ACTION RECOGNITION USING BAG OF FEATURES EXTRACTED FROM A BEAM OF TRAJECTORIES

Thanh Phuong Nguyen, Antoine Manzanera

ENSTA Paristech, Boulevard des Maréchaux, 91762 Palaiseau, France
{thanh-phuong.nguyen,antoine.manzanera}@ensta-paristech.fr

ABSTRACT

A new spatio temporal descriptor is proposed for action recognition. The action is modelled from a beam of trajectories obtained using semi dense point tracking on the video sequence. We detect the dominant points of these trajectories as points of local extremum curvature and extract their corresponding feature vectors, to form a dictionary of atomic action elements. The high density of these informative and invariant elements allows effective statistical action description. Then, human action recognition is performed using a bag of feature model with SVM classifier. Experimentations show promising results on several well-known datasets.

Index Terms— action recognition, spatio temporal feature, bag of features, dominant point, semi dense point tracking, trajectory beam

1. INTRODUCTION

Action recognition in videos has been a very active research domain in computer vision for recent years. Many approaches have been introduced, but a reliable application in real conditions is still a big challenge.

One popular approach for action recognition consists in analyzing local spatio temporal interest point features. These features are designed to be invariant to certain geometric transforms, for robustness purposes with respect to scale, rotation, or view point. Laptev [1] detected space-time interest points in videos extending Harris corner criteria from 2D images to 3D. Similarly, Dollár [2] extracted spatio temporal keypoints in the energy map referred to as cuboids by performing symmetric temporal Gabor filtering. It avoids problem of sparse corner detection reported in [1]. Willems [3] performed an extension of the Hessian saliency measure to detect dense and scale-invariant spatial temporal interest points. Chakraborty [4] proposed a surround suppression of detected interest points combined with spatial and temporal constraints to be robust with respect to camera motion and background cluster.

In our work, the motion in videos is described based on the semi dense point tracking, which produces a beam of particles trajectories. We propose to use dominant point extraction and quantization to characterize local motion in each trajectory. A bag of feature model is then applied to model the action. The classification is done by using a SVM classifier. We present different experimentations, first on a classic dataset (KTH) and next on a recent and more challenging dataset (UCF Youtube).

The rest of this paper is organized as follows. Section 2 describes in brief the semi dense point tracking [5] used in our approach. Section 3 presents the extraction of dominant points from the beam of trajectories. Section 4 details the action descriptor based on bag of features approach. Several experiments are shown in Section 5.

2. MOTION REPRESENTATION USING A BEAM OF TRAJECTORIES

Trajectories are compact and rich information source to represent motion in videos, and have been used already for action recognition [6]. Generally, to obtain reliable trajectories, the spatial information is dramatically reduced to a small number of keypoints, and then it may be hazardous to compute statistics on the set of trajectories. In this work we use the semi dense point tracking method [5] which is a trade-off between long term tracking and dense optical flow, and allows the tracking of a high number of weak keypoints in a video in real time, thanks to its high level of parallelism.

It contains two main steps. First, weak keypoints are detected from a saliency function whose purpose is to discard only the points whose matching will necessarily be ambiguous at all scales. Each keypoint is characterized by a short feature vector (16 dimensions) calculated from Bresenham circles at two different radii (3 and 6). Then, the keypoints are tracked using multiscale search algorithm based on a prediction combining the estimated velocities of each particle and the dominant (global) acceleration. Using GPU implementation, this method can handle 10 000 points per frame at 55 frames/s on 640×480 videos. In addition, it is robust to sudden camera motion change.

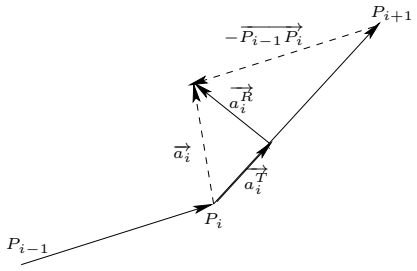


Fig. 1. Tangential and radial acceleration of a point.

3. TRAJECTORY FEATURE POINTS

As shown by Attneave [7], the local maxima of the curvature function contain rich information about the shape. Thus, many approaches for dominant point detection are based on curvature. In our approach, the dominant points are extracted from the beam of trajectories in the 2d+t space, corresponding to the positions of tracked particles in the different frames.

3.1. Dominant point detection

In the 2D+t space, the curvature is physically related to radial acceleration. Each trajectory is smoothed by a Gaussian kernel of parameter σ. Suppose that the location of point P in image at frame i is P_i. So, $\overrightarrow{P_{i-1}P_i}$ (resp. $\overrightarrow{P_iP_{i+1}}$) is the velocity vector of P at frame $i-1$ (resp. i). The acceleration $\overrightarrow{a_i} = \overrightarrow{P_iP_{i+1}} - \overrightarrow{P_{i-1}P_i}$ is decomposed in two orthogonal components: the tangential acceleration $\overrightarrow{a_i^T}$, in the direction of the displacement, and the radial acceleration $\overrightarrow{a_i^R}$, whose norm corresponds to the curvature multiplied by the squared velocity norm (see Figure 1).

The dominant points are defined as the local maxima of the radial acceleration. To eliminate non-significant dominant points, a threshold value is used, set to 0.25 pixel/frame^2 in our implementation.

3.2. Description of feature point

For each detected point P, a description vector is constructed using information extracted from a portion of the point trajectory centered on P. Let E_1 and E_2 be the positions of P, respectively w_σ frames before and w_σ frames after, with $w_\sigma = \lfloor 6\sigma \rfloor$ (see Figure 2).

First, the radial acceleration of P is considered. Second, the mean and standard deviation of speeds, tangent and radial accelerations of P in its successive positions from E_1 to P are recorded, and also from P to E_2. Then, the come in and go out directions ($\overrightarrow{E_1P}$ and $\overrightarrow{PE_2}$), and so the angle $\widehat{E_1PE_2}$ are recorded. Finally, a feature vector of 14 dimensions is constructed for each dominant point.

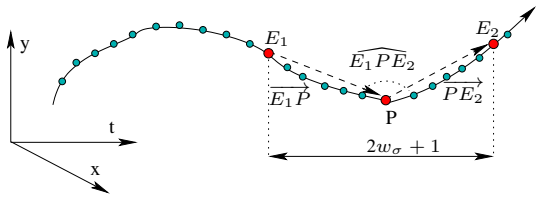
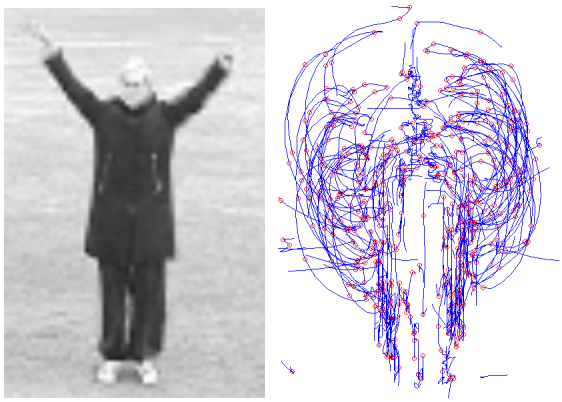


Fig. 2. Extraction of feature vector using a portion of trajectory centered on a dominant point P.



(a) (b)

Fig. 3. (a) A hand waving action; (b) The motion is modelled as a beam of trajectories (in blue), from which the dominant points (in red) are extracted.

3.3. Multiscale approach

Our approach is multiscale thanks to the use of different Gaussian kernels in the smoothing of trajectories. In our implementation, we apply 5 different scales (σ from 1.0 to 5.0). The detected dominant points are different from one scale to the other, so the dimension of descriptors do not change but every dominant point has a characteristic scale. Finally, the set of dominant points form a multiscale outline of motion elements, since it contains geometric and physical features at different scales.

3.4. Properties of dominant points

The use of the such dominant points as atomic elements for action representation can be justified as follows:

Rich information about action. The proposed descriptor captures information on the velocity and acceleration of the movement around salient points: it represents elementary motion elements that may be used as a base to describe complex actions.

Robustness to geometric transforms. The detection of dominant points is invariant to geometric transforms such as rotation, scaling or translation. Hence the descriptor should not change much when the location and scale of the action

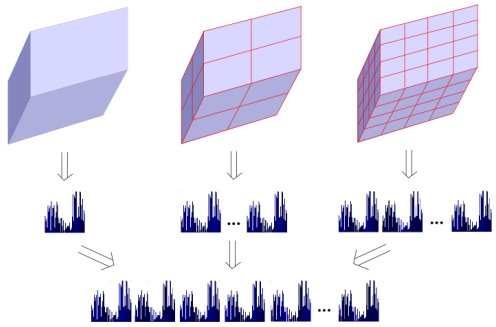


Fig. 4. Modelling of action using histogram concatenation.

change.

Robustness to appearance variation. By design of the weak keypoints detection, which is normalized by the contrast, the descriptor must be robust against illumination change and should not depend much on the appearance of the moving objects.

4. ACTION RECOGNITION USING A BAG OF FEATURES

We introduce a representation of action based on bag of features (BoF) model using dominant points of the beam of trajectories (see also Figure 3).

Vocabulary construction

In the first step, a visual vocabulary is constructed. We employ k-means clustering on the spatio temporal features of extracted dominant points for the training videos. Then each visual word corresponds to a characteristic motion element. Therefore, at the recognition step, each detected dominant point shall be assigned to the closest vocabulary word using Euclidean distance.

Action representation

In our approach, an action is represented by the histogram of visual words on a spatio temporal volume. The considered volume can be either the entire sequence or a set of subsequences defined by a spatio temporal grid. Then, all histograms are concatenated into one vector that is then normalized for action representation. Figure 4 shows how to construct the action description using three different grids.

Classification

We choose the SVM classifier [8] of Vedaldi et al. which approximates a large scale SVMs using an explicit feature map for the additive class of kernels to perform action classification. Generally, it is much faster than classical non linear SVMs and it can be used in large scale problems.

5. EXPERIMENTATIONS

We evaluate our descriptor on two well-known datasets. The first one (KTH) [9] is a classic dataset, used to evaluate many action recognition methods. The second one (UCF Youtube) [10] is a more realistic and challenging dataset.

Table 1. Confusion matrix on KTH dataset.

	Box.	Clap.	Wave	Jog.	Run.	Walk.
Boxing	100	0	0	0	0	0
Clapping	0	100	0	0	0	0
Waving	0	2.5	97.5	0	0	0
Jogging	0	0	0	95	2.5	2.5
Running	0	0	0	17.5	82.5	0
Walking	0	2.5	0	2.5	0	95

Table 2. Comparison on UCF Youtube.

Our method	[11]	[12]	[10] [1]	[10] [2]
65.1	64	64	65.4	71.2

5.1. Parameter settings

There are several parameters in our method. First, for the detection of dominant points, the threshold of radial acceleration value was set to 0.25 pixel/frame^2. Second, the number of visual words used in k-means clustering. It should depend on the variability of gestures and motion elements that may appear in the actions. It has been empirically set to 70 for KTH, and 1 000 for UCF Youtube. Third, there is the configuration of spatio-temporal grids used to construct the histograms of visual words. In our implementation, we used 3 grids: 1x1x1, 2x2x2 and 4x4x4.

5.2. Experimentation on KTH dataset

The dataset contains 25 people for 6 actions (running, walking, jogging, boxing, hand clapping and hand waving) in 4 different scenarios (indoors, outdoors, outdoors with scale change and outdoors with different clothes). Different people perform the same action at different orientations and speeds. In contains 599 videos, of which 399 are used for training, and the rest for testing. As designed by [9], the test set contains the actions of 9 people, and the training set corresponds to the 16 remaining persons. Table 1 shows the confusion matrix obtained by our method on the KTH dataset. The ground truth is read row by row. The average recognition rate is 95 % which is comparable to the state-of-the-art approaches. The main error factor comes from confusion between jogging and running, which is a typical problem in reported methods.

5.3. Experimentation on UCF Youtube dataset

The UCF Youtube dataset records 11 categories (basketball shooting, cycling, diving, golf swinging, horse back riding,

[1] Static features: SIFT+HAR+HES+MSER

[2] Static features+Dynamic features (Gaussian and Gabor filters+PCA)

Table 3. Comparison on KTH.

Ours	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
95	94.5	93.2	93.5	94.7	95.1	93.8	94.2	98.2

Table 4. Confusion matrix on UCF Youtube dataset. The columns (resp. rows) from left (resp. top) to right (resp. bottom): basketball, biking, diving, golf swing, horse riding, soccer juggling, swing, tennis swing, trampoline jumping, volleyball spiking, walking with dog.

46.2	0	9.6	1.9	0	1.9	0	17.3	5.8	17.3	0
0	51.9	3.7	0	18.5	0	0	7.4	0	0	18.5
0	0	73.3	6.7	0	1.7	5.0	3.3	5.0	3.3	1.7
0	0	4.0	82.0	0	0	2.0	12.0	0	0	0
1.2	2.3	0	0	91.9	0	0	1.2	2.3	1.2	0
0	3.5	5.3	0	5.3	66.7	14.0	0	5.3	0	0
2.2	0	2.2	6.7	15.6	2.2	57.8	0	6.7	2.2	4.4
5.1	1.7	1.7	13.6	8.5	6.8	0	59.3	0	1.7	1.7
0	0	2.2	0	2.2	0	6.7	0	82.2	0	6.7
10.0	0	10.0	5.0	2.5	0	0	12.5	0	60.0	0
0	15.2	4.3	4.3	28.3	8.7	6.5	4.3	4.3	0	23.9

soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking with a dog), and contains 1 600 video sequences. Each category is divided into 25 groups sharing common appearance properties (actors, background, or other). Following [10], we used 9 groups out of the 25 as test and the 16 remaining groups as training data. This dataset is much more challenging than KTH because of its large variability in terms of viewpoints, backgrounds and camera motions. Table 4 shows the confusion matrix obtained by our method. As it can be seen on Table 2, our mean recognition rate (65.1 %) is comparable to existing methods with a much shorter descriptor and lower computational cost.

6. CONCLUSIONS

We have proposed a new approach for action recognition based on features extracted from a beam of trajectories. Thanks to the high density of reliable trajectories, we can use statistical bag-of-features method using spatio temporal histograms of motion elements. Thanks to its low computational complexity, this method can be applied online for real recognition applications. In the future, we will also investigate dominant point selection process like in [4], to enhance the robustness of the descriptor to camera motion and background clutter. The combination with other features will also be the subject of future works.

Acknowledgement

This work is part of an ITEA2 project, and is supported by french Ministry of Economy (DGCIS).

7. REFERENCES

[1] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005, pp. 65–72.

[3] G. Willem, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*, 2008, LNCS, pp. 650–663.

[4] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzalez, and F. Xavier Roca, "A selective spatio-temporal interest point detector for human action recognition in complex scenes," in *ICCV*, 2011, pp. 1776–1783.

[5] M. Garrigues and A. Manzanera, "Real time semi-dense point tracking," in *ICIAR (1)*, Aurélio J. C. Campilho and Mohamed S. Kamel, Eds., 2012, vol. 7324 of *LNCS*, pp. 245–252.

[6] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011, pp. 3169–3176.

[7] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, no. 3, pp. 183–193, 1954.

[8] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *PAMI*, vol. 34, no. 3, pp. 480–492, 2012.

[9] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *ICPR*, 2004, pp. 32–36.

[10] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from video "in the wild"," in *CVPR*, 2009, pp. 1996–2003.

[11] Z. Lu, Y. Peng, and H. H. S. Ip, "Spectral learning of latent semantics for action recognition," in *ICCV*, 2011, pp. 1503–1510.

[12] M. Bregonzio, J. Li, S. Gong, and T. Xiang, "Discriminative topics modelling for action feature selection and recognition," in *BMVC*, 2010, pp. 1–11.

[13] M. J. Roshtkhari and M. D. Levine, "A multi-scale hierarchical codebook method for human action recognition in videos using a single example," in *CRV*, 2012, pp. 182–189.

[14] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *CVPR*, 2009, pp. 1948–1955.

[15] A. Yao, J. Gall, and L. J. Van Gool, "A hough transform-based voting framework for action recognition," in *CVPR*, 2010, pp. 2061–2068.

[16] T. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Integrating local action elements for action analysis," *CVIU*, vol. 116, pp. 378–395, 2011.

[17] H. J. Seo and P. Milanfar, "Action recognition from one example," *PAMI*, vol. 33, no. 5, pp. 867–882, 2011.

[18] J. Wang, Z. Chen, and Y. Wu, "Action recognition with multiscale spatio-temporal contexts," in *CVPR*, 2011, pp. 3185–3192.

[19] J. Liu and M. Shah, "Learning human actions via information maximization," in *CVPR*, 2008, pp. 1–8.

[20] S.Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *CVPR*, 2012, pp. 1234–1241.