Revisiting LBP-Based Texture Models for Human Action Recognition*

Thanh Phuong Nguyen[1], Antoine Manzanera[1], Ngoc-Son Vu[2], and Matthieu Garrigues[1]

[1] ENSTA-ParisTech, 828, Boulevard des Maréchaux, 91762 Palaiseau, France
[2] LIRIS, INSA Lyon, 20, Avenue Albert Einstein, 69621 Villeurbanne, France

Abstract. A new method for action recognition is proposed by revisiting LBP-based dynamic texture operators. It captures the similarity of motion around keypoints tracked by a realtime semi-dense point tracking method. The use of self-similarity operator allows to highlight the geometric shape of rigid parts of foreground object in a video sequence. Inheriting from the efficient representation of LBP-based methods and the appearance invariance of patch matching method, the method is well designed for capturing action primitives in unconstrained videos. Action recognition experiments, made on several academic action datasets validate the interest of our approach.

Keywords: action recognition, local binary pattern, dynamic texture,...

1 Introduction

Human activity recognition has been an active research topic in recent years due to its interesting application domains such as video surveillance, human computer interaction, video analysis, and so on. Many approaches have been introduced using different video features for action representation, we refer to [1] for a comprehensive survey. However a robust and real time method for action recognition with unconstrained videos is still a difficult challenge.

An interesting approach is to consider the action as a texture pattern, and to apply dynamic or static texture based methods to action modelling and recognition. Thanks to the effective properties of Local Binary Patterns (LBP) for texture representation, several LBP-based methods have also been proposed in the past for action recognition. Kellokumpu et al. [2] used dynamic texture operator (LBP-TOP) to represent human movement. They also presented another approach [3] using classical LBP on temporal templates (MEI and MHI images [4]) that were introduced to describe motion information from images. All extracted features in the two methods are then modelled using HMM (Hidden Markov Model). Mattivi and Shao [5] presented a different method using LBP-TOP to describe cuboids detected by Dollar's feature detector. Recently, Yeffet

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

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 286–293, 2013.
© Springer-Verlag Berlin Heidelberg 2013

and Wolf proposed LTP (Local Trinary Patterns) [6] that combines the effective description of LBP with the adaptivity and appearance invariance of patch matching methods. They capture the motion effect on the local structure of self-similarities considering 3 neighbourhood circles at a spatial position and different instants. Kliper-Gross et al. developed this idea by capturing local changes in motion directions with Motion Interchange Patterns (MIP) [7]. Nanni et al. [8] improved LBP-TOP using ternary units in the encoding step.

In this paper, we revisit dynamic texture based methods for action recognition. We are inspired by 2 popular LBP based representation: uniform LBP for texture coding and LTP for motion coding. We propose a new self-similarity operator to capture spatial relations in a trajectory beam, by representing the similarity of motion between the tracked point along its trajectory, and its neighbourhood. The semi-dense point tracker computes the displacement of many points in real time, then we apply self-similarity operator on appearance information to represent the motion information of a larger zone surrounding the trajectory. Our method can be seen as a hybrid solution between optical flow methods and dynamic texture based approaches. The rest is organised as follows. Section 2 briefly presents the basic material. The next section proposes our approach for action representation. The last sections are experiments and conclusions.

2 Basic Materials

2.1 LBP Based Operators

Uniform LBP. Local Binary Patterns [9] were introduced by Ojala et al. Their idea is to capture the local structures of texture images using binary patterns obtained by comparing a pixel value with its surrounding neighbours. LBP operator has two important properties: it is invariant to monotonic gray scale changes, and its complexity is very low. As a consequence, LBP-based approaches are suitable for many applications, aside from texture recognition. A LBP is called uniform if the number of binary transitions (from 0 to 1, from 1 to 0) while scanning the circle clockwise is at most 2. The uniform pattern coding ($LBP^{u2}_{n,r}$, corresponding to ignoring the non uniform patterns) is widely used in real applications because it reduces significantly the length of feature vectors while capturing important texture primitives (see Fig. 1).

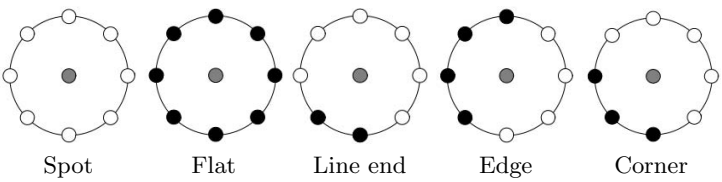

Fig. 1. Texture primitives corresponding to Uniform LBPs [9]

LTP. Local Trinary Patterns [6] use sum of squared differences (SSD) between patches centred at different space and time locations. Let $\mathrm{SSD}^{\Delta_{\mathbf{x}}}_{\Delta_t}$ be the SSD between the patch centred at pixel $\mathbf{x}$ at frame t and the patch centred at pixel $\mathbf{x} + \Delta_{\mathbf{x}}$ at frame $t + \Delta_t$. One ternary code $\{-1, 0, 1\}$ is obtained for each shift direction $\Delta_{\mathbf{x}}$, by comparing $\mathrm{SSD}^{\Delta_{\mathbf{x}}}_{-\Delta_t}$ and $\mathrm{SSD}^{\Delta_{\mathbf{x}}}_{+\Delta_t}$.

2.2 Motion Representation Using a Beam of Dense Trajectories

Trajectories are compact and rich information source to represent motion in videos, and have been used already for action recognition [10]. 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 [11] (see also Fig. 2) 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.

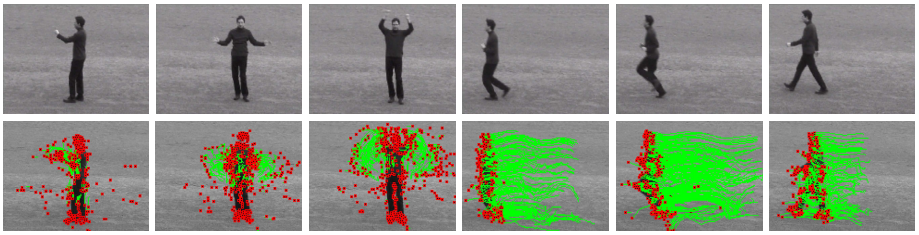

Fig. 2. Several actions of KTH dataset and their corresponding beam of trajectories. Red points represent tracked particles, green curves describe their trajectories.

3 Action Descriptor Using Spatial Motion Patterns

We present now our descriptor for action representation. The input data is the semi-dense trajectory beam described in Section 2. The classic approach to build motion information from optical flow is to consider histogram of optical flow (HOOF). This approach is simple to compute but neglects the spatio-temporal relation between moving points. One popular but limited solution is to consider the extracted histograms in different sub-volumes defined by a spatio-temporal grid. In this section, we introduce a descriptor that addresses more finely this problem. Briefly, the motion information is exploited at different context levels: (1) *Point level*; (2) *Local spatio-temporal level*; (3) *Regional to global spatio-temporal level.* This is detailed hereafter.

3.1 Point Level

Let $\overrightarrow{p_t}$ be the 2d displacement of the point between frames t and $t+\delta$. The first part of the encoding is simply a dartboard quantisation of vector $\overrightarrow{p_t}$ (see Fig. 3). 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 direction angle, 4 bins for norm.

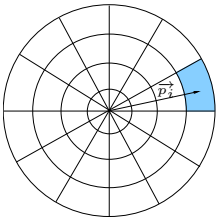

Fig. 3. Dartboard quantisation of the motion vector

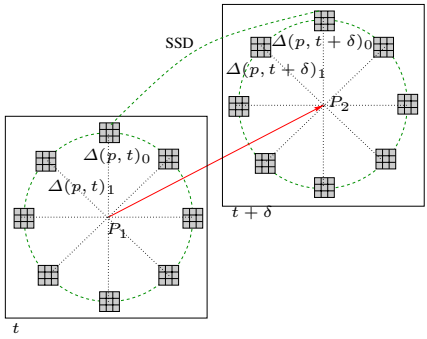

Fig. 4. The SMP descriptor is calculated at each tracked keypoint, along its trajectory. The consistency of motion in every direction is checked by computing the SSD between the corresponding image patches.

3.2 Local Spatio-temporal Level

At the local spatio-temporal level, we use an LBP-based dynamic texture to capture the relations between a point and its neighbours. Our idea is to capture the inter-trajectory relations among a beam of trajectories. We propose to combine the LBP-based self-similarity operator [9] and the appearance invariance of patch matching method inspired by [6]. This operator, called Spatial Motion Pattern (SMP), is presented below.

Spatial Motion Patterns

Consider a point p that moves from position P_1 at frame t to position P_2 at frame $t+\delta$, provided by the semi dense tracker [11]. The similarity of motion between this point and its neighbours is obtained by considering the $2\times n$ patches sampled from circles centred at P_1 and P_2 in their corresponding frames (see Fig. 4). Every index $i \in \{0, n-1\}$ represents a direction, which is encoded by 0 if the motion in this direction is similar to the motion of the centre point, and by 1 otherwise. Following [6], SSD (sum of square difference) score is used as similarity measure to check the consistency of motion.

Let $\{\Delta(p,t)_i\}_{i=0}^{n-1}$ be the set of n patches surrounding particle p at frame t. The corresponding SMP codeword $(b_0, b_1, \ldots, b_{n-1})$ is constructed as follows:

$$b_i = \begin{cases} 1 & \text{If } SSD\big(\Delta(p,t)_i, \Delta(p,t+\delta)_i\big) \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

where δ is the time interval between two frames, τ is the SSD threshold.
Our local descriptor differs significantly from LTP in several aspects:

- *Encoding process.* Unlike [6], our descriptor uses only 2 bits. The encoding of LTP is done by comparing SSD scores between neighbouring patches of past and future frames, and the centre patch of the middle frame. Our method estimates the SSD scores between two corresponding patches in two consecutive frames.
- *Neighbouring configuration.* LTP used three circles centred at the same position in 2D space. In our approach, the two neighbouring circles are centred at the tracked position of each keypoint.
- *Interpretation.* LTP aims to represent motion information at a given position, whereas in our case, the motion information is already known, the SMP is interpreted as a local disparity map of velocities around each trajectory.

Properties of Spatial Motion Patterns
Inheriting from [6, 9], Spatial Motion Patterns have attractive properties:

- *Simple computation.* They use SSD scores on small image patches. In addition, the calculation is only applied on tracked keypoints, not on the whole image, avoiding many irrelevant calculations.
- *Appearance invariance.* This property is due to: (1) the LBP based encoding and (2) the basic information which only relates to the trajectory, not to the appearance.

SMP uniform patterns (SMP^{u2}) captures local primitives action in a similar way as LBP uniform patterns (LBP^{u2}). They detect the motions between foreground objects and the background in videos, and more generally, between two rigid parts of a moving object. We can point out the relation between SMP^{u2} and action primitives as follows (see also Fig. 1).

- *Spot:* A small foreground object move on the background.
- *Flat:* A big rigid part of a moving object.
- *Line end:* End of a thin foreground object.
- *Edge:* Border between two parts of a moving object, or between a foreground object and the background.
- *Corner:* A corner of a rigid part of a moving object.

Fig. 5 illustrates the interpretation of SMP uniform patterns (SMP^{u2}).

It is also worth mentioning that, unlike many other methods, the more complex the background, the more efficiently should the SMP describe the rigid parts of the moving object.

Fig. 5. SMP^{u2} configurations allow to determine the shape of the rigid parts of the moving object around the keypoints (in red points). In the neighbouring circles, image patches in green (resp. blue) indicates that they belong to the same rigid part of the moving object as the keypoint (resp. another rigid part or the background).

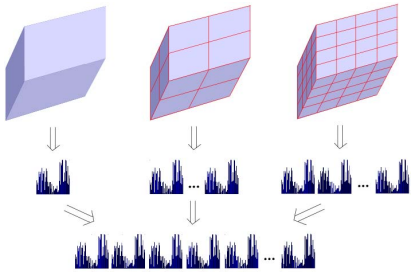

Fig. 6. Action modelling by SMP histogram concatenation

3.3 Regional to Global Spatio-temporal Level

In this context, a pyramidal bag of feature (BoF) [12] is used to represent action by histograms of codewords made of the two previous primitives (motion code and spatial motion patterns) on spatio-temporal volumes. All histograms are concatenated into one vector that is then normalised for action representation. Fig. 6 shows how to construct the action description using three different grids.

4 Experimentation on Human Action Classification

4.1 Classification

To perform action classification, we choose the SVM classifier of Vedaldi et al. [13] 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.

4.2 Experimentation

We evaluate our descriptor on two well-known datasets. The first one (KTH) [14] is a classic dataset, used to evaluate many action recognition methods. The second one (UCF Youtube) [15] is a more realistic and challenging dataset.

Parameter Settings. There are several parameters concerning the construction of SMP. Like [6], we compute SSD score on image patch of size 3×3 with threshold $\tau = 1000$ that represents 0.17% maximal value of SSD. The time interval δ is set to 1. Because every tracked keypoint already represents a certain

spatial structure, the radius of SMP must be sufficiently large to better capture the geometric shape of rigid parts of moving object surrounding the keypoints. In our implementation, we consider 16 neighbours sampled on a circle of radius 9. In addition, only uniform patterns ($SMP^{u2}_{16,9}$) are considered. To construct the histograms of codewords, we used 3 spatiotemporal grids: $1 \times 1 \times 1$, $2 \times 2 \times 2$ and $3 \times 3 \times 3$.

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). It contains 599 [1] videos, of which 399 are used for training, and the rest for testing. As designed by [14], 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 by row. The average recognition rate is 93.33 % which is comparable to the state-of-the-art of LBP-based approaches (see Table 2). We remark that unlike [2, 3] that work on segmented box, our results are obtained directly on unsegmented videos. Applying the same pre-processing step would probably improve our result.

Table 1. Confusion matrix on KTH dataset

	Box.	Clap.	Wave	Jog.	Run.	Walk.
Boxing	97.5	2.5	0	0	0	0
Clapping	2.5	97.5	0	0	0	0
Waving	2.5	0	97.5	0	0	0
Jogging	0	0	0	95.0	0	5.0
Running	0	0	0	12.5	80.0	7.5
Walking	0	0	0	10.0	0	90.0

Table 2. Comparison on KTH dataset

Method	Result	Method	Result
Ours	93.33	[6]	90.17
[3]	90.8	[7]	93.0
[5]	88.38	[2]	93.8

Table 3. Comparison on UCF Youtube

Our method	[16]	[17]	[15]
72.07	64	64	71.2

Experimentation on UCF Youtube Dataset. The UCF Youtube dataset records 11 categories (basketball shooting, cycling, diving, golf swinging, horse back riding, 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). It is much more challenging than KTH because of its large variability in terms of viewpoints, backgrounds and camera motions. Following the experimental protocol proposed by the authors [15], we used 9 groups out of the 25 as test and the 16 remaining groups as training data. Our mean recognition rate on UCF Youtube dataset is 72.07 % (see Table 3), which outperforms recent methods.

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

5 Conclusions

We have presented a new method for action recognition based on semi-dense trajectory beam and the LBP philosophy. Its main idea is to capture spatial relation of moving parts around the tracked keypoints, along their trajectories. Our descriptor is designed to capture geometric shape of the rigid parts of moving object in unconstrained videos with complex background. In the future, we are interested in several perspectives related to this method such as multi-scale SMPs, and extension to moving backgrounds.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. 16, 16:1–16:43 (2011)
2. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. In: BMVC (2008)
3. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Texture based description of movements for activity analysis. In: VISAPP (2), pp. 206–213 (2008)
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. PAMI 23, 257–267 (2001)
5. Mattivi, R., Shao, L.: Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 740–747. Springer, Heidelberg (2009)
6. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: ICCV, pp. 492–497 (2009)
7. Kliper-Gross, O., Gurovich, Y., Hassner, T., Wolf, L.: Motion interchange patterns for action recognition in unconstrained videos. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 256–269. Springer, Heidelberg (2012)
8. Nanni, L., Brahnam, S., Lumini, A.: Local ternary patterns from three orthogonal planes for human action classification. Expert Syst. Appl. 38, 5125–5128 (2011)
9. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24, 971–987 (2002)
10. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR, pp. 3169–3176 (2011)
11. Garrigues, M., Manzanera, A.: Real time semi-dense point tracking. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part I. LNCS, vol. 7324, pp. 245–252. Springer, Heidelberg (2012)
12. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2), pp. 2169–2178 (2006)
13. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. PAMI 34, 480–492 (2012)
14. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: ICPR, pp. 32–36 (2004)
15. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from video "in the wild". In: CVPR, pp. 1996–2003 (2009)
16. Lu, Z., Peng, Y., Ip, H.H.S.: Spectral learning of latent semantics for action recognition. In: ICCV, pp. 1503–1510 (2011)
17. Bregonzio, M., Li, J., Gong, S., Xiang, T.: Discriminative topics modelling for action feature selection and recognition. In: BMVC, pp. 1–11 (2010)