3D Trajectories for Action Recognition

Michal Koperski, Piotr Bilinski, François Bremond

3D TRAJECTORIES FOR ACTION RECOGNITION

Michal Koperski, Piotr Bilinski, Francois Bremond

INRIA Sophia Antipolis, STARS team
2004 Route des Lucioles, BP93, 06902 Sophia Antipolis, France
{Michal.Koperski, Piotr.Bilinski, Francois.Bremond}@inria.fr

ABSTRACT

Recent development in affordable depth sensors opens new possibilities in action recognition problem. Depth information improves skeleton detection, therefore many authors focused on analyzing pose for action recognition. But still skeleton detection is not robust and fail in more challenging scenarios, where sensor is placed outside of optimal working range and serious occlusions occur. In this paper we investigate state-of-the-art methods designed for RGB videos, which have proved their performance. Then we extend current state-of-the-art algorithms to benefit from depth information without need of skeleton detection. In this paper we propose two novel video descriptors. First combines motion and 3D information. Second improves performance on actions with low movement rate. We validate our approach on challenging MSR DailyActivty3D dataset.

Index Terms— Computer Vision, Action Recognition

1. INTRODUCTION

Human action recognition has been an active research topic for many years. It has also found applications such as: video surveillance, video data indexing, patient monitoring, human-computer interface. Apart from many contributions in last years, action recognition is still a challenging task. The major problem is to find model which would be discriminative, but still flexible enough to handle intraclass variation, because same actions can be done in different way.

Recent studies have shown that methods based on local space-time approach (such as trajectories) and Bag-of-Words, reached high accuracy rate. These methods model motion by detecting points of interest on each frame and then compute trajectories by tracking them in time space. However this methods fail to recognize similar actions, as they ignore spatial relationship between features. This problem has been addressed by Bilinski *et al.* [1]. They use head as reference point and compute relative position of trajectories according to head position. Even though this method improves action recognition accuracy, it still has problems using 2D information to distinguish actions which are performed in depth plane.

Recently, thanks to rapid development in cost effective depth sensors it is feasible to capture real-time depth information. Compared to conventional RGB cameras, the depth camera has several advantages: depth images are insensitive to changes in lighting conditions, moreover depth information simplifies task of object segmentation, which eases human skeleton detection. Recently many authors [2], [3] have focused on action recognition using pose detection. Unfortunately skeleton detection is not robust when occlusions occur and it can lead to inaccurate or even missing detentions. What is more, direct employing state-of-the-art local interest points detectors on depth map (instead of skeleton detector) is not feasible due to high noise of depth map comparing to RGB sequences [3]. Recently many dedicated depth map descriptors have been proposed [3], [4], [5], but they are still limited in terms of high order search space and sensitivity to depth map noise or missing measurements.

To address the above limitations of methods based on RGB sequences and depth sequences we propose a novel method which can benefit from both data sources. As detection of interest points is easier on RGB sequences we propose to extend Wang *et al.* [6] "Dense trajectories" method by adding depth information to each trajectory point. Such approach improves detection accuracy of actions mainly performed in depth plane. To model spatial relationship we employ [1] "Relative trajectories". We propose to use head as center of dynamic coordinate system. To improve discriminative power of such descriptor the positions of head and trajectories is given in 3D (x, y, z) space.

Descriptors based on trajectory detection require a certain amount of movement in processed video, because all computed features depend directly on move. In case where given action needs only little movement or action is occluded, proposed descriptors fail. To overcome this issue we propose a novel descriptor which combines features computed on RGB sequences and depth map. We use SURF key point detector [7] on RGB sequences and on each detected SURF point we compute Local Depth Pattern based on depth map.

We evaluate our approach on challenging MSR DailyActivity3D dataset. The experiments shows that our approach improves action recognition performance without need of skeleton detection.

The contribution of this paper are summarized as follows:

- We propose a new way to combine motion and depth information.
- We improve discriminative power of dense trajectories by adding depth information.
- We use relative trajectories to encode spatial information of features.
- We propose a novel descriptor for actions with low movement rate.
- Our approach does not need skeleton detection. Which is a very important feature as in real life video surveillance environment, reliable skeleton detection is not feasible due to camera viewpoint angle or high depth map noise.

2. OUR APPROACH

To extract trajectories information we employ (similarly to [6]) dense trajectories sampled on RGB sequences. For each frame we sample feature points with step W pixels. Such extracted points are tracked using optical flow and median filter kernel. We limit length of trajectory to L frames to avoid drifting problem. If trajectory exceeds length L we remove it from tracking process. To assure dense coverage, in situation when we detect that there are no tracked points in $W \times W$ area, we sample a new point from this area and add it to tracking process.

2.1. Trajectory Shape Descriptor (TSD)

Local motion patterns can be encoded by the shape of trajectory. We describe shape of the trajectory of length L as sequence $S = (\Delta P_t, \Delta P_{(t+1)}, ..., \Delta P_{t+L-1})$, where ΔP_t is a displacement vector $\Delta P_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. Vector S is normalized by sum of the magnitudes of the displacement vectors. Thus for each video v we obtain descriptor set:

$$\Omega_v = \{S_1, S_2, ..., S_{N_v}\} \tag{1}$$

where N_v is number of trajectories detected in video v.

2.2. 3D Trajectory Shape Descriptor (3DTSD)

To improve discriminative power of Trajectory Shape Descriptor we add depth information z to each trajectory point $p = (x, y)$. The z value is computed as mean of area of size N around point p. Such approach allows to estimate depth value of missing measurement points and to filter out noise which is an issue in depth map. It is a very important step as displacement vector cannot be computed for points which do not have measured depth. In addition, trajectories are very often detected at the edges, where depth missing values are most likely to present. Before computing displacements, trajectories which contain at least one point with missing depth value are being removed. For remaining trajectories 3D Trajectory Shape Descriptor is computed in the analogical way to Trajectory Shape Descriptor. Thus for each video v we obtain descriptor set:

$$\Psi_v = \{S_1, S_2, ..., S_{N_v}\} \tag{2}$$

where N_v is number of trajectories which remained in the video v.

2.3. Relative Trajectory Descriptor (RTD)

Both Trajectory Shape Descriptor and 3D Trajectory Shape Descriptor encodes only displacement information ignoring important spatial position of trajectory. A common way to resolve this issue is to use either spatio-temporal grids or multi-scale pyramids. But those methods provide only coarse information. To resolve this problem similarly to [1] we employ relative trajectories. We use detected head as a center of dynamic coordinate system. Such descriptor encodes both shape characteristics and spatial position, this approach helps to distinguish similar trajectories detected at different positions. Given trajectory $t_i = [(x_j, y_j), ..., (x_{j+L}, y_{j+L})]$ where L is the length of trajectory and head trajectory $t_h = [(x'_j, y'_j), ..., (x'_{j+L}, y'_{j+L})]$. We can define relative trajectory as:

$$R = [(x_j, y_j, ..., x_{j+L}, y_{j+L}) - (x'_j, y'_j, ..., x'_{j+L}, y'_{j+L})]. \tag{3}$$

Thus for each video v we obtain descriptor set:

$$\Phi_v = \{R_1, R_2, ..., R_{N_v}\} \tag{4}$$

where N_v is a number of trajectories detected in video v.

RTD descriptor may be combined with other non-relative descriptors. Such solution would allow action recognition even if head detection is missing.

2.4. 3D Relative Trajectory Descriptor (3DRTD)

To improve discriminative power of Relative Trajectory Descriptor we add depth information to each trajectory point and each head trajectory point. We do it in same way as in 2.2. Thus for each video v we obtain descriptor set:

$$\Pi_v = \{RD_1, RD_2, ..., RD_{N_v}\} \tag{5}$$

where N_v is number of trajectories extracted from the video v.

2.5. Trajectory Appearance Descriptors

Descriptors proposed in sections: 2.1, 2.2, 2.3, 2.4 provide only trajectory shape information. To empower the motion

information of dense trajectories, similarly to [6] we compute HOG, HOF, MBH (motion boundary histogram) in spatio-temporal volume around each trajectory point. The volume size is $N \times N$ pixels and L frames long.

2.6. SURF Key Points Appearance Descriptor

Descriptors proposed in previous sections rely on trajectory detection. This implies requirement that there must be certain amount of movement in the processed video to compute enough discriminating features. In case where given action needs only little movement or action is occluded proposed descriptors fail.

To overcome this issue we propose another descriptor which will catch appearance features even when there are no trajectories detected. To achieve this goal we propose to use SURF key point detector [7]. On each sampled frame SURF key points are detected inside bounding box of detected person. For each detected key point SURF descriptor is computed based on RGB appearance.

2.7. Local Depth Pattern Descriptor

To improve discriminating power similarly to [8] we compute Local Depth Pattern in the neighborhood of detected SURF points. At given frame f for each detected SURF point p we divide space around p into $N_x \times N_y$ cells. Each cell has size of $S_x \times S_y$ pixels. Then we compute average depth value for each cell. To create the feature vector we compute difference between each cell pair.

2.8. Action recognition

We use standard bag-of-words approach. First we construct codebook for each descriptor using k-means algorithm. We empirically set the size of codebook to 4000 words. Descriptors are then assigned to the closest word from codebook in means of Euclidean distance. For each video histogram of occurrences of codebook words is computed. Which is then normalized by L1 norm. For classification we use a non-linear SVM with χ^2-kernel.

As mentioned before trajectory based descriptors fail when there is no enough motion, due to either characteristics of action or occlusions. To distinguish actions with high and low movement and benefit from both trajectories descriptor and key points descriptor we use hierarchical framework (see figure 1). First we compute descriptor which combines Trajectory Shape Descriptor (TSD) and 3D Trajectory Shape Descriptor (3DTSD) and train classifier which distinguish actions with high movement from actions with low movement rate. Then for actions with low movement rate we combine SURF Key Points Descriptor with Local Depth Pattern Descriptor and then train classifier using described bag-of-words approach. For high movement rate actions we re-use descriptor from top hierarchy level and train classifier.

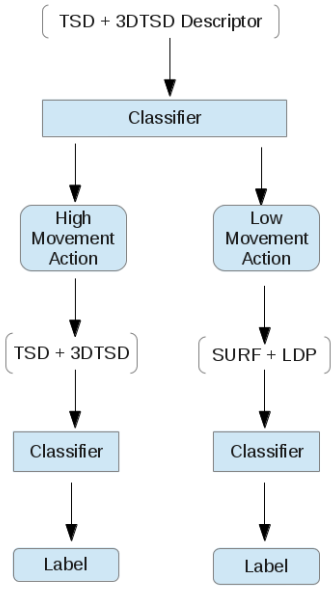

Fig. 1. Classification framework

3. EXPERIMENTS

In this section we evaluate the performance of our approach on challenging MSRDailyActivity3D dataset. We compare the results to state-of-the-art methods.

3.1. MSRDailyActivity3D

In this section we evaluate our approach on MSRDailyActivity3D [2] dataset. The data set consists of 16 actions performed by 10 subjects. Each action is performed in standing and sitting position which brings additional intraclass variation. We use Leave One Subject Out setup, where in each split one person is selected for testing and training is performed on the remaining subjects. In this case we have 10 splits. For computing 2D dense trajectories, HOG, HOF and MBH features we use LEAR's implementation [1]. The following parameters values has been: $W = 5$, $L = 15$, $N = 32$, as they gave the best experimental results.

We have selected the actions such: *write on paper, use laptop, sit still* as low movement actions, due to low number of detected trajectories (see figure 3). *Play game, play guitar* are also considered as low movement actions as most trajectories are detected on head or due to whole body move which is not discriminative for those actions. In *play game* action subject plays on pad involving small fingers movement, the same case is with *play guitar* action. The remaining actions such as *drink, eat, read book, call phone, use vacuum,*

[1] `http://lear.inrialpes.fr/people/wang/dense_trajectories` (Second version)

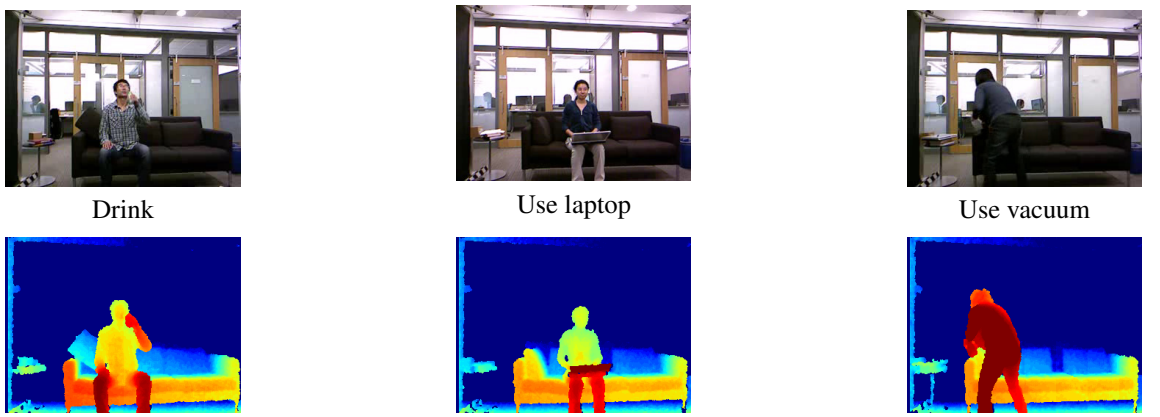

Fig. 2. Sample frames from MSR DailyActivity3D dataset

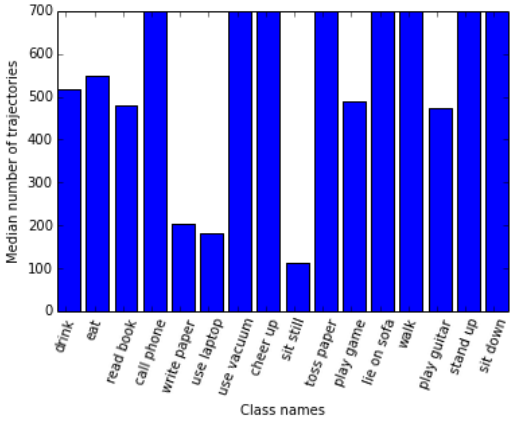

Fig. 3. Median of number of detected trajectories per class.

cheer up, toss paper, lie down on sofa, walk, stand up, sit down are considered as high movement actions. The top level classier (see figure 1) achieved 90% of accuracy. In table 2 we provide comparison of different descriptors combination performance for high movement actions. In table 1 we provide performance comparison for low movement actions.

The final results are in table 3. Note that methods printed in italic require skeleton detection which is not the case in our approach. It is a very important feature as in real life video surveillance environment reliable skeleton detection is not feasible due to camera viewpoint angle or high depth map noise (as people often stay outside of optimal sensor range).

Acknowledgments

This work is supported by Dem@care and SafEE projects.

Method	Accuracy
SURF	0.57
Local Depth Pattern (LDP)	0.56
SURF + LDP	**0.60**

Table 1. Low movement action classification - descriptors comparison for DailyActivity3D datset.

Method	Accuracy
Trajectory Shape Desc (TSD)	0.78
3D Trajectory Shape Desc (3DTSD)	0.74
TSD + 3DTSD	**0.85**
TSD + Relative Trajectory Descriptor (RTD)	0.83
3DTSD + 3D Relative Trajectory Descriptor (3DRTD)	0.83
HOG	0.68
HOF	0.80
MBH	0.84

Table 2. High movement actions classification - descriptors comparison for DailyActivity3D datset.

Method	Accuracy
Dynamic Temporal Warping [9]	*0.54*
HON4D [3]	*0.80*
Actionlet Ensemble [2]	*0.85*
TSD (without hierarchical classifier)	0.58
3DTSD (without hierarchical classifier)	0.55
TSD + RTD (without hierarchical classifier)	0.65
TSD + 3DTSD (without hierarchical classifier)	0.63
Our Approach (with hierarchical classifier)	**0.72**

Table 3. Recognition Accuracy Comparison for DailyActivity3D datset. Methods in italic require full skeleton detection.

4. REFERENCES

[1] Piotr Bilinski, Etienne Corvee, Slawomir Bak, and Francois Bremond, "Relative dense tracklets for human action recognition," in *10th IEEE International Conference on Automatic Face and Gesture Recognition*, Shanghai, China, Apr. 2013, pp. 1–7.

[2] Ying Wu, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, Jun 2012, pp. 1290–1297.

[3] Omar Oreifej and Zicheng Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, Portland, USA, Jun 2013, pp. 716–723.

[4] Simon Hadfield and Richard Bowden, "Hollywood 3d: Recognizing actions in 3d natural scenes," in *Proceeedings, conference on Computer Vision and Pattern Recognition*, Portland, Oregon, Jun 2013, pp. 3398–3405.

[5] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *Proc. IEEE International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, San Francisco, USA, Jun 2010, pp. 9–14.

[6] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun 2011, pp. 3169–3176.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008.

[8] Yang Zhao, Zicheng Liu, Lu Yang, and Hong Cheng, "Combing rgb and depth map features for human activity recognition," in *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 2012, pp. 1–4.

[9] Meinard Müller and Tido Röder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Aire-la-Ville, Switzerland, Switzerland, 2006, pp. 137–146.