

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

Author1's name

Adress 1

Author2's name

Adress 2

Author3's name

Adress 3

Abstract

Dense trajectory-based approaches on 2D videos have been demonstrated state-of-the-art at action recognition since it can capture most discriminative motions. In this work, we extend the approach on depth video and show its effectiveness on action recognition. We extract dense trajectories from 2D videos transformed from depth video and apply trajectory-aligned descriptors to calculate motion features. Further, we present a projection method to view actions under different directions that provide additional information for recognizing actions more exactly. We evaluate this approach on framework of action recognition using the benchmark MSR Action 3D and MSR Activity Daily 3D datasets. Evaluation results show that our proposed approach is effective for action recognition on depth video and outperforms the state-of-the-art methods.

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

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, and complicated environment (e.g background clutter and illumination variation), action recognition is still a challenging problem. In order to solve this problem, related approaches can be divided into three major directions, including silhouette-based [3–6], salient point-based [7–12] and trajectory-based [13–15]. All the three approaches, especially, 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].

In the past decades, most studies in human action recognition mainly investigate on video sequences captured by traditional 2D cameras. Although, there are many advanced approaches for action recognition in domain of 2D videos, 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 [20–27]. However, in 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 dense trajectory-based approach leverages dense sampling to keep most discriminative trajectories in video. Therefore, in order to effectively exploit this approach on depth video, it is necessary to exactly extract discriminative

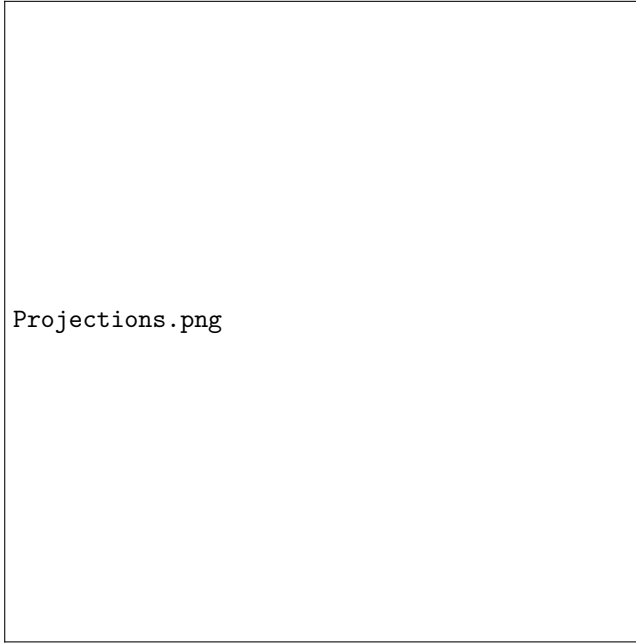


Figure 1: Illustration of our trajectory-based approach. The original sequence of depth maps is projected onto three orthogonal planes to form intensity videos. After that, the dense trajectory motion features are calculated for each representation.

trajectories in depth video. To perform this requirement, a straightforward method is to consider depth value as intensity value and adapts extracting dense trajectories on 2D transformed videos. However, unlike RGB information, depth information is textureless and more unstable. Besides, depth noise usually appears in any environment due to several reasons, such as systematic errors and non-systematic errors, as described in [28]. Thus, applying the dense trajectory-based approach always deals with confused cases of action classes. 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 view 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 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 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 ap-
70 proach 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 recog-
nizing complicated motions, the salient point-based approach deals with several
challenges, due to the lack of relationship of salient points. In recent studies [13–
75 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.

80 For the first direction, Yang.X et al. [24] 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)
[29] are computed from the DMM to represent an action video. Another ap-
85 proach proposed by Xia.L and Aggarwal.J.K [26] presents a filtering method to
extract spatio-temporal interest points from depth videos (DSTIPs). In this ap-
proach, 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
90 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.

95 For the second direction, [20] 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 [20], [21, 22, 25] use occupancy patterns to represent features in action videos.

100 Vieira et al. [22] 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. [25] presented semi-local features called Random
105 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. [21] designed a feature to describe the local
110 “depth appearance” for each 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
115 actionlet ensemble model to represent each action video.

Recently, Oreifej and Liu [27] 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.
120 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.

125 Inspired by results of Shotton et al. [30] and Xia.L et al. [31], the work by Yang et al. [23] developed the EigenJoints features based on skeleton in-

formation from RGBD sequences. The features contain three feature channels: posture, motion and offset. The posture and motion features represent spatial and temporal information, respectively. The offset features encode the difference between a pose with the initial pose in assumption that the initial pose is neutral. The three channels, then, are normalized and reduced by applying PCA method to obtain the EigenJoints descriptor.

Different from the previous approaches, we use a dense trajectory-based approach for action recognition. We do not care to segment human body like [20, 24]. Besides, skeleton extraction used in [21, 23] is not also required in our work. We only investigate the benefit of generating intensity representations from depth data, as mentioned in [20, 24]. Moreover, we leverage the effectiveness of trajectory feature to represent an action video. In our best knowledge, no method 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 motion feature 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. Our dense trajectory-based approach 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
155 methods, such as [13, 14] used KLT tracker [32], 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
160 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
165 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:

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

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 [33] is adopted to compute the dense optical flow. And to avoid a drifting problem, a
170 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 trajectory-aligned descriptors including the HOG [29], the
175 Histogram of Optical Flow (HOF) [9], and the Motion Boundary Histogram (MBH) [34]. 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
180 n_τ segments). The default settings of these parameters are $N = 32$ pixels, $L =$

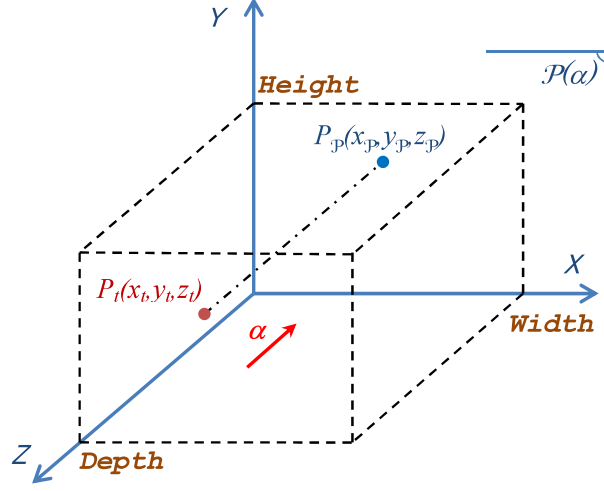


Figure 2: An illustration of the projection. Point $P_{\mathcal{P}}$ is the projection of point P_t along a view direction α onto a view plane $\mathcal{P}(\alpha)$.

15 frames, $n_{\sigma} = 2$, and $n_{\tau} = 3$.

According to the authors [9, 16, 35, 36], 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
 185 our experiment on all the three descriptors when compared to the depth-based state-of-the-art methods.

3.2. Proposed Method to Adapt Dense Trajectory-based Approach for Depth Video

Our proposed method to adapt the dense trajectory-based approach for hu-
 190 man action recognition in 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 $\mathcal{P}(\alpha) : ax + by + cz + d = 0$, 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:

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

195 where,

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

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

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

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

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

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
 200 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
 205 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:

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

210

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

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

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

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

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

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

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
220 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
225 2D transformed video $\mathcal{R}(\alpha_i)$, $i = \overline{1..N}$, corresponding feature representation $F(\alpha_i) = (b_{\alpha_i}^1, b_{\alpha_i}^2, \dots, b_{\alpha_i}^K)$ 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 uni-
230 modal features before learning, then, is used to generate feature representation $\mathcal{F} = (F(\alpha_1), \dots, F(\alpha_N))$ 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. Be-
235 sides, we perform the one-vs-all strategy for multi-class classification.

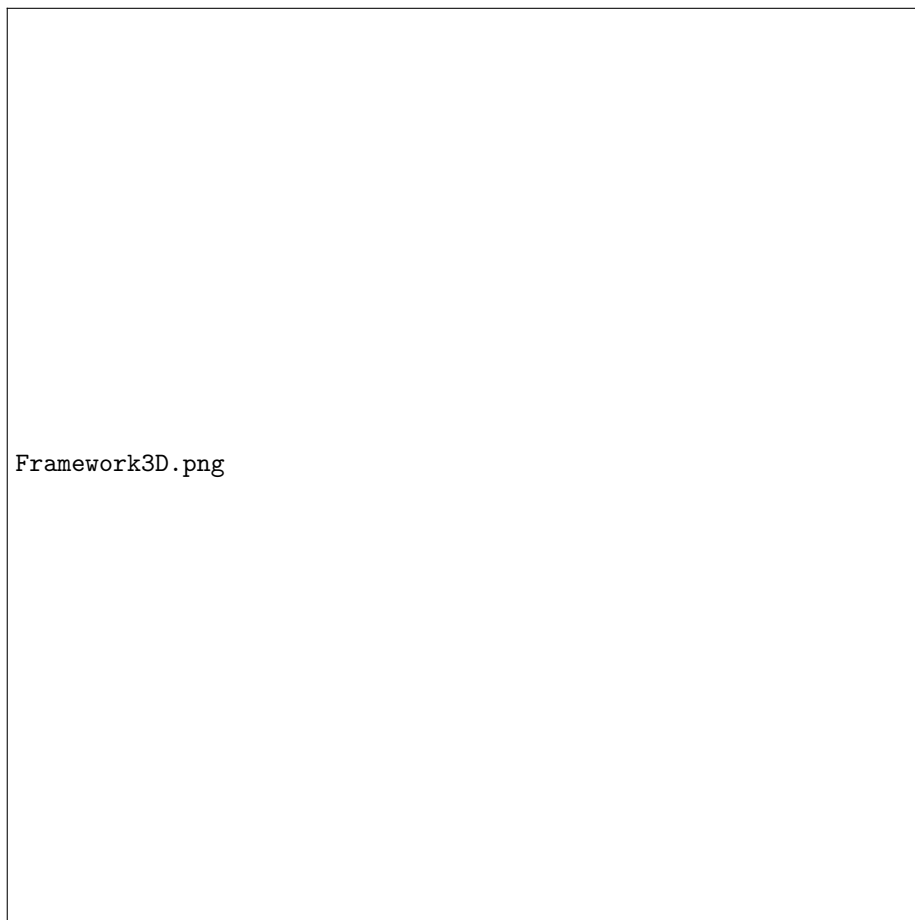


Figure 3: Our Framework Overview

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 [21, 23]. In fact, extracting skeleton exactly is still an completely unsolved problem, due to the challenges, such as cluttered background,
240 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
245 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

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

250 In order to conduct a fair comparison, we use the same experimental settings as [20, 21, 23, 24, 26, 27]. 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
255 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).

4.2. Evaluation Framework

Figure 3 shows our evaluation framework for the trajectory-based features.
260 We perform experiments using the proposed approach and compare with the state-of-the-art methods on depth data. We use the application available on-line¹ to extract dense trajectories and aligned-descriptors. Experimental results reported in section 5 attach to the MBH descriptor. The HOG, HOF descrip-

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

tors will be mentioned in the section 6. To quantize a large number of features
265 obtained by densely sampling, the BoW model is applied. At first, in each in-
tensity 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.

In order to classify actions, in our implementation, we use the libSVM library
270 published online by author². We adopt the format requirements of the library
to synchronize the annotation and the data. For testing, 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.

5. Experimental Results

275 This section presents the experimental results for applying our proposed ap-
proach on MSR Action 3D dataset. All experimental results are reported under
the settings mentioned in section 4.1. Besides, in comparison with the state-of-
the-art methods, our reported result is calculated on front representation only.
In addition, an evaluation related to selecting compensation information from
280 other representations will be also mentioned. All the results are compared in
terms of recognition accuracy. The best performance is highlighted in bold.

5.1. Recognize Actions from An Intensity Representation

Table 3 shows evaluation results of our trajectory-based approach and the
state-of-the-art methods in terms of average accuracy on three action subsets of
285 MSR Action 3D dataset (Table 2). The compared methods are based on various
feature representations, such as silhouette features [20, 24], skeletal joint features
like [21, 23], local occupancy patterns [22, 25], normal orientation features [27]

²<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Method	Accuracy (%)
Bag of 3D Points [20]	74.70
STOP [22]	84.80
EigenJoints [23]	82.33
Random Occupancy Patterns [25]	86.50
Local Occupancy Patterns [21]	88.20
Depth Motion Maps-based HOG [24]	91.63
Histogram of Oriented 4D Normals [27]	88.89
Depth Cuboid Similarity Feature [26]	89.30
Ours	94.53

Table 3: Comparison of accuracy on MSR Action 3D dataset. Notice that experimental results reported in this table is based on front representation only. Besides, we also use MBH descriptor only to calculate trajectory features.

and cuboid similarity features [26]. Interestingly, under the same setting (i.e cross subject test), the result table indicates that our approach beats all of
290 them. Besides, the results also show that there is significant difference of the performance between our method and the rest.

Consider the accuracy results on action subsets reported in Table 4, we found that two subsets AS1, AS2 contain many confused actions. Considering confusion matrices as showed in Table 5, some action-pairs are confused, such
295 as: *hammer* (a03) and *forward punch* (a05) in AS1, or *side-boxing* (a12) and *hand catch* (a04) in AS2. When analyzing the confused actions, we found that the main cause is due to similar motions of actions in the same view direction (i.e front representation). Besides, since depth data is textureless, it makes recognizing more difficult. That is reason why we need compensate information
300 from other intensity representations.

Action Subsets		
AS1	AS2	AS3
92.45	92.04	99.11

Table 4: Accuracy of our method on three action subsets.

5.2. *Compensate Motion Information from Other Representations*

In this part, we conduct experiments based on compensating information from other representations (i.e side and top) for the front representation. We report the experimental results on three action subsets and the average of the three subsets. Figure 4 shows a better view in comparing the fusion of front, side and top representations with the separate representations. Expectedly, the average recognition accuracy of the fusion, which is 96.67% accuracy, is better than the average recognition accuracy of the representations on the average of the three action subsets. Obviously, our proposed approach outperforms the mentioned state-of-the-art methods.

Besides, 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.

	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 5: Confusion matrices on three action subsets. Notice that action names are identified by indices of actions in table 1

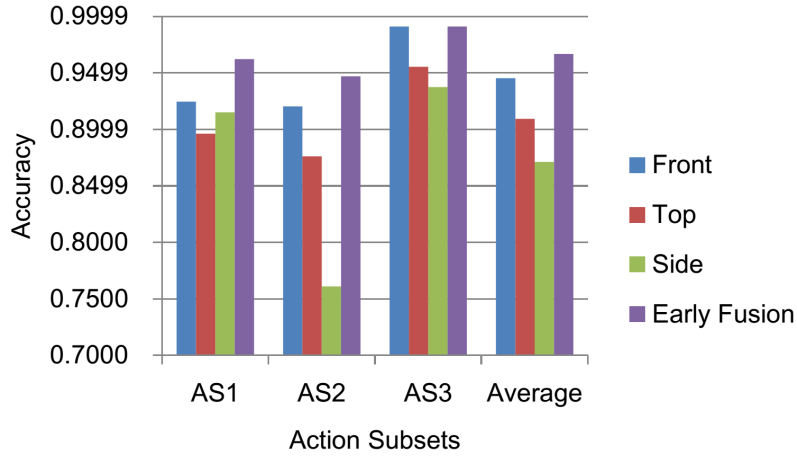


Figure 4: Comparison of recognition accuracy by using the early fusion scheme on intensity representations.

6. Discussions

6.1. The Impact of Our Method on Descriptors

For intensity data, according to [16] MBH is the best feature descriptor for
 320 dense trajectories. Therefore, in previous experiments, we only use MBH de-
 scriptor 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 exper-
 iments on these descriptors to answer this issue.

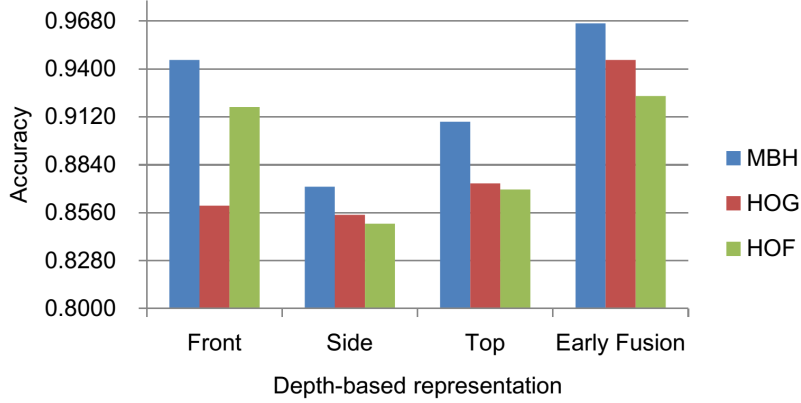


Figure 5: Comparison of recognition accuracy on trajectory-aligned descriptors.

325 In this part, we report the average recognition accuracies on the three de-
 scriptors and on the separate representations (i.e. front, side and top) as well
 as the fusion of the three representations. 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
 330 improved. The results indicate that the performances of HOG and HOF, re-
 spectively 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
 335 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,
 340 we perform experiments on representation combinations with front representa-
 tion. Experimental results are reported in figure 6. In this experiment, the

recognition accuracies of combinations are calculated on each intensity representations and the fusion.

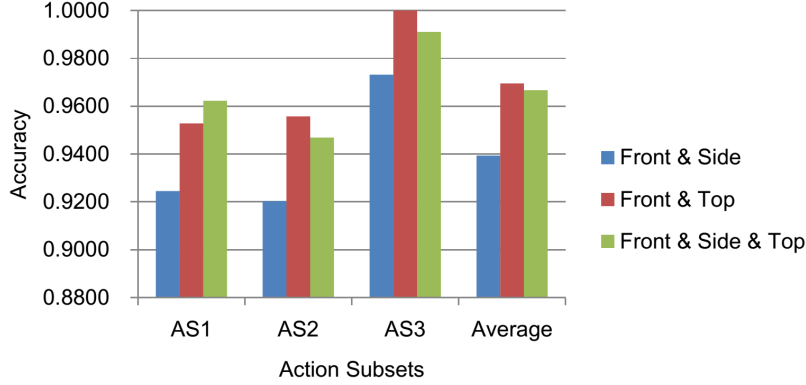


Figure 6: Comparison of recognition accuracy on combinations of intensity representations.

In order to conduct the experiments, we create combinations: front and side,
 345 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 perfor-
 mance, 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
 350 but still ensures a convincing performance.

6.3. MSR Daily Activity 3D Dataset

The MSR Daily Activity 3D dataset is proposed by [21], which includes 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
 355 different distances to the camera. Table 6 shows a comparison of recognition
 accuracies 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 and use MBH descriptor to describe motion feature. In

addition, we follow the experimental settings as described in [21]. In condition
 360 of only using depth data, [21, 26, 27] report a unexpected performance. In [26],
 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 [26] and evaluate
 on original MSR Daily Activity 3D dataset.

Method	Accuracy
LOP [21]	42.5
HON4D [27]	52
DSTIP&DCSF [26]	56.88
Ours	62.5

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

Although our method outperforms all the state-of-the-art methods, it is not
 365 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*
 370 *a paper* and *using a laptop*, 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 are
 375 small and textureless, so, it is very difficult 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 interactive
 objects must be necessary.



(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: Some sample actions on MSR Daily Activity 3D dataset.

380 6.4. Early versus Late Fusion in Our Approach

In terms of the fusion, [37] provided an interesting work. In this work, authors evaluated semantic concepts on two fusion schemes: early fusion and late fusion. They conducted experiments on the 2004 TRECVID benchmark dataset for visual modality and textual modality. Results indicated that the performance of the late fusion scheme is better than the performance of the early fusion scheme for most concepts. This evaluation are also applied for several multimodal-based analysis systems. However, the conclusion is reasonable or not for our approach, when considering each intensity representation as a modality. In order to answer this issue, we perform similar experiments on the late fusion scheme. In the experimental settings, we use the MBH descriptor to represent motion features and work on representation combinations: (front and side), (front and top) and (front, side and top). Experimental results in comparison between the early fusion scheme and the late fusion scheme are showed in figure 8.

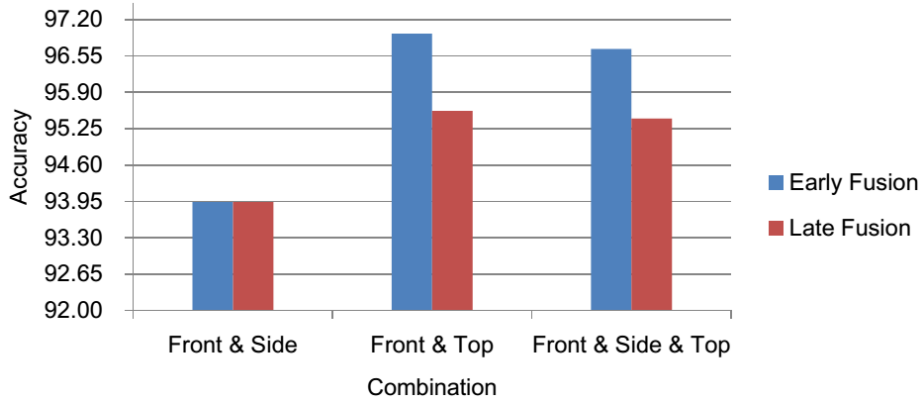


Figure 8: Comparison of recognition accuracy on the early and late fusion schemes.

395 Figure 8 indicates that both of the fusion schemes obtain significant improvements. However, the early fusion scheme gets better performances. Actually, we know that disadvantage of the early fusion approach is the difficulty to create

a good feature representation, due to the semantic difference of modalities. To deal with this challenge, the late fusion approach is used to convert the representations into the same type of semantics (i.e. probability score). In our approach, due to the similarity of semantics between modalities (i.e. features to represent motion information), the performance of the early fusion approach will tend to be better than the one of the late fusion approach. Besides, the achieved results from combinations confirm again that selecting representations to merge motion information is not a trivial task.

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 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-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.

- 450 [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. Marszaek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: *Proceedings of the British Machine Vision Conference (BMVC08)*, Leeds, United Kingdom, September 2008, 2008, pp. 995–1004.
- 455 [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.
- 460 [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.
- 465 [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.
- 470 [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.
- 475

- [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: Kins, and med, in: TRECVID workshop, 2012.
- 480 [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.
- 485 [20] 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.
- [21] 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.
- 490 [22] 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.
- 495 [23] 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.
- 500 [24] 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.
- [25] 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.

- [26] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for
505 activity recognition using depth camera, in: Computer Vision and Pattern
Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 2834–
2841.
- [27] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activ-
ity recognition from depth sequences, in: Computer Vision and Pattern
510 Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 716–723.
- [28] S. Foix, G. Alenya, C. Torras, Lock-in time-of-flight (tof) cameras: a survey,
Sensors Journal, IEEE 11 (9) (2011) 1917–1926.
- [29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection,
in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE
515 Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.
- [30] 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.
- [31] L. Xia, C.-C. Chen, J. Aggarwal, Human detection using depth informa-
520 tion by kinect, in: Computer Vision and Pattern Recognition Workshops
(CVPRW), 2011 IEEE Computer Society Conference on, IEEE, 2011, pp.
15–22.
- [32] 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.
- 525 [33] G. Farnebäck, Two-frame motion estimation based on polynomial expan-
sion, in: Image Analysis, Springer, 2003, pp. 363–370.
- [34] 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.

- 530 [35] 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.
- [36] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the
wild, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009.
535 IEEE Conference on, IEEE, 2009, pp. 1996–2003.
- [37] C. G. Snoek, M. Worring, A. W. Smeulders, Early versus late fusion in
semantic video analysis, in: Proceedings of the 13th annual ACM interna-
tional conference on Multimedia, ACM, 2005, pp. 399–402.