Action recognition in videos using frequency analysis of critical point trajectories

Cyrille Beaudry, Renaud Péteri, Laurent Mascarilla

ACTION RECOGNITION IN VIDEOS USING FREQUENCY ANALYSIS OF CRITICAL POINT TRAJECTORIES

Cyrille Beaudry, Renaud Péteri, Laurent Mascarilla

MIA - Univ. La Rochelle
Avenue Michel Crépeau
17042 La Rochelle, France

ABSTRACT

This paper focuses on human action recognition in video sequences. A method based on the optical flow estimation is presented, where critical points of the flow field are extracted. Multi-scale trajectories are generated from those points and are characterized in the frequency domain. Finally, a sequence is described by fusing this frequency information with motion orientation and shape information. Experiments show that this method has recognition rates among the highest in the state of the art on the KTH dataset. Contrary to recent dense sampling strategies, the proposed method only requires critical points of motion flow field, thus permitting a lower computation time and a better sequence description. Results and perspectives are then discussed.

Index Terms— Action recognition in videos, critical points, frequency analysis of motion trajectories.

1. INTRODUCTION

Action recognition is an active field of research in computer vision. Laptev et Lindberg [3] were the first authors to propose a temporal extension of the 2D Harris-Laplace interest points detector. In [1], the authors proposed the *cuboïd* detector, where interest points are estimated from temporal Gabor filters and 2D spatial gaussian filters. A temporal extension of the 2D detection based on the Hessian operator for detecting relevant image blobs is proposed in [11]. The efficiency of densely sampling the spatio-temporal domain has been stressed in [10] for human action recognition. Several methods are making the choice of a uniformly dense point selection rather than a sparse estimation of interest points. The drawback of it is the increase in the computation complexity. In [9], authors extend the dense selection approach by estimating the trajectories of these points along the sequence. [5] and recently [8] use the characterization of these trajectories as a powerful discriminating factor. These recent approaches have shown the relevance of estimating point trajectories for action recognition.
This paper presents an approach based on the use of optical flow and critical point tracking at different scales. It aims at going beyond the concept of spatio-temporal points by considering their trajectories as a motion feature. These trajectories are described by their Fourier transform coefficients and are made invariant to changes in scale, rotation and translation. They are also robust to small motion perturbations.
This paper is divided as follows. Section 2 details the estimation of critical points and multi-scale trajectories. Section 3 exposes the descriptors used for critical points. The use of Fourier coefficients to characterize multi-scale trajectories and to combine frequency information with shape and motion is then presented. In section 4, experimental results on the KTH dataset, with a comparison between our approach and the state of the art are provided.

2. CRITICAL POINTS AND TRAJECTORIES

2.1. Critical points of a vector field

Critical points are extracted from a robust optical flow estimation method that uses a median filter at each iteration step [7].

The divergence and curl of the optical flow are first computed. Let a flow field $\mathbf{F} = (u_t, v_t)$ with u_t and v_t being the horizontal and vertical components of the flow. The curl and divergence of $\mathbf{F}$ are:

$$Rot(\mathbf{F}) = \nabla \wedge \mathbf{F} = \frac{\partial v_t}{\partial x} - \frac{\partial u_t}{\partial y}$$

$$Div(\mathbf{F}) = \nabla \cdot \mathbf{F} = \frac{\partial u_t}{\partial x} + \frac{\partial v_t}{\partial y}$$

Both values characterize the way a vector field evolves in time:
- the curl gives information on how a fluid may rotate locally.
- the divergence represents to what extent a small volume around a point is a source or a sink for the vector field. Points with high divergence or curl are typical of high local deformations of the flow field. We use these features to find local areas of potential movements of interest (Figure 1). Spatio-temporal interest points are then the extrema of curl and divergence, and correspond to some critical points of the estimated flow.

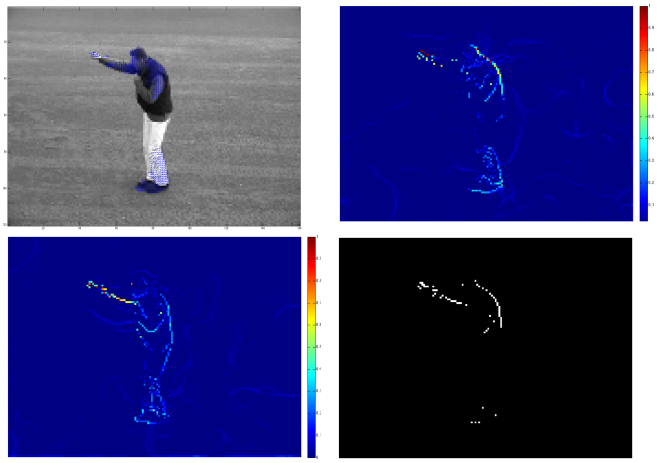

Fig. 1. From top left to bottom right: optical flow, curl value map, divergence value map, map of extrema. Estimated points correspond to actions performed by the subject.

2.2. Extraction and characterization of multi-scale trajectories

In order to analyse typical movements in videos, trajectories are estimated from critical points using the dense trajectory approach [9]. These points are tracked in the video by using a median filter on the optical flow. Given an optical flow field $\mathbf{F} = (u_t, v_t)$, position of a point $P_t = (x_t, y_t)$ at frame t is estimated at $t+1$ as the point $P_{t+1} = (x_{t+1}, y_{t+1})$ such that:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + Med_F(V_{(x_t,y_t)})$$

with Med_F, a spatial median filter applied on F at $V_{(x_t,y_t)}$ which is a neighborhood centered on P_t.

A spatio-temporal pyramidal approach has been choosen to analyse the different motion frequencies of extracted trajectories.

A spatio-temporal dyadic subdivision is first performed on sequences with a spatio-temporal gaussian kernel to suppress high frequencies. Optical flow is then estimated on these resulting sequences. Each sub-sequence corresponds to a scale of the pyramid. The dyadic subdivision allows to obtain trajectories with the same length but for different frequencies. This part is detailed thereafter.

The deformation of the flow can be related to a movement with a characteristic spatio-temporal scale. Trajectories extracted from this movement also have one or several characteristic frequencies, and this multi-scale approach enables to deal with larger frequency intervals (Figure 2).

Critical points are extracted from each scale of the pyramid and trajectories are computed from these points, so called "multi-scale trajectories". The size of trajectories is proportional to the length of the sequence and the scale of the pyramid. For a sequence of N frames, the size T_s of trajectories is computed such that:

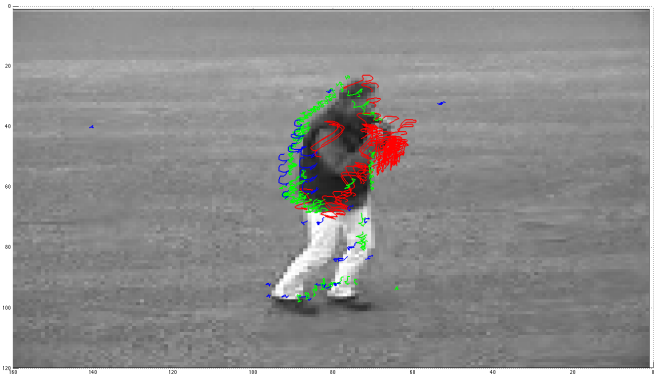

Fig. 2. Red trajectories are computed from high motion frequencies (fist), while green and blue trajectories are computed on motion with lower frequencies (legs).

$$T_s = l_1 \cdot (2^{s-1}) \cdot N$$

where s is the pyramid scale and l_1 that is a threshold empirically fixed.

When the size of a trajectory is larger than this threshold, the trajectory is automatically cut such that it satisfies a size criterion. If its size is smaller, it is removed. This condition allows to keep short trajectories and to avoid problem of drifting during the tracking. All extracted trajectories have the same size but correspond to different motion frequencies.

3. DESCRIPTOR COMPUTED FROM CRITICAL POINTS AND TRAJECTORIES.

3.1. The HOG/HOF descriptor

The descriptor used for critical points is the HOG/HOF descriptor [10]. It is based on shape information (histogram of 2D gradient orientations) and motion orientation (histogram of optical flow orientations). It is a classical descriptor that has been proven to be very efficient in computer vision.

3.2. Trajectory descriptors based on Fourier coefficients

Multi-scale trajectories obtained are described by their Fourier coefficients. A robust action recognition method should extract descriptors with low intra-class variablity by ensuring invariances to different kind of transformations. The choice of Fourier coefficients is motivated by invariances which are easy to obtain in the frequency domain.

Given a trajectory T_N with N sequential points:

$T_N = [P_1, P_2, ..., P_t, ..., P_N]$

P_t being a point of the trajectory at position (x_t, y_t).

The Fourier transform of trajectory T_N is:

$TF[T_N] = [X_0, X_1, ..., X_k, ..., X_{N-1}]$ with:

$$X_k = \sum_{n=0}^{N-1} e^{\frac{-i2\pi kn}{N}} \cdot P_n \ , k \in [\![0, N-1]\!]$$

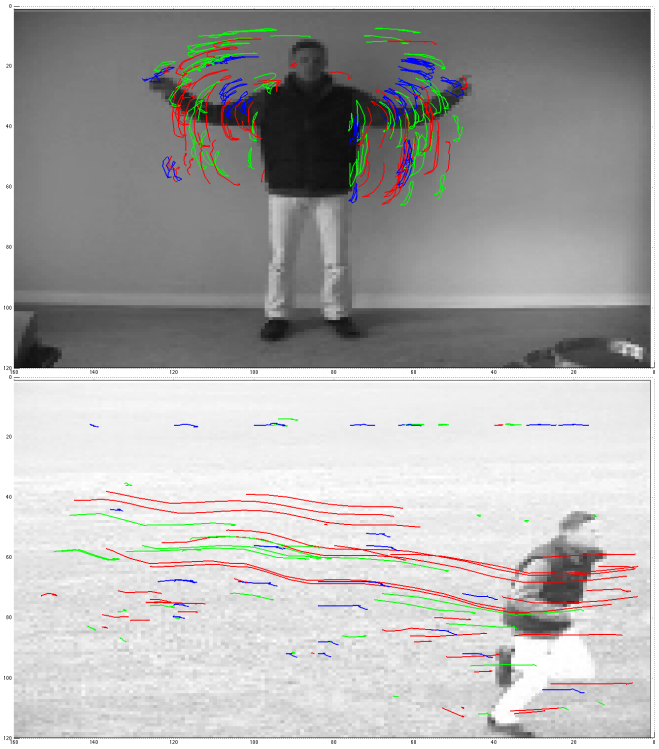

Fig. 3. Example of trajectories extracted from actions "hand-waving" and "running". In both cases, trajectories coincides with the observed motion.

To obtain translation invariance, the mean point value on this trajectory T_N is subtracted to each point (x_n, y_n).

$\tilde{x_n} = x_n - \sum_{t=1}^{N} \frac{x_t}{N}$ et $\tilde{y_n} = y_n - \sum_{t=1}^{N} \frac{y_t}{N}$

To obtain rotation invariance, trajectories T_N are considered as complex number vectors:

$T_{iN} = [P_{i1}, P_{i2}, ..., P_{it}, ..., P_{iN}]$

with $P_{it} = \tilde{x}_t + i\tilde{y}_t$ being the complex representation of point P_t. For a trajectory $T_{\theta iN}$ which represents a rotation by θ of the initial trajectory T_{iN}, the modulus of the Fourier transform of $T_{\theta iN}$ and T_{iN} are equal, giving rotation invariance.

Scale invariance is insured by normalizing the Fourier transform with the first non-zero frequency component:

$\tilde{X_k} = \frac{X_k}{|X_0|}$, $k \in [\![0, N-1]\!]$

Finally, descriptors based on the Fourier coefficients (FCD) are:

$FCD_{[T_{iN}]} = [|\tilde{X_0}|, |\tilde{X_1}|, ..., |\tilde{X_k}|, ..., |\tilde{X_{N-1}}|]$, $k \in [\![0; N-1]\!]$ with:

$X_k = \sum_{n=0}^{N-1} e^{\frac{-i2\pi kn}{N}} \cdot P_{in}$, $k \in [\![0, N-1]\!]$

All trajectories having the same size N, the FCD descriptor is also a fixed size.

Trajectories are finally smoothed by removing Fourier coefficients corresponding to high frequencies, which are assimilated to noise or tracking drift. This processing improves robustness with respect to small motion perturbations.

4. EVALUATION OF THE METHOD FOR ACTION RECOGNITION

4.1. Database used for assessing the method

4.1.1. The KTH Dataset

The KTH dataset [6] consists of six human action classes: "walking", "jogging", "running", "boxing", "waving" and "clapping". Each action is performed several times by 25 subjects with four different scenarios : outdoors, outdoors with scale variation, outdoors with different clothes and indoors. All sequences were shot with homogeneous backgrounds and a static camera at 25 fps. This dataset contains 600 videos.

4.2. Experiments

To evaluate performances of our method for action recognition, we use the bag of features approach [12]. This method has shown performance for text and image recognition [12] and is now commonly used for action recognition in videos.

The multi-channel approach [4, 9] is used to obtain a spatio-temporal bag of features. The video is subdivided with a grid structure and the bag of feature approach is computed on each grid cell. This permits a more localized approach for the bag of feature. A grid structure is called a channel. The spatio-temporal bag of feature approach uses several kinds of channels to combine information.

In the literature [9], a channel c is noted $hx \times vy \times tz$ such that:

-x is the number of horizontal subdivisions h.

-y is the number of vertical subdivisions v.

-z is the number of temporal subdivisions t.

A supervised SVM classification is then employed with a multi-dimensional gaussian kernel which allows a robust channel combination [12]:

$K(x_i, x_j) = \exp(-\sum_{c \in C} \frac{1}{A_c} D(H_i^c, H_j^c))$

where H_i^c and H_j^c are respectively the histograms of videos x_i and x_j and correspond to a channel c. [4, 9, 12]. $D(H_i^c, H_j^c)$ is the χ^2 distance and A_c a normalizing coefficient [12].

The classifier is trained on each descriptor. Two methods have been used to combine the results. The first is to concatenate channels obtained for each descriptor. The second method is the fusion of probabilities estimated by the multi-class Adaboost algorithm [2].

4.3. Results

The performance of our approach is evaluated on the KTH dataset, which is a classic benchmark for action recognition.

The method is applied on 120 frames for each video. Threshold l_1 is set to 0.15, which results in trajectories computed on 18 frames.

The influence on the recognition rate of the numbers of critical points and Fourier coefficients has been evaluated. As suggested by Fig 4, a maximum of 500 critical points with 80% of Fourier coefficients is kept.

Only one scale was used for the KTH dataset to compute trajectories. Results are indeed suggesting that a multi-scale analysis does not improve significantly the results on this database. We have finally used two channels, $h1 \times v1 \times t1$ and $h2 \times v1 \times t1$ for the multi-channel approach. [9].

Recognition rates obtained with different feature combinations of our approach are shown in Table 1.

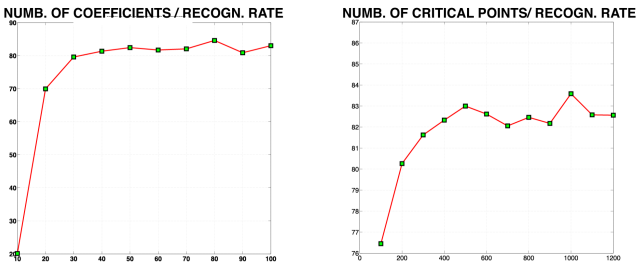

Fig. 4. Number of Fourier coefficients and critical points kept on the mean recognition rate. After a certain threshold, the mean recognition rate does not evolve significantly.

Descriptor	KTH Dataset
FCD	85.47%
HOG/HOF	91.98%
Concatenation	94.49%
Adaboost	**95.32%**

Table 1. Recognition rate with our approach for different kinds of descriptors.

4.3.1. Discussion

The FCD descriptor alone gives a satisfying recognition rate. It is however less efficient than the HOG/HOF descriptor. It can be explained by the fact that the KTH dataset cannot be totally discriminated by its frequency content. We also note that actions like "boxing" and "handshaking" have a similar frequency content (20.6% of confusion between those two classes). However, classes such as "running","jogging" and "walking" are visually similar but are performed at different frequencies: they are here strongly discriminated by the FCD descriptor (only 4.6% of confusion between those three classes). For the fusion method, Adaboost gives better results than the classical channel concatenation. It permits to obtain a recognition rate superior to most recent methods (Table 2). This illustrates the complementarity of the combined informations. The recent trajectory methods are close in terms of results [9, 5]. However, they are much more complex to implement, especially [5] which uses a pattern tracking for trajectory estimation or [9] which computes thirty channels, contrary to our approach which uses only two channels.

The method is computed with Matlab on a server with 2 Quadcore CPU at 3.1GHZ and 24 GB RAM. It takes 2.03 sec/frame to compute the optical flow and 1.71 sec/video to process the features.

Method	KTH Dataset
Williems *et al.* [11]	88.7%
Dollar *et al.* [1]	89.1%
Laptev *et al.*[4]	92.1%
Wang *et al.* [9]	94.2%
Raptis *et al.* [5]	94.8%
Our approach	**95.32%**
Vrigkas [8]	98.3%

Table 2. Recognition rate from the literature on KTH.

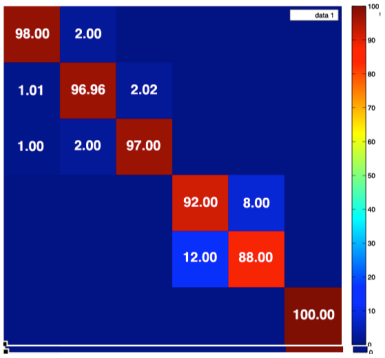

Fig. 5. Confusion matrix.

5. CONCLUSION

This paper presents a new approach for human action recognition in videos. These videos are characterized by critical points estimated from the optical flow field and by trajectories extracted from these points. Our results show that frequency information extracted from trajectories, combinated with motion and shape information, gives recognition rate among the highest on the KTH dataset (Table 2). Being a non dense sampling method, it allows lower computing complexity compared to other related methods (17.5 features/frame compared to 205.1 features/frame for [9]).

Current experiments on other complex datasets revealed the interest of multi-scale trajectory when combining information. Improvement of trajectory estimation, contextual information integration and recognition of complex activities are part of the ongoing work.

6. REFERENCES

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65 – 72, oct. 2005.

[2] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009.

[3] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, September 2005.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR 2008*, pages 1 –8, june 2008.

[5] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Proceedings of the 11th ECCV: Part I*, ECCV'10, pages 577–590, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36 Vol.3, 2004.

[7] D. Sun, S. Roth, and M.J. Black. Secrets of optical flow estimation and their principles. In *CVPR 2010. IEEE Conference on*, pages 2432–2439, 2010.

[8] M. Vrigkas, V. Karavasilis, C. Nikou, and A. Kakadiaris. Matching mixtures of curves for human action recognition. *CVIU*, 119(0):27 – 40, 2014.

[9] H. Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR. 2011 IEEE Conference on*, pages 3169 –3176, june 2011.

[10] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *University of Central Florida, U.S.A*, 2009.

[11] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th ECCV: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *CVPR Workshop, 2006. CVPRW '06. Conference on*, pages 13–13, 2006.