Pairwise features for human action recognition

Anh-Phuong Ta Christian Wolf Guillaume Lavoué Atilla Baskurt Jean-Michel Jolion
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
{anh-phuong.ta,christian.wolf,guillaume.lavoue,atilla.baskurt,jean-michel.jolion}@liris.cnrs.fr

Abstract

Existing action recognition approaches mainly rely on the discriminative power of individual local descriptors extracted from spatio-temporal interest points (STIP), while the geometric relationships among the local features[1] *are ignored. This paper presents new features, called pairwise features (PWF), which encode both the appearance and the spatio-temporal relations of the local features for action recognition. First STIPs are extracted, then PWFs are constructed by grouping pairs of STIPs which are both close in space and close in time. We propose a combination of two codebooks for video representation. Experiments on two standard human action datasets: the KTH dataset and the Weizmann dataset show that the proposed approach outperforms most existing methods.*

Key-words: action recognition, local features, pairwise features.

1. Introduction

Based on the features used for recognition, existing action recognition methods can be broadly divided into two categories: local approaches [1, 9, 12, 14] and holistic approaches [8, 6, 17, 16]. However, some methods do not neatly fall into these categories, e.g Sun et al. [15] combine the local and holistic features. Most of the holistic-based approaches need the pre-processing of input data such as background subtraction or tracking. The local-based approaches overcome some limitations by exploiting robust descriptors extracted from interest points. Most of these methods are based on bag-of-words models, which have been very successful for text analysis, information retrieval and image classification. Inspired by this, a number of works have shown very good results for human action recognition [1, 9, 12, 14] using the bag of words (BoW) models. However, BoW models discard the spatio-temporal layout of the local features which may be almost as important as the features themselves.

To overcome the limitations of the BoW models, some researchers concentrate on exploiting information on the spatial and temporal distribution of interest points. Liu and Shah [7] explore the correlation of video-word clusters using a modified correlogram. Gilbert et al. [3] spatially concatenate corner descriptors detected on different regions and apply data mining techniques to construct compound features. Zhang et al. [18] introduce the concept of motion context to capture both spatial and temporal distribution of motion words. Oikonomopoulos et al. [10] encode the spatial co-occurrences of pairs of codewords and propose a probabilistic spatiotemporal voting framework for localizing and recognizing human activities in unsegmented image sequences. Ryoo and Aggarwal [11] present a spatio-temporal relationship matching method through defining the spatial and temporal predicates.

The methods mentioned above are mainly based on the discriminative power of individual local features for codebook construction and thus the performance depends on the local features used. In this paper[2] we present new features, called pairwise features (PWF), which capture both appearance and geometric relationships among local features. Our method differs from the state-of-the-art methods in two main points. First, our feature is not limited to the appearance information but also includes geometric information. Second, our method incorporates both the spatial and temporal relationships among local features in a significant manner, i.e such relationships are constrained to time and space.

The rest of this paper is organized as follows. In

[1] We consider the features extracted from interest points as local features

[2] This project was financed through the French National grant ANR-CaNaDA *Comportements Anormaux : Analyse, Détection, Alerte*, No. 128, which is part of the call for projects CSOSG 2006 *Concepts Systèmes et Outils pour la Sécurité Globale.*

section 2, we introduce our new features and propose a novel action representation exploiting our features. The experimental results are presented in section 3. Finally we conclude and give some perspectives of this work.

2. Pairwise features for action recognition

2.1 Overview of the proposed method

The spatio-temporal interest points (STIPs) are first extracted from the video sequences, then PWFs[3] are constructed from them. We apply the bag of words (BoW) model using PWFs for video quantization, which requires creating a visual vocabulary. To this end, we generate two codebooks (vocabularies) according to the appearance and the geometric similarity of the PWFs. A video sequence is then represented by two histograms of visual words. These two histograms are combined into a single feature vector, and Support Vector Machines (SVM) are used for action classification.

2.2 Pairwise descriptor

We propose new features, namely pairwise features (PWF), which encode both STIP descriptors and the spatio-temporal relations among the STIPs. Essentially, two STIPs are connected to form a PWF if they are adjacent in space and in time. Intuitively, two STIPs that are close both in space and in time should belong to the same human activity. A spatio-temporal detector usualy detects interest points locating salient changes in a video sequence, and descriptors are extracted around these interest points. Thus, in general a local feature contains two types of information: appearance information, and its space-time coordinate information. We denote a spatio-temporal local feature as $f = (f_{des}, f_{loc})$ where f_{des} is an arbitrary appearance descriptor and f_{loc} is its space-time coordinates. Let $f_1 = (f_{1des}, f_{1loc})$ and $f_2 = (f_{2des}, f_{2loc})$ be two STIPs, a PWF $p = (F_{des}, \mathbf{F}_{vec})$ is established if the conditions below are satisfied:

a) $d_s(f_{1loc}, f_{2loc}) \leq t_s$, AND
b) $d_t(f_{1loc}, f_{2loc}) \leq t_t$

where F_{des} is a concatenation of f_{1des} and f_{2des}; $\mathbf{F}_{vec}$ is a geometric vector from the first location (f_{1loc}) to the second one (f_{2loc}) in temporal order. If f_1 and f_2 are in the same frame, the STIPs in the PWF are ordered from left to right, according to their x coordinate; t_s is a spatial threshold, t_t is a temporal threshold; $d_s(.,.)$ and $d_t(.,.)$ are spatial distance and temporal distance functions, respectively. We can imagine each PWF as a segment in 3D space (see figure 1 for an illustration). As shown in figure 1, apart from spatio-temporal relationships, we are also motivated to take into account spatial relationships (i.e., PWFs in the same frame).

[3]pairwise features

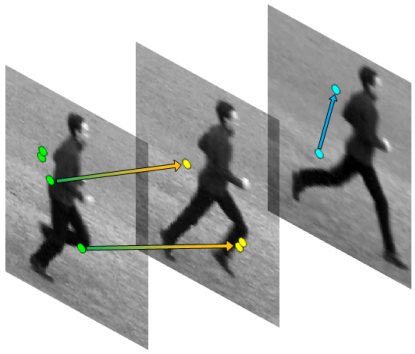

Figure 1: Illustration of several pairwise features as segments in space-time.

2.3 Comparing PWFs

Once the PWFs are constructed, the video sequence is considered as a collection of PWFs (segments). In order to compare different PWFs, we need a similarity measure. Note that a PWF p contains not only the appearance descriptor F_{des} but also the geometric information $\mathbf{F}_{vec}$, where the geometric information is translation invariant. Since the physical meanings are different, it is not really feasible to construct a single codebook based on the descriptors of PWFs. We propose to generate two codebooks: the first codebook is generated by clustering only the appearance descriptors of the PWFs, and the second one is generated through the clustering of the geometric descriptors of the PWFs. We present hereafter a simple similarity for the geometric descriptors of the PWFs. Let p^i and p^j be two PWFs, a measure $d_g(p^i, p^j)$ of the geometric similarity between two PWFs is given as follows:

$$d_g(p^i, p^j) = \frac{(\mathbf{F}^i_{vec} - \mathbf{F}^j_{vec})^T \Sigma^{-1} (\mathbf{F}^i_{vec} - \mathbf{F}^j_{vec})}{||\mathbf{F}^i_{vec}|| + ||\mathbf{F}^j_{vec}||} \quad (1)$$

where Σ is a matrix capturing variations and correlations between the data dimensions similar to a covariance matrix; In our case we propose to set it to a diagonal matrix diag($[\lambda_s, \lambda_s, \lambda_t]$) where λ_s and λ_t are thus weights adjusting the differences between the spatial and the temporal domains. Supposing the existence of a distance function d_d between the two descriptor parts F^i_{des} and F^j_{des} of the PWF p^i and p^j, the two codebooks

can be constructed by using a clustering algorithm like K-means.

By mapping the PWFs extracted from a video to the vocabularies, a video sequence is represented by the two histograms of visual words, which are denoted as H^d and H^g. We introduce a combination of these two histograms to form a feature vector as input to a classifier such as SVM:

$$H = \{\alpha * H^d, (1-\alpha) * H^g\} \quad (2)$$

where α is a weighting coefficient. The advantage of this combination method is that the two histograms do not necessarily have the same size.

3. Experiments

3.1 Experimental results

Our experiments are carried out on the standard KTH and Weizmann human action datasets. We perform leave-one-out cross-validation and report the average accuracies. Our feature can be applied to any STIP detector, but for the purpose of testing our method we apply the Dollar detector [1] to detect STIPs and use the cuboid descriptor (i.e, flattening gradients of the cuboid into a vector) from [1] as the basis feature for our PWF. K-means clustering is applied to define the visual vocabulary, with 512 visual words for KTH and 250 for Weizmann, respectively. For constructing PWFs, t_s was set to 20 and t_t was 5. To reduce complexity, we select the 1500 most significant PWFs for each video sequence, where significance is measured as the product of the response functions of the two interest points corresponding to the PWF. Recognition was performed using a non-linear SVM with a chi-squared kernel. To adjust the differences between the spatial and the temporal domains, λ_s and λ_t were set to 7 and 1, respectively for all experiments.

Figure 2 shows the performance obtained on the KTH dataset. It can be observed (fig. 2a) that even with only geometric descritors used for recognition, the result is very good. This result is very promising and suggests that it is possible to avoid the non-trivial problem of choosing an appearance descriptors for action recognition, i.e exploiting only geometric distribution among STIPs. The best result obtained for the KTH dataset is reported in figure 2d, which indicates that the appearance and the geometric descriptor of the PWF are complementary to each other. To evaluate the stability of our method, we also performed the tests with $\lambda_s = \lambda_t = 1$, the accuracies moderately decrease by roughly 1% for both datasets.

Table 1: Comparison of our method with different methods, tested on KTH and Weizmann datasets.

Method	KTH	Weizmann
Our method	**93.0**	**94.5**
Dollar et al. [1]	81.2	-
Niebles et al. [9]	83.3	90.0
Savarese et al. [12]	86.8	-
Gilbert et al. [3]	89.9	-
Oikonomopoulos et al. [10]	80.5	-
Zhang et al. [18]	91.3	-
Ryoo and Aggarwal [11]	93.8	-
Liu and Shah [7]	94.2	-
Sun et al. [15]	94.0	97.8
Laptev et al. [5]	91.8	-
Fathi and Mori [2]	90.5	100.0
Gorelick et al. [4]	-	99.6
Schindler and Gool [13]	92.7	100.0

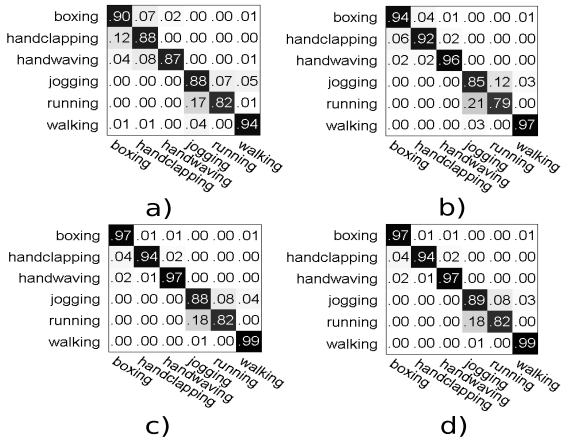

Figure 2: Confusion matrices on the test KTH dataset: a) $\alpha = 0$ and accuracy (ac) $= 88.2\%$; b) $\alpha = 1.0$ and ac $= 90.5\%$; c) $\alpha = 0.5$ and ac $= 92.8\%$; d) $\alpha = 0.6$ and ac $= 93.0\%$;

3.2 Comparison to the state-of-the-art

Table 1 presents a comparison of our results with state-of-the-art results. This table is divided into three groups: the first group consists of the methods which can be directly compared with ours, i.e use the same features (cuboid descriptor) and the same experimental setup for action classification; the second one includes the methods that exploit spatio-temporal relations among STIPs or among visual words. Among the methods in the first group, our method obtained the best accuracy for both datasets. Our method outperforms most existing methods which take into account the spatio-temporal relationships (second group). Note that the results which are most close to ours, i.e the work from

Table 2: Testing the performance of the feature parts (appearance, geometry) of the PWF in a new experimental set-up: learning on one dataset and testing on another one.

Feature part	Learning	Testing	ac.
Appearance descriptor	KTH	WM	40.0
	WM	KTH	48.2
Geometric descriptor	KTH	WM	70.0
	WM	KTH	83.3

Ryoo and Aggarwal [11], requires encoding more semantic information about the interactions (17 in total) between STIPs. On the KTH dataset, our recognition rate of 93.0% is very close to the current best rate of 94.2% . It should be noted that the selection of different numbers of codewords sometimes lead to different recognition rates. Note that we cannot directly compare to results reported in the last group of table 1, because they exploit both holistic and local representation [15], included more data given by segmentation masks [4, 2], or use another experimental set-up.

Beyond this straightforward comparison, our method offers an important advantage over all other methods: the geometric descriptor of our PWF is a generic feature, which keeps its discriminative power well across different datasets. To verify this fact, we performed an experimental evaluation through training on one dataset and testing on another one. More precisely, the experimental tests are carried out using three common actions which exist in both datasets: handwaving, running, walking. Because there are only 29 videos of these actions in the Weizmann dataset, we randomly select 30 videos containing these three actions from the KTH dataset to perform our tests. Table 2 shows the results obtained using each feature part of the PWFs alone. From this table, it can be seen that the appearance descriptors of the PWFs failed, while the geometric descriptors work very well. These results are very interesting for future research in human action recognition. Note that the video sequences in the Weizmann dataset are of much shorter duration compared to the ones in the KTH dataset. This explains why there are large differences in the performance in our second test (e.g. from $\%70$ to 83%).

4 Conclusions

In this paper, we propose a new feature for action recognition. Our method differs significantly from previous local-based approaches in that our feature is a semi-local representation, which encodes both the appearance and geometric information among local features. Through experiments, we have proved that exploiting the location information of local features which have been ignored in the literature, gives valuable improvements to the conventional BoW models. Our ongoing work aims at extending this method to capture more complex geometric structures among the local features.

References

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. pages 65–72. VS-PETS, 2005.

[2] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.

[3] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *ECCV*, pages 222–233, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, December 2007.

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

[6] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *CVPR*, 2008.

[7] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

[8] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. *In CVPR, 2008*.

[9] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV, 2008*.

[10] A. Oikonomopoulos, I. Patras, and M. Pantic. An implicit spatiotemporal shape model for human activity localization and recognition. *In CVPR*, 0:27–33, 2009.

[11] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[12] S. Savarese, A. Del Pozo, J. Niebles, and F. Li. Spatial-temporal correlatons for unsupervised action classification. In *In WMVC*, pages 1–8, 2008.

[13] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *CVPR*. IEEE Press, June 2008.

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32–36 Vol.3, September 2004.

[15] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *CVPR Workshop*, pages 58–65, 2009.

[16] L. Wang, X. Geng, C. Leckie, and K. Ramamohanarao. Moving shape dynamics: A signal processing perspective. In *CVPR*, 2008.

[17] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, pages 1–7, 2007.

[18] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia. Motion context: A new representation for human action recognition. In *ECCV (4)*, pages 817–829, 2008.