage. We obtain the three datasets in the cross-subject setting. The evaluation protocol used in our selection experiments was mean average precision (mAP). Suppose that a group of viewpoints has been obtained (the pool of multiple viewpoints). Our objective is to select a few viewpoints that can form an optimal combination to represent actions from depth videos. We use a greedy method to select representative viewpoints (Algorithm 1). This method iteratively adds viewpoints from the pool of multiple viewpoints (*POOL*) to the final viewpoint set (*SEL*). In each step of the selection, classifiers for each candidate viewpoint set are learned and evaluated. In the first iteration, classifiers for each single candidate viewpoint are learned, and the viewpoint with the best evaluation performance is added to the final viewpoint set. The learning and evaluation step is repeated using the current selected viewpoint set until a stopping condition is reached. The algorithm stops at one of the following conditions: (1) the number of collected viewpoints is greater than a given threshold, or (2) none of the viewpoints that could be added yields a benefit. The outcome of this algorithm is an optimal subset of viewpoints selected from the pool.

3.4 Feature Representation

In order to represent videos, we use the bag-of-words model [24]. So far, we have selected M viewpoints. For each viewpoint we train a codebook with 2000 codewords using the k-means algorithm. In order to reduce the computational cost of clustering, we used another dataset, the Northwestern-UCLA Multiview Action 3D dataset [25], to

Algorithm 1: Greedy method of viewpoint selection

input : A set of available viewpoints - $POOL$
Number of selected viewpoints - M
output: A set of selected viewpoints - SEL , $SEL \subseteq POOL$

```

1   $SEL \leftarrow \emptyset$ ;
2   $mAP\_max \leftarrow -inf$ ;
3  for  $t \leftarrow 1$  to  $M$  do
4       $change \leftarrow false$ ;
5       $viewID \leftarrow 0$ ;
6      foreach viewpoint  $v \in POOL$  do
7           $SEL\_temp \leftarrow SEL \cup v$ ;
8           $mAP\_SEL \leftarrow Compute\_mAP(SEL\_temp)$ ;
9          if  $mAP\_SEL > mAP\_max$  then
10              $mAP\_max \leftarrow mAP\_SEL$ ;
11              $viewID \leftarrow index(v)$ ;
12              $change \leftarrow true$ ;
13         end
14     end
15     if  $change$  then
16          $v^* \leftarrow POOL(viewID)$ ;
17          $POOL \leftarrow POOL \setminus v^*$ ;
18          $SEL \leftarrow SEL \cup v^*$ ;
19     else
20          $break$ ;
21     end
22 end

```

Fig. 5 Example frames showing actions in the MSR Action 3D dataset [3].

build universal view-codebooks. Different from the datasets used to evaluate our method, each activity in this dataset consists of many complicated and arbitrary movements. Therefore, this dataset contains a large number of movements that cover all of the movements in the evaluation datasets. Because of the characteristics of hand gestures in the MSR Gesture 3D dataset, the codebooks were only used for training and testing with MSR Action 3D and 3D Action Pairs.

For each video, we quantize its local descriptors to form M video representations, corresponding to M viewpoints. After that, we concatenate these representations to obtain the final feature representation. We use the libSVM library [26] with the histogram intersection kernel (see Eq. (2)) for classification. This kernel can be defined as follows. Suppose that a and b are the histograms of two action

Table 1 Twenty actions in MSR Action 3D dataset.

Index	Action Name	Index	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

videos (i.e., feature representations after the viewpoint selection). Both histograms consist of n bins and a_i , b_i are denoted as the i -th bins of a and b respectively. Let us assume that a and b have been normalized.

$$K(a, b) = \sum_{i=1}^n \min(a_i, b_i), a_i \geq 0, b_i \geq 0 \quad (2)$$

In order to recognize $\mathcal{N}$ action categories, we build $\mathcal{N}$ corresponding binary classifiers (one-versus-all classifiers) $f^{(j)}$, $j = 1..N$. With each test sample x , we obtain $\mathcal{N}$ scores from all classifiers. Then, we select the action category that has the maximum score as the decision for the test sample x . The predicted category of x can be also defined by:

$$\hat{j} = \arg \max_j f^{(j)}(x), j = 1..N \quad (3)$$

4. Experiment

4.1 Datasets

4.1.1 MSR Action 3D Dataset

MSR Action3D dataset consists of temporally segmented action sequences captured by a depth camera. This dataset is challenging because the actions contained in it often appear similar. This dataset contains 20 action types, as described in Table 1. The actions were performed two or three times by ten subjects and reasonably cover various movements of the limbs, torso, and their combinations. We evaluated our method in two different experimental settings.

1. *Set.1*: We divided the twenty actions into three subsets in as the original setting [3]. Each subset consisted of eight actions (Table 2). The AS1 and AS2 subsets grouped actions that had similar movements. The AS3 subset grouped complex actions together. For instance, the *hammer* action seems to be confused with the *forward punch* action in AS1, and the *hand catch* and *side boxing* actions are similar looking movements in AS2.
2. *Set.2*: This experimental setting was the same as the one described in [16]. We evaluated our method on a total of twenty actions. This setting is considered more challenging than the three action subsets due to the greater number of action classes.

Table 2 The three action subsets used in the experiments.

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 3 Twelve gestures in MSR Gesture 3D dataset.

Index	Gesture Name	Index	Gesture Name
1	Bathroom	7	Past
2	Blue	8	Pig
3	Finish	9	Store
4	Green	10	Where
5	Hungry	11	J
6	Milk	12	Z

Fig. 6 Example frames showing hand gestures from the MSR Gesture 3D dataset [16].

Our experiments were cross-subject tests; i.e., one half of the subjects were used for training, the rest for testing.

4.1.2 MSR Gesture 3D Dataset

The Gesture3D dataset [16] is a hand gesture dataset of depth sequences captured by a depth camera. It contains twelve American Sign Language gestures (see Table 3). Examples of these gestures are shown in Fig. 6. Ten subjects performed each gesture two or three times. In total, the dataset contains 333 depth sequences. The main challenge here is self-occlusion. We used the experimental settings in [16] (i.e., leave-one-subject-out cross-validation) to evaluate our approach.

4.1.3 3D Action Pairs Dataset

The 3D Action Pairs dataset [18] is a new type of action dataset. Different from the other two datasets, this dataset contains pairs of actions in which the motion and shape cues are similar, but different in their order of performance (see Fig. 7). It is a challenging task to differentiate the prominent cues within each pair. There are twelve actions, as shown

Fig. 7 Example frames showing pairs of the 3D Action Pairs dataset [18]. Each column shows two frames from a pair of actions. For example, in the first column, both actions, i.e., *Pick up a box* and *Put down a box*, have similar motion information and shape cues; however, they are performed in different spatio-temporal orders.**Table 4** Twelve actions in 3D Action Pairs dataset.

Index	Action Name	Index	Action Name
1	Pick up a box	7	Wear a hat
2	Put down a box	8	Take off a hat
3	Lift a box	9	Put on a backpack
4	Place a box	10	Take off a backpack
5	Push a chair	11	Stick a poster
6	Pull a chair	12	Remove a poster

in Table 4. Each action was performed three times by ten subjects. Actions from the first five subjects were used for testing, the rest for training.

4.2 Experimental Results

This section presents the experimental results of our approach on the three aforementioned datasets. First, we describe the experimental results showing the influence of single projections in action recognition. Second, we discuss the role of compensating information from multiple projections with naive selection. Finally, we compare our MPGS method with the previous and state-of-the-art methods. The experimental results are reported in terms of recognition accuracy. The best performance is highlighted in bold.

4.2.1 Action Recognition from Single Projections

We evaluated the dense trajectory-based approach, as described in Sect. 3.2, for action recognition on single projections. Figure 8 presents the recognition performance of the single projections on three benchmark datasets. For each dataset, the experimental results indicate that the recognition performance varies over projections. Additionally, the recognition performance reaches maximum at different viewpoints (or projections). For example, on MSR Action 3D, we obtain maximum accuracies of 91.4% and 88.4% with experimental settings 1 and 2, respectively, at viewpoint-13. In contrast, the maximum accuracy is 89.3% at viewpoint-11 on MSR Gesture 3D and 93.3% at viewpoint-10 on 3D Action Pairs.

Moreover, there are some actions with similar move-

Fig. 8 Recognition performance (accuracy) of the single projection-based method on MSR Action 3D, MSR Gesture 3D and 3D Action Pairs.

Table 5 Recognition performance of combinations of multiple projections.

Dataset	Average Accuracy (%)				
	Front	Side	Top	Front + Side + Top	All Viewpoints
MSR Action 3D (Set.1)	91.4	83.5	85.7	96.1	95.5
MSR Action 3D (Set.2)	88.4	77.5	78.2	93.1	92.0
MSR Gesture 3D	83.3	77.9	77.6	88.6	91.6
Action Pairs 3D	93.3	76.1	89.4	93.3	84.4

ments in the same view direction. As shown in Fig. 4, High Arm Wave and Two Hand Wave are two such examples. Therefore, it is necessary to combine multiple projections.

4.2.2 Action Recognition from Multiple Projections

Here, we describe experiments demonstrating the effectiveness of combining multiple projections. We compare two combination strategies: (1) combining front, side and top projections as proposed in [3], [4]. (Note that these projections correspond to viewpoints 13, 4 and 1 in our method); (2) combining all available projections. For both strategies, the action in a depth sequence is represented by concatenating the feature vectors computed from the corresponding projections.

Table 5 shows the recognition performance of combinations of multiple projections. The results for the first strategy indicate that the recognition performance of front, side and top combination is better than that of the single projections. This means we can achieve an improvement if we can employ much more discriminative motion information from other projections.

However, compensating information from unreliable sources can compromise the discrimination ability. As presented in Table 5, although the combination of all viewpoints yields a comparative improvement over front, side and top combination on MSR Gesture 3D, it does not so on MSR Action 3D and 3D Action Pairs. For example, for 3D Action Pairs, the recognition performance declines in comparison with the single viewpoint-based method, because the information from all projections is fused together.

Table 6 Recognition performance of a combination of selected projections.

Dataset	Average Accuracy (%)	#Viewpoints
MSR Action 3D (Set.1)	97.1	7
MSR Action 3D (Set.2)	97.9	7
MSR Gesture 3D	92.8	3
3D Action Pairs	96.7	3

These results show that it is necessary to select a subset of projections for action recognition.

4.2.3 Action Recognition from Selected Projections

Here, we report the experimental results of our projection selection method (MPGS). The results (Table 6) show that MPGS can significantly improve the recognition accuracy by effectively exploiting much more discriminative motion information using the pool of multiple projections. Obviously, it is not necessary to extract motion information from all projections in the pool. The motion information provided from our MPGS is robust enough to accurately recognize actions.

In addition, the computational cost is significantly lowered as a result of using a small yet robust combination of selected projections. For MSR Action 3D, the compact subset of projections includes viewpoints: 13, 7, 15, 9, 1, 11, and 3 in the selection order of Algorithm 1. For 3D Action Pairs, the selected subset includes viewpoints: 1, 10, and 13. For MSR Gesture 3D, the number of viewpoints in the final selected subset depends on by the leave-one-subject-out experimental setting, and Table 6 shows the maximum

Table 7 Comparison with state-of-the-art methods.

Method	MSR Action 3D (Set.1)	MSR Action 3D (Set.2)	MSR Gesture 3D	3D Action Pairs
Bag of 3D Points [3]	74.70	-	-	-
Space-Time Occupancy Patterns [15]	84.80	-	-	-
Eigenjoint [21]	82.33	-	-	-
Random Occupancy Patterns [16]	-	86.50	88.50	-
Local Occupancy Patterns [17]	-	88.20	-	-
Depth Motion Maps-based HOG [4]	91.63	-	-	-
Histogram of Oriented 4D Normals [18]	-	88.89	92.50	96.67
DSTIPs & DCSF [13]	-	89.30	-	-
DL-GSGC [22]	96.70	-	-	-
Our MPGS	97.90	97.10	92.80	96.70

Fig. 9 Recognition performance of projection-based method on MSR Action 3D, MSR Gesture 3D and 3D Action Pairs.

number. These results have shown that the selected viewpoints are different among the datasets. The reason is that the confusing actions are different from dataset to dataset. Therefore, different viewpoints are selected to discriminate different confusing cases. For example, for MSR Action 3D dataset, accuracy is improved because viewpoint-7 is selected to better differentiate confusing actions such as ‘hammer’ and ‘forward punch’. Meanwhile, for 3D Action Pairs dataset, for confusing actions such as ‘put down a box’ and ‘wear a hat’, viewpoint-7 is not selected.

The selected subsets for each benchmark dataset indicate that MPGS can cover a wide range of viewpoints in which the selected viewpoints are not necessarily equally spaced. Thus, the viewpoints together can effectively compensate information. Furthermore, notice that the compact subsets cover the front (13), side (4) and top (1) viewpoints or their neighbors. This shows that the front, side and top combination of viewpoints is complementary, but not sufficient. In general, MPGS is much more effective than the single projection, specific projection, and pool of multiple projections methods, as shown in Fig. 9.

Figures 10 and 11 show the confusion matrices corresponding to the evaluations on MSR Action 3D and 3D Action Pairs. It is clear that our approach significantly improves recognition performance on easily confused actions. For example, (Lift a box; Place a box) and (Stick a poster; Remove a poster) are easily confused action pairs in 3D Action Pairs, and (Hand catch; Side boxing) and (Draw x; Draw tick) are easily confused actions in MSR Action 3D. However, there are also some samples which are incorrectly

predicted even though an optimal subset of viewpoints is selected. The misidentification occurs due to the arbitrariness in action execution. In addition, occlusion is also a reason. It can be either self-occlusion or common occlusion caused by surrounding objects. Besides, our method proposes a solution to exploit a pool of multiple viewpoints. It does not guarantee a global optimization but it is more effective than the aforementioned baselines, such as using single viewpoint, specific viewpoints, and all viewpoints.

4.2.4 Comparison with State-of-the-Art Methods

Table 7 shows the outcome of the comparison with state-of-the-art methods. The compared methods use various feature representations, such as, silhouette-based features [3], [4], skeleton joint-based features [21], occupancy pattern-based features [15]–[17], normal orientation-based features [18], and depth cuboid similarity features [13]. The approaches in [15]–[17] outperform those in [21] since 3D cloud points are less sensitive to occlusions and provide additional shape information compared with skeleton joints. Unlike [15]–[17] employs both depth and skeleton information to extract motion and shape features. However, these features are simply concatenated; therefore their relations are not encoded. Additionally, skeleton information is not effective for human hand gestures, the skeleton information-based methods cannot be used for hand gesture-based applications. In depth motion maps [4], depth sequences are projected onto three planes where the temporal order of shape/motion cues is eliminated and thus these methods suffer from inner-paired confusion. MPGS employs dense trajectories; therefore, it captures both motion and shape features. Moreover, our method operates in a pool of multiple viewpoints; thus it can exploit more discriminative information. As shown in Table 7, MPGS achieves the best performance on all three benchmarks under all settings.

In addition, our method effectively exploits dense trajectory-based features. Moreover, it is not influenced by preprocessings, such as human body segmentation and skeleton extraction, which play important roles in some state-of-the-art methods. This means our approach would have numerous practical applications.

Fig. 10 Confusion matrix of our method on MSR Action 3D (experimental setting 2).

Fig. 11 Confusion matrix of our method on 3D Action Pairs.

5. Discussion

5.1 Feature Representation

To the best of our knowledge, selecting effective features from depth data using a deep neural network (DNN) has not been well-studied. However, DNN features have been successfully applied for various tasks. In practice, we can utilize our multiple projection method as a data augmentation strategy and train a DNN model with the augmented data. Due to the robustness of DNN, it is expected to select good

features for this task. However, this is not the focus of this paper, which is a greedy method to select a complementary viewpoints. Moreover, designing and training a good DNN model from scratch is not a trivial task.

5.2 Threshold of Selected Viewpoints

In Fig. 12, we show the influence of the threshold of selected viewpoints M on the performance of our algorithm. These results are reported on MSR Action 3D dataset and 3D Action Pairs dataset. Our algorithm stops with a small number of selected viewpoints. Specifically, when the threshold

Fig. 12 Classification accuracy on the threshold of selected viewpoints.

$M \geq 8$, our algorithm does not yield a better performance on the datasets. In this case, the second stopping condition will be satisfied. In the experiments, we set the threshold to be the number of all available viewpoints. In practice, threshold M is used to control a trade-off between the reasonable computational cost and the good performance on new datasets.

6. Conclusion

We presented a method using a pool of multiple projections for recognizing actions in 3D videos. The key idea is to exploit depth information for better use with state-of-the-art features for 2D video such as dense trajectories. To this end, 3D data is projected onto multiple planes from corresponding viewpoints and features are extracted from these planes. We devised a greedy method to select an optimal subset of viewpoints for fusion that yields superior performance compared with the baseline methods we tested. Moreover, our contributions are all complementary and lead to outstanding results in comparison with the state-of-the-art, as demonstrated by our extensive experiments on benchmark datasets.

In this study, we only investigated action recognition using depth data obtained from one camera. Therefore, self-occlusion is still a problem that can compromise the discrimination power. In order to overcome this challenge, it is necessary to fully capture 3D information from motions. One possible way forward is to use wearable sensors for capturing 3D motion information of human body parts [27]. However, it is not easy to apply the sensors in practice. Another way is to exploit motion information using multiple cameras. With multiple cameras, we can collect much more discriminative motion information, and this should lead to an improvement in recognition accuracy. In the future work, we will try this idea within our multiple projection-based framework.

Acknowledgments

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2013-26-01.

References

- [1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol.43, no.3, Article No.16, 2011.
- [2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol.115, no.2, pp.224–241, 2011.
- [3] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.9–14, 2010.
- [4] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," *Proc. 20th ACM International Conference on Multimedia*, pp.1057–1060, 2012.
- [5] H. Wang, A. Kläser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," *IEEE Conference on Computer Vision & Pattern Recognition*, pp.3169–3176, Colorado Springs, United States, June 2011.
- [6] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol.103, no.1, pp.60–79, 2013.
- [7] H. Wang and C. Schmid, "Action recognition with improved trajectories," 2013 IEEE International Conference on Computer Vision (ICCV), pp.3551–3558, 2013.
- [8] D. Oneata, M. Douze, J. Revaud, S. Jochen, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, K. McGuinness, S. Chen, N.E. O'Connor, K. Chatfield, O. Parkhi, R. Arandjelovic, A. Zisserman, F. Basura, and T. Tuytelaars, "Axes at trecvid 2012: KIS, INS, and MED," *TRECVID Workshop*, 2012.
- [9] P. Natarajan, P. Natarajan, S. Wu, X. Zhuang, A. Vazquez-reina, S.N. Vitaladevuni, C. Andersen, R. Prasad, G. Ye, D. Liu, S.F. Chang, I. Saleemi, M. Shah, Y. Ng, B. White, L. Davis, A. Gupta, and I. Haritaoglu, "Bbn viser trecvid 2012 multimedia event detection and multimedia event recounting systems," *NIST TRECVID Workshop*, 2012.
- [10] S. Phan, T.D. Ngo, V. Lam, S. Tran, D.-D. Le, D.A. Duong, and S. Satoh, "Multimedia event detection using segment-based approach for motion feature," *Journal of Signal Processing Systems*, vol.74, no.1, pp.19–31, 2014.
- [11] C.-Q. Le, T.D. Ngo, D.-D. Le, S. Satoh, and D.A. Duong, "Human action recognition from depth sequences using multi-projection based representation," 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), pp1–6, 2015.
- [12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, pp.886–893, 2005.
- [13] L. Xia and J.K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2834–2841, 2013.
- [14] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65–72, 2005.
- [15] A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, and M.F.

Campos, "Stop: Space-time occupancy patterns for 3D action recognition from depth map sequences," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science*, vol.7441, pp.252–259, Springer, Berlin, Heidelberg, 2012.

- [16] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," *Computer Vision – ECCV 2012, Lecture Notes in Computer Science*, vol.7573, pp.872–885, Springer, Berlin, Heidelberg, 2012.
- [17] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1290–1297, 2012.
- [18] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.716–723, 2013.
- [19] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol.56, no.1, pp.116–124, 2013.
- [20] L. Xia, C.-C. Chen, and J.K. Aggarwal, "Human detection using depth information by Kinect," *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.15–22, 2011.
- [21] X. Yang and Y.L. Tian, "Eigenjoints-based action recognition using naïve-Bayes-nearest-neighbor," *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.14–19, 2012.
- [22] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," *2013 IEEE International Conference on Computer Vision (ICCV)*, pp.1809–1816, 2013.
- [23] G. Farneback, "Two-frame motion estimation based on polynomial expansion," *Image Analysis, Lecture Notes in Computer Science*, vol.2749, pp.363–370, Springer, Berlin, Heidelberg, 2003.
- [24] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," *Proc. Ninth IEEE International Conference on Computer Vision*, pp.1470–1477, 2003.
- [25] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2649–2656, 2014.
- [26] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol.2, no.3, Article No.27, 2011.
- [27] H.-T. Cheng, F.-T. Sun, M. Griss, P. Davis, J. Li, and D. You, "NuActiv: Recognizing unseen new activities using semantic attribute-based learning," *Proc. 11th Annual International Conference on Mobile Systems, Applications, and Services*, pp.361–374, 2013.

Chien-Quang Le received his BS and MS degrees in Computer Science from the University of Science, Hue City, Vietnam in 2005 and 2007, respectively. He currently is a PhD student at the Graduate University for Advanced Studies (SOKENDAI), Japan. His research interests include pattern recognition, computer vision, and multimedia analysis.

Sang Phan is a project researcher at the National Institute of Informatics (NII), Japan. He obtained his PhD degree at The Graduate University for Advanced Studies (SOKENDAI) in 2015. Before that, he received his BS and MS degrees in 2009 and 2012 respectively from the University of Science, Vietnam National University, Ho Chi Minh City. His research interests include computer vision and multimedia analysis.

Thanh Duc Ngo obtained his BS degree in Computer Science from the University of Science, Ho Chi Minh City, Vietnam in 2006, and his PhD degree in 2013 from The Graduate University for Advanced Studies (SOKENDAI), Japan. He is currently a lecturer at the Multimedia laboratory, the University of Information Technology, Ho Chi Minh City, Vietnam. His research interests include pattern recognition, computer vision and multimedia analysis.

Duy-Dinh Le received his BS and MS degrees in 1995 and 2001, respectively, from the University of Science, Ho Chi Minh City, Vietnam, and his PhD degree in 2006 from The Graduate University for Advanced Studies (SOKENDAI), Japan. He is currently an associate professor at the National Institute of Informatics (NII), Japan. His research interests include semantic concept detection, video analysis and indexing, pattern recognition, machine learning, and data mining.

Shin'ichi Satoh is a professor at the University of Tokyo, Japan. He received his BE degree in 1987, ME and PhD degrees in 1989 and 1992 from the University of Tokyo. His research interests include video indexing and understanding, high-dimensional data structure, drawing understanding and image similarity matching. He was a visiting scientist at the Robotics Institute, Carnegie Mellon University, from 1995 to 1997. He is a member of IPSJ, ITEJ, IEEE-CS, and ACM.

Duc Anh Duong is a professor at the University of Information Technology, Ho Chi Minh City, Vietnam. He obtained his B.Sc. and M.Sc. in Computer Science from the University of Ho Chi Minh City in 1990 and 1995, respectively. He received his Ph.D. degree in Mathematics from the University of Science, VNU-HCM, Vietnam in 2002. He was a visiting researcher at Japan Advanced Institute of Science and Technology (JAIST), Japan from 2008 to 2009. His research interests include image processing, computer vision and pattern recognition, cryptography and security, geographic information systems, and computer graphics.

He is currently the Vice-President of Vietnam National University, Chair of Program of Information Technology and Electronics of Ho Chi Minh City, and Chair of Program of Information Security of Ho Chi Minh City, Vietnam. He is a member of the ACM and IEEE.