Motion Trend Patterns for Action Modelling and Recognition

Thanh Phuong Nguyen, Antoine Manzanera, and Matthieu Garrigues

ENSTA-ParisTech, 828 Boulevard des Maréchaux, 91762 Palaiseau CEDEX, France
{thanh-phuong.nguyen,antoine.manzanera,matthieu.garrigues}@ensta-paristech.fr

Abstract. A new method for action modelling is proposed, which combines the trajectory beam obtained by semi-dense point tracking and a local binary trend description inspired from the Local Binary Patterns (LBP). The semi dense trajectory approach represents a good trade-off between reliability and density of the motion field, whereas the LBP component allows to capture relevant elementary motion elements along each trajectory, which are encoded into mixed descriptors called Motion Trend Patterns (MTP). The combination of those two fast operators allows a real-time, on line computation of the action descriptors, composed of space-time blockwise histograms of MTP values, which are classified using a fast SVM classifier. An encoding scheme is proposed and compared with the state-of-the-art through an evaluation performed on two academic action video datasets.

Keywords: Action Recognition, Semi dense Trajectory field, Local Binary Pattern, Bag of Features.

1 Introduction

Action recognition has become a very important topic in computer vision in recent years, due to its applicative potential in many domains, like video surveillance, human computer interaction, or video indexing. In spite of many proposed methods exhibiting good results on academic databases, action recognition in real-time and real conditions is still a big challenge. Previous works cannot be evoked extensively, we refer to [1] for a comprehensive survey. In the following, we will concentrate on the two classes of method most related to our work: trajectory based modelling, and dynamic texture methods.

An important approach for action representation is to extract features from point trajectories of moving objects. It has been considered for a long time as an efficient feature to represent action. Johansson [2] showed that human subjects can perceive a structured action such as walking from points of light attached to the walker's body. Messing et al. [3], inspired by human psychovisual performance, extracted features from the velocity histories of keypoints using KLT tracker. Sun et al. [4] used trajectories of SIFT points and encoded motion in three levels of context information: point level, intra-trajectory context and inter-trajectory context. Wu et al. [5] used a dense trajectory field obtained by

tracking densely sampled particles driven by optical flow. They decomposed the trajectories into camera-induced and object-induced components using low rank optimisation. Then a set of global measures coming from the theory of chaotic dynamical systems are used to describe the action. Wang et al. [6] also used a dense trajectory field. They encoded the action information using histograms of the differential motion vectors computed along the boundary of the moving objects. Those works have shown the benefits of using dense motion features with respect to the sparse approaches, when using histogram based action descriptors.

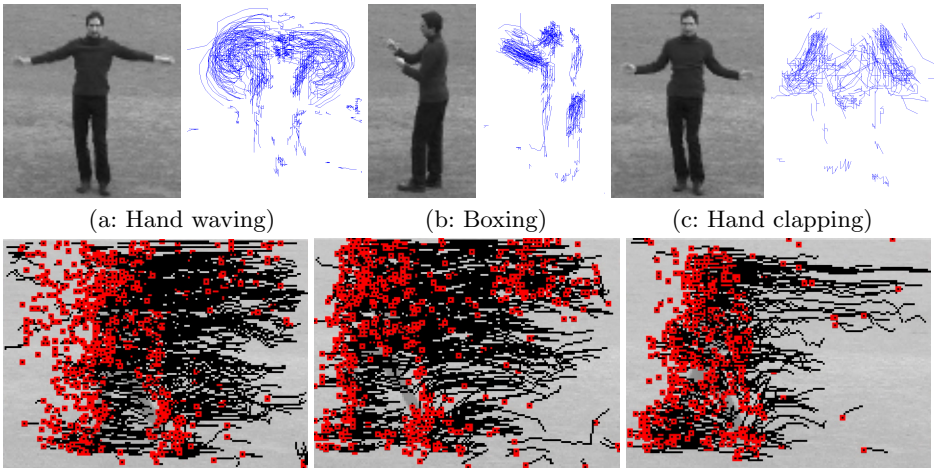
On the other hand, the LBP representation [7] was introduced for texture classification. It captures local image structure thanks to a binary sequence obtained by comparing values between neighbouring pixels. Due to its nice properties in terms of contrast invariance and computation time, LBP is very attractive for many applications, including action recognition. Zhao and Pietikäinen proposed an extension (LBP-TOP) [8] to dynamic texture by computing LBP on Three Orthogonal Planes (TOP), that was used by Kellokumpu et al. [9] in 3d space-time to represent human movement. In another approach [10], they used classical LBP on temporal templates (MEI and MHI, 2d images whose appearance is related to motion information). In these methods, the action is modelled using Hidden Markov Model to represent the dynamics of the LBPs. Recently, Yeffet and Wolf proposed LTP (Local Trinary Patterns) [11] that combines the effective description of LBP with the flexibility and appearance invariance of patch matching methods. They capture the motion effect on the local structure of self-similarities considering 3 neighbourhood circles at different instants. Kliper-Gross et al. extended this idea to Motion Interchange Patterns [12], which encodes local changes in different motion directions.

In this paper, we present a novel representation of human actions based on elementary motion elements called Motion Trend Patterns (MTP), that capture local trends along trajectories obtained by semi-dense point tracking. It combines the effective properties of the two previously presented techniques. The semi dense point tracking allows to obtain a large number of trajectories with a much smaller computational cost than fully dense tracking. We encode local direction changes along each trajectory using an LBP based representation. The combination of these approaches allows a real-time, on-line computation of action descriptors. The remaining of the paper is organised as follows: Section 2 summarises the computation of semi-dense trajectories. Section 3 details our elementary motion element, the MTP descriptor. Section 4 introduces the action modelling and its application to recognition. The last sections are dedicated to experiments, evaluation and discussion.

2 Semi Dense Beam of Trajectories

Trajectories are compact and rich information source to represent activity in video. 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 [13] 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. 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 changes thanks to a dominant acceleration estimation. Figure 1 shows several actions represented by their corresponding beams of trajectories.



(a: Hand waving) (b: Boxing) (c: Hand clapping)

(d: Jogging) (e: Running) (f: Walking)

Fig. 1. Actions from the KTH dataset represented as beams of trajectories. For actions (d-f), only the most recent part of the trajectory is displayed.

3 Motion Trend Patterns

We describe hereafter our MTP descriptor for action modelling. The input of the MTP is the previously described beam of trajectories, and no appearance information is used. An MTP descriptor is produced for every frame and for every point which belongs to a trajectory. It has two components: the motion itself, represented by the quantised velocity vector, and the motion local trend, represented by polarities of local direction changes.

3.1 Encoding of motion

Let $\overrightarrow{p_i}$ be the 2d displacement of the point between frames $i-1$ and i. The first part of the encoding is simply a dartboard quantisation of vector $\overrightarrow{p_i}$ (see Fig. 2). In our implementation, we used intervals of $\pi/6$ for the angles and 2 pixels for the norm (the last interval being $[6, +\infty[$), resulting in 12 bins for the angle, 4 bins for the norm.

3.2 Encoding of motion changes

Inspired by LBP, we encode elementary motion changes by comparing the motion vector $\overrightarrow{p_i}$ with its preceding and following velocities along the trajectory: $\{\overrightarrow{p_{i-1}}, \overrightarrow{p_{i+1}}\}$. Encoding the sign of the difference can be applied to the 2 components: (1) the norm, where it relates to tangential acceleration, and (2) the direction, where it relates to concavity and inflexion. The two can be encoded in a binary pattern. It turned out from our experiments that the use of the norm did not improve the results with respect to using the direction only, and then we only consider direction changes in MTP proposed hereafter.

Motion Trend Patterns (MTP): The local direction trend is encoded by the signs of the 2 differences between the direction $\angle\overrightarrow{p_i}$ and the directions of its 2 preceding and following motion vectors. This encoding corresponds to the local trend of the motion direction in terms of concavity and inflexion, as illustrated by Fig. 3, which shows the 4 possible configurations of MTP, for a fixed value of the quantised motion vector.

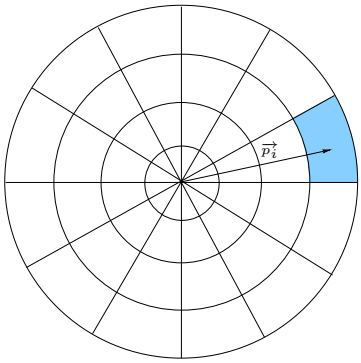


Fig. 2. Dartboard quantisation of the motion vector.

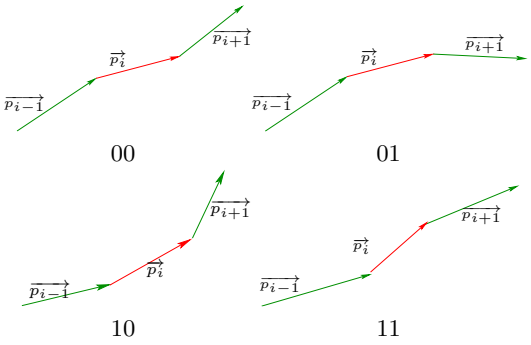


Fig. 3. Possible configurations of MTPs.

3.3 Properties of our descriptor

Several good properties of our action descriptor can be pointed out:

- *Low computation cost.* It is based on the semi dense beam of trajectories whose computation is very fast [13]. Thanks to the low complexity of LBP-based operators, the calculation of MTPs is also very fast and then our system is suitable for action recognition in real-time.
- *Invariance to monotonic changes of direction.* It inherits from LBPs their invariance to monotonic changes, which in our case correspond to changes in the curvature of concavities and inflexions.

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

4 Modelling and Recognition of Actions

Motion words and Action representation
The MTP descriptors represent elementary motion elements, or *motion words*. The number of different motion words is 48×2^2. Following hierarchical bag-of-features approach [14], we model an action by the space-time distribution of motion words: let B be the 3d space-time bounding box of the action. Histograms of MTPs are calculated in B and in sub-volumes of B, using regular pyramidal sub-grids, and the different histograms are concatenated as shown on Fig. 4. In our experiments we used 3 different sub-grids: $1 \times 1 \times 1$, $2 \times 2 \times 2$ and $4 \times 4 \times 4$, resulting in 73 histograms.

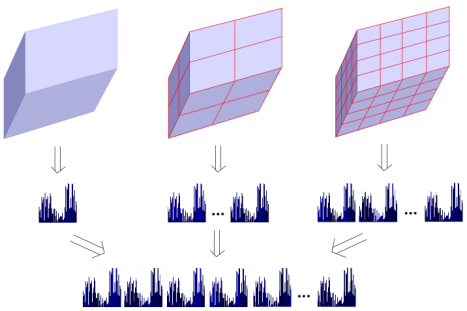


Fig. 4. Action modelling by concatenation of MTP histograms.

Classification
To perform action classification, we choose the SVM classifier of Vedaldi et al. [15] which approximates a large scale support vector machines using an explicit feature map for the additive class of kernels. Generally, it is much faster than non linear SVMs and it can be used in large scale problems.

5 Experiments

We evaluate our descriptor on two well-known datasets. The first one (KTH) [16] is a classic dataset, used to evaluate many action recognition methods. The second one (UCF Youtube) [17] is a more realistic and challenging dataset.
Extraction of semi-dense trajectories We have studied the influence of the extraction of semi-dense trajectories on the performance of our model. We

changed the parameters of the semi dense point tracker [13] to modify the number of trajectories obtained on the video. What we observe is that as long as the average matching error does not increase significantly, more we have trajectories, the better is the recognition rate. The improvement can be raised up to 5-6 % for KTH dataset. Table 1 shows the recognition rate obtained on this dataset, for different average number of tracked trajectories. In our experiments, the average number is set to 5 000 for a good result.

Table 1. Recognition rate on the KTH dataset in function of the number of trajectories.

Mean number of trajectories per video	1 240	2 700	3 200	5 271	7 763
Recognition rate	87.5	88.33	90	**92.5**	90.83

Experiments on KTH dataset The KTH 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). It contains 599 [1] videos, of which 399 are used for training, and the rest for testing. As designed by [16], the test set contains the actions of 9 people, and the training set corresponds to the 16 remaining persons.

Table 2 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 92.5 % which is comparable to the state-of-the-art, including LBP-based approaches (see Table 3). We recall that unlike [9, 10], which works on segmented boxes, our results can be obtained on line on unsegmented videos, using a preprocessing step to circumscribe the interest motions in space-time bounding boxes. The main error factor comes from confusion between jogging and running, which is a typical problem in reported methods. Table 6 presents the recognition rates obtained separately by the different components of our method: motion only, MTP only and both components (motion words). Clearly, quantised motion provides more information than the MTP component, but combining these complementary components allows to improve by 2 % the recognition rate.

Experimentation on UCF Youtube dataset This dataset [17] contains 1 600 video sequences with 11 action categories. Each category is divided into 25 groups sharing common appearance properties (actors, background, or other). Following the experimental protocol proposed by the authors [17], 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; our mean recognition rate (65.63 %) is comparable to recent methods (see Table 5).

[1] It should contain 600 videos but one is missing

Table 2. Confusion matrix on KTH dataset.

	Box.	Clap.	Wave	Jog.	Run.	Walk.
Boxing	97.5	2.5	0	0	0	0
Clapping	7.5	92.5	0	0	0	0
Waving	0	2.5	97.5	0	0	0
Jogging	0	0	0	95.0	2.5	2.5
Running	0	0	0	12.5	82.5	5.0
Walking	0	0	0	10.0	0	90.0

Table 3. Comparison with other methods on KTH dataset.

Method	Result	Method	Result
Our method	92.5	[11]	90.17
[18]	82.36	[12]	93.0
[19]	88.38	[9]	93.8
[10]	90.8	[17]	90.5
[6]	95.0		

Table 4. Confusion matrix on UCF. Ground truth (by row) and predicted (by column) labels are: basketball, biking, diving, golf swing, horse riding, soccer juggling, swing, tennis swing, trampoline jumping, volleyball spiking, walking with dog.

48.98	0	2.05	0	0	8.16	14.28	10.20	0	16.33	0
0	70.21	0	0	8.51	0	17.02	0	2.13	0	2.12
4.92	0	90.16	0	0	1.64	0	0	0	0	3.28
0	0	3.70	83.33	0	7.41	3.71	0	0	1.85	0
1.64	4.92	0	0	73.77	4.92	0	1.64	0	4.92	8.20
0	0	5.26	8.77	1.75	64.91	7.02	1.75	0	1.75	8.77
1.96	5.88	0	0	0	7.84	56.86	1.96	11.76	5.88	7.84
1.64	4.92	0	1.64	1.64	1.64	0	78.69	3.28	1.64	4.92
0	0	0	0	0	9.10	13.64	15.91	56.82	0	4.54
11.36	0	6.82	4.54	6.82	2.27	0	4.54	0	59.10	4.54
4.348	15.22	0	4.38	8.69	2.17	17.39	4.35	2.17	2.17	39.13

6 Conclusions

We have presented a new action model based on semi dense trajectories and LBP-like encoding of motion trend. It allows to perform on line action recognition on unsegmented videos at low computational complexity.

For the future, we are interested in using other variants of LBP operator. A temporal multi-scale approach for MTP encoding will also be considered. Furthermore, we will address the effects of moving camera in the performance of our model, in order to deal with uncontrolled realistic videos.

Acknowledgement

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

Table 5. Comparison on UCF Youtube.

Our method	[20]	[21]	[17]	[22]
65.63	64	64	71.2	56.8

Table 6. Experimentation on KTH using different components.

Motion	Motion changes	Motion words
90.42	84.58	92.5

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. **43** (2011) 16:1–16:43
2. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception and Psychophysics **14** (1973) 201–211
3. Messing, R., Pal, C., Kautz, H.A.: Activity recognition using the velocity histories of tracked keypoints. In: ICCV'09. (2009) 104–111
4. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR. (2009) 2004–2011
5. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: ICCV. (2011) 1419–1426
6. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR. (2011) 3169–3176
7. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24** (2002) 971–987
8. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. PAMI **29** (2007) 915–928
9. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. In: BMVC. (2008)
10. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Texture based description of movements for activity analysis. In: VISAPP (2). (2008) 206–213
11. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: ICCV. (2009) 492–497
12. Klipper-Gross, O., Gurovich, Y., Hassner, T., Wolf, L.: Motion interchange patterns for action recognition in unconstrained videos. In Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: ECCV. Volume 7577 of LNCS. (2012) 256–269
13. Garrigues, M., Manzanera, A.: Real time semi-dense point tracking. In Campilho, A.J.C., Kamel, M.S., eds.: ICIAR (1). Volume 7324 of LNCS. (2012) 245–252
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178
15. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. PAMI **34** (2012) 480–492
16. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: ICPR. (2004) 32–36
17. J. Liu, J.L., Shah, M.: Recognizing realistic actions from video "in the wild". In: CVPR. (2009) 1996–2003
18. Tabia, H., Gouiffès, M., Lacassagne, L.: Motion histogram quantification for human action recognition. In: ICPR. (2012) 2404–2407
19. Mattivi, R., Shao, L.: Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In Jiang, X., Petkov, N., eds.: CAIP. Volume 5702 of LNCS. (2009) 740–747
20. Lu, Z., Peng, Y., Ip, H.H.S.: Spectral learning of latent semantics for action recognition. In: ICCV. (2011) 1503–1510
21. Bregonzio, M., Li, J., Gong, S., Xiang, T.: Discriminative topics modelling for action feature selection and recognition. In: BMVC. (2010) 1–11
22. Wang, S., Yang, Y., Ma, Z., Li, X., Pang, C., Hauptmann, A.G.: Action recognition by exploring data distribution and feature correlation. In: CVPR. (2012) 1370–1377