Local polynomial space-time descriptors for actions classification

Olivier Kihl, David Picard
ETIS/ ENSEA - Université Cergy-Pontoise
CNRS, UMR 8051, France
olivier.kihl@ensea.fr, picard@ensea.fr

Philippe-Henri Gosselin
INRIA Rennes Bretagne Atlantique
France
philippe.gosselin@inria.fr

Abstract

In this paper we propose to tackle human actions indexing by introducing a new local motion descriptor. Our proposed descriptor is based on two modeling, a spatial model and a temporal model. The spatial model is computed by projection of optical flow onto bivariate orthogonal polynomials. Then, the time evolution of spatial coefficients is modeled with a one dimension polynomial basis. To perform the action classification, we extend recent still image signatures using local descriptors to our proposal and combine them with linear SVM classifiers. The experiments are carried out on the well known KTH dataset and on the more challenging Hollywood2 action classification dataset and show promising results.

1 Introduction

Space-time feature descriptors [12, 6, 9] have become essential tools in action classification. In this paper we propose a new space-time motion descriptor based on polynomial decomposition of the optical flow. Our descriptor is localized spatially and temporally in a space-time tube, in order to capture characteristic atoms of motion. We propose to model the vector field of motion between two frames using projection on orthogonal polynomials. Then, we model the evolution of polynomial coefficient along time. We name this novel descriptor Series of local Polynomial Approximation of optical Flow ($SoPAF$).

The paper is organized as follows. In section 2 we present the most popular space-time feature descriptors in the literature. Then, in section 3 we present our descriptor. Finally, in section 4 we carry out experiments on two well known action classification datasets.

2 Related work

The recognition of human action and activity is an important area in several fields such as computer vision, machine learning and signal processing. A popular way of comparing videos is to extract a set of descriptors from video and then find a transformation that maps the set of descriptors into a single vector and then measure the similarity between the obtained vectors (for example in [12]). In the past ten years, several descriptors have been proposed.

2.1 Video descriptors

Several descriptors used for action classification consist in the extention to video of still image descriptors, in particular the well known (Lowe 2D) SIFT descriptor [7]. This descriptor relies on a histogram of orientation of gradient. Many other descriptors are close to the SIFT. The most commonly used are the Histogram of oriented gradient (HOG) [1], the Histogram of Oriented Flow (HOF) [1] and the Motion Boundary Histogram (MBH) [1]. HOG is very similar to SIFT, but is not only computed on salient points. In the same way, Dalal et al also propose the Histogram of Oriented Flow (HOF) [1] which is the same as HOG but applied to optical flow instead of the gradient. They also propose the Motion Boundary Histogram (MBH) that model the spatial derivative of each component of the optical flow vector field with a HOG.

Recently, Wang et al [12] propose to model these descriptors along dense trajectories. The time evolution of trajectories, HOG, HOF and MBH is modelled using a space time grid along trajectories. To our knowledge, they obtained state of the art results.

The descriptor we propose here is inpired from HOF and dense trajectories. Our model is based on a polynomial approximation of the field rather than a histogram. We approximate the time evolution thanks to an approach based on function regression. In our approach, decriptors are not extracted only on salient points, we use a dense extraction with a regular spatial step between each descriptor.

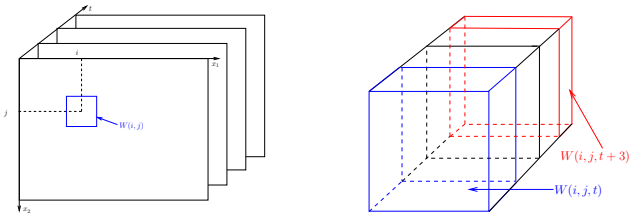
2.2 Signatures

Once a set of descriptors is obtained from the video, a signature has to be computed for the whole video. The most common method for computing signatures is called the "Bag of Words" (BoW) approach [10]. Given a dictionary of descriptor prototypes (usually by clustering a large number of descriptors), the histogram of occurrences of these prototypes within the video is computed.

In this paper, we consider a compressed version of VLAT [8] which is known to achieve performances close to state of the art in still images classification when very large sets of descriptors are extracted from the images. This method uses an encoding procedure based on high order statistics deviation of clusters. In our case, the dense sampling both in spatial and temporal directions leads to highly populated sets, which is consistent with the second order statistics computed in VLAT signatures.

3 Series of Polynomial Approximation of Flow (SoPAF)

We propose to model the vector field of motion between two frames using projection on an orthogonal basis of polynomials. This polynomial model is used in [4] to recognize movements in a video. The modeling is applied to the entire field and each frame is processed separately.



(a) (b)

Figure 1: Localisation in the space and space-time domain ; (a) Localisation in the space domain ; (b) Localization example in the space-time domain with $\tau = 3$

Since motion can succesfully be modeled by polynomials, we propose to use such models on a local neighborhood in order to obtain densely extracted local motion descriptors. We use two successive polynomial models. At first, we model the spatial vector field. Then, time evolution of spatial coefficients are modeled.

3.1 Spatial modeling using a polynomial basis

Let us consider the descriptor $\mathbf{M}(i,j,t)$ located in frame at coordinates (i,j) and in video stream at time t. Descriptors are computed using space and time neighborhood around location (i,j,t), denoted as window $W(i,j,t)$. An example of $W(i,j,t)$ is shown in Figure 1a. We propose to model the vector field of motion inside the window $W(i,j,t)$ by a finite expansion of orthogonal polynomials. Let us define the family of polynomial functions with two real variables as follows:

$$P_{K,L}(x_1,x_2) = \sum_{k=0}^{K}\sum_{l=0}^{L} a_{k,l}\ x_1^k\ x_2^l \tag{1}$$

where $K \in \mathbb{N}^+$ and $L \in \mathbb{N}^+$ are respectively the maximum degree of the variables (x_1,x_2) and $\{a_{k,l}\}_{k\in\{0..K\},l\in\{0..L\}} \in \mathbb{R}^{(K+1)\times(L+1)}$ are the polynomial coefficients. The global degree of the polynomial is $D = K + L$.

Let $\mathcal{B} = \{P_{k,l}\}_{k\in\{0..K\},l\in\{0..L\}}$ be an orthogonal basis of polynomials. A basis of degree D is composed by n polynomials with $n = (D+1)(D+2)/2$ as follows:

$$\begin{aligned}\mathcal{B} = \quad & \{P_{0,0}, P_{0,1}, \cdots, P_{0,L}, P_{1,0}, \cdots \\ & \cdots, P_{1,L-1}, \cdots, P_{K-1,0}, P_{K-1,1}, P_{K,0}\}\end{aligned} \tag{2}$$

We can create an orthogonal basis using the following three terms recurrence:

$$\begin{cases} P_{-1,l}(\mathbf{x}) = 0 \\ P_{k,-1}(\mathbf{x}) = 0 \\ P_{0,0}(\mathbf{x}) = 1 \\ P_{k+1,l}(\mathbf{x}) = (x_1 - \lambda_{k+1,l})P_{k,l}(\mathbf{x}) - \mu_{k+1,1}P_{k-1,l}(\mathbf{x}) \\ P_{k,l+1}(\mathbf{x}) = (x_2 - \lambda_{k,l+1})P_{k,l}(\mathbf{x}) - \mu_{k,l+1}P_{k,l-1}(\mathbf{x}) \end{cases} \tag{3}$$

where $\mathbf{x} = (x_1,x_2)$ and the coefficients $\lambda_{k,l}$ and $\mu_{k,l}$ are given by

$$\begin{aligned} \lambda_{k+1,l} &= \frac{\langle x_1 P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l}(\mathbf{x})\|^2} & \lambda_{k,l+1} &= \frac{\langle x_2 P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l}(\mathbf{x})\|^2} \\ \mu_{k+1,l} &= \frac{\langle P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k-1,l}(\mathbf{x})\|^2} & \mu_{k,l+1} &= \frac{\langle P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l-1}(\mathbf{x})\|^2} \end{aligned} \tag{4}$$

and $\langle \cdot \mid \cdot \rangle$ is the usual inner product for polynomial functions:

$$\langle P_1 \mid P_2 \rangle = \iint_\Omega P_1(\mathbf{x})P_2(\mathbf{x})w(\mathbf{x})d\mathbf{x} \tag{5}$$

with w the weighting function that determines the polynomial family and Ω the spatial domain covered by the window $W(i,j,t)$. We use Legendre polynomials ($w(\mathbf{x}) = 1, \forall \mathbf{x}$).

Using this basis, the approximation of the horizontal motion component $\mathcal{U}$ is:

$$\tilde{\mathcal{U}} = \sum_{k=0}^{D}\sum_{l=0}^{D-k} \tilde{u}_{k,l}\frac{P_{k,l}(\mathbf{x})}{\|P_{k,l}(\mathbf{x})\|} \tag{6}$$

The polynomial coefficients $\tilde{u}_{k,l}$ are given by the projection of component $\mathcal{U}$ onto normalized $\mathcal{B}$ elements:

$$\tilde{u}_{k,l} = \frac{\langle \mathcal{U} \mid P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l}(\mathbf{x})\|} \tag{7}$$

Similarly, vertical motion polynomial coefficients $\tilde{v}_{k,l}$ are given by computing the projection of vertical component $\mathcal{V}$ onto $\mathcal{B}$ elements. Using the polynomial basis $\mathcal{B}$ of degrre D, the vector field asociated to window $W(i,j,t)$ is modelled by $(D+1)\times(D+2)$ coefficients.

3.2 Time modeling using a polynomial basis

Since an action is performed along more than two frames, we propose to model motion information in longer space-time volumes.

Let us consider the descriptor located in frame at coordinates (i,j) and in video stream at time t_0. We consider the same spatial domain as previously defined (see Figure 1a). Moreover, we now consider the space-time tube defined by all the window $W(i,j,t_0)$ to $W(i,j,t_0+\tau)$, with τ being the length of our descriptors temporal domain (see Figure 1b). For each frame at time t between t_0 and $t_0+\tau$, we propose to model the vector field of motion inside the windows $W(i,j,t)$ of the tube by the coefficients $\tilde{u}_{k,l}$ and $\tilde{v}_{k,l}$, as defined in the previous section.

Then all coefficients $\tilde{u}_{k,l}(i,j,t)$ (respectively $\tilde{v}_{k,l}(i,j,t)$) for $t = t_0$ to $t = t_0+\tau$ are grouped in a vector defined as

$$\mathbf{u}_{k,l}(i,j,t_0) = [\tilde{u}_{k,l}(i,j,t_0), \ldots, \tilde{u}_{k,l}(i,j,t_0+\tau)] \tag{8}$$

We then model the time evolution of the coefficients $\tilde{u}_{k,l}(i,j,t)$ (resp. $\tilde{v}_{k,l}(i,j,t)$) by projecting $\mathbf{u}_{k,l}(i,j,t_0)$ (resp. $\mathbf{v}_{k,l}$) onto a one dimension orthogonal function basis. Here, we use Legendre polynomial basis of degree d defined by

$$\begin{cases} P_{-1}(t) = 0 \\ P_0(t) = 1 \\ T_n(t) = (t - \langle tP_{n-1}(t) | P_{n-1}(t)\rangle)P_{n-1}(t) - P_{n-2}(t) \\ P_n(t) = \dfrac{T_n(t)}{|T_n|} \end{cases} \tag{9}$$

Using this basis with degree d, the approximation of $\mathbf{u}_{k,l}(i,j,t)$ is:

$$\tilde{\mathbf{u}}_{k,l}(i,j,t) = \sum_{n=0}^{d} \tilde{u}_{k,l,n}(i,j,t) \frac{P_n(t)}{\|P_n(t)\|} \tag{10}$$

The model has $d+1$ coefficients $\tilde{\mathbf{u}}_{k,l}(i,j,t)$ given by

$$\tilde{u}_{k,l,n}(i,j,t) = \frac{\langle \mathbf{u}_{k,l}(i,j,t) \mid P_n(t) \rangle}{\|P_n(t)\|} \tag{11}$$

The time evolution of a given coefficient $\tilde{u}_{k,l}(i,j)$ (respectively $\tilde{v}_{k,l}(i,j)$) is given by the vector $\mathbf{m}_{l,k}(i,j,t_0)$ (respectively $\mathbf{n}_{l,k}(i,j,t_0)$) as defined in equation (12)

$$\mathbf{m}_{l,k}(i,j,t_0) = [\tilde{u}_{k,l,0}(i,j,t_0), \tilde{u}_{k,l,1}(i,j,t_0), \ldots, \tilde{u}_{k,l,d}(i,j,t_0)] \tag{12}$$

The feature descriptor $\nu(i,j,t_0)$ for the whole space-time volume beginning at time t_0 and centered at position (i,j) is given by

$$\begin{aligned}\nu(i,j,t_0) = [&\mathbf{m}_{0,0}, \mathbf{m}_{0,1}, \cdots, \mathbf{m}_{0,L}, \mathbf{m}_{1,0}, \cdots, \mathbf{m}_{1,L-1}, \cdots \\ &\cdots, \mathbf{m}_{K-1,0}, \mathbf{m}_{K-1,1}, \mathbf{m}_{K,0}, \mathbf{n}_{0,0}, \mathbf{n}_{0,1}, \cdots \\ &\cdots, \mathbf{n}_{0,L}, \mathbf{n}_{1,0}, \cdots, \mathbf{n}_{1,L-1}, \cdots \\ &\mathbf{n}_{K-1,0}, \mathbf{n}_{K-1,1}, \mathbf{n}_{K,0}]\end{aligned} \tag{13}$$

Here, $\mathbf{m}_{k,l}(i,j,t_0)$ and $\mathbf{n}_{k,l}(i,j,t_0)$ are written $\mathbf{m}_{k,l}$ and $\mathbf{n}_{k,l}$ to simplify.

The size of the descriptor $\nu_(i,j,t_0)$ is $(D+1) \times (D+2) \times d$.

3.3 Trajectories

As proposed in [12], we use trajectories to follow the spatial position of the window along time axis.

In our case the window $W(i_1, j_1, t_0+1)$ at time t_0+1 is selected as the best matching block with respect to the window $W(i_0, j_0, t_0)$ from time t_0. This matching is performed using a three step search block matching method from [5]. The temporal evolution of spatial coefficients is thus modeled on tubes instead of volumes.

4 Experiments

We carry out experiments[1] on two well known human action recognition datasets. The first one is the KTH dataset [9], and the second one is the Hollywood2 Human Actions dataset [6]. The degree of the spatial polynomial basis is set to 3 and the degree of time polynomial basis is set to 5. The spatial size of space-time volumes are set to 25×25 pixels and the length is set to 10. The spatial step for dense extraction is set to 5 pixels and the time step is set to 5 frames. We use a Horn and Schunk optical flow algorithm [3] for motion extraction with 25 iteration and the regularization λ parameter is set to 0.1. We extract the motion fields at 5 scales for KTH and 8 for Hollywood2, the scale factor is set to 0.8.

For experiments, we use VLAT indexing method to obtain signatures from SoPAF descriptors. We train a linear SVM for classification.

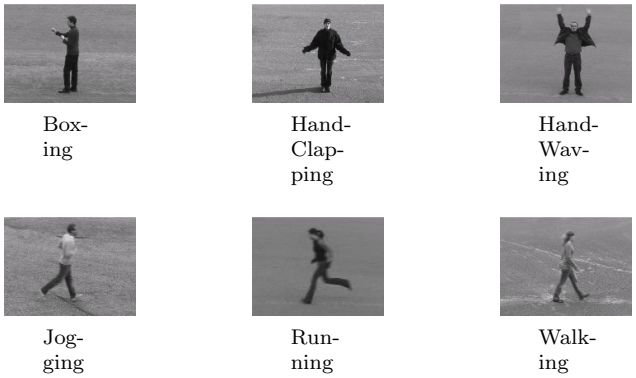


Figure 2: Example of videos from KTH

Figure 3: Example of videos from Hollywood2 dataset

4.1 KTH dataset

The KTH dataset [9] contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping (Figure 2). These actions are done by 25 different subjects in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, inside. For experiments, we use the same experimental setup as in [9, 12], where the videos are divided into a training set (8 persons), a validation set (8 persons) and a test set (9 persons).

For experiments on KTH dataset, the best hyper-parameters are selected through cross-validation using the official training and validation sets. The results were obtained on the test set.

4.2 Hollywood dataset

The Hollywood2 [6] dataset consists of a collection of video clips and extracts from 69 films in 12 classes of human actions (Figure 3). It accounts for approximately 20 hours of video and contains about 150 video samples per actions. It contains a variety of spatial

[1]Software can be downloaded at http://http://www.vlat.fr

scales, zoom camera, deleted scenes and compression artifact which allows a more realistic assessment of human actions classification methods. We use the official train and test splits for the evaluation.

4.3 Experimental results

In this section we show results on the two datasets. For each, we show results with our SoPAF descriptor alone and with a SoPAF and HOG descriptors combination. Let us note that our approach uses linear classifiers, and thus leads to better efficiency both for training classifiers and classifying video shots, on the contrary to methods [12] and [2].

On Table 1 we show the results obtained on KTH dataset, and compare them to recent results from the literature. We obtain good results only using the proposed SoPAF descriptors. When using SoPAF and HOG combination we obtain 94.1% multiclass accuracy, which is near state of the art performances while still using a linear classifier and combining less descriptors.

Table 1: Classification accuracy on the KTH dataset ; ND means the number of descriptors used ; NL stands for non-linear classifiers ; * In [2], the same feature is iteratively combined with itself 3 times

Method	ND	NL	Results
Wang [12]	4	X	94.2%
Gilbert [2]	$\simeq 3^\star$	X	94.5%
HOG	1		91.5%
SoPAF	1		93.4%
SoPAF+HOG	2		94.1%

On Table 2 we show results obtained on Hollywood2 dataset. With our SoPAF descriptor, we obtain results slightly better than the related HOF descriptors of [12] (and [2]) while using a linear classifier.

When combining SoPAF and HOG, we obtain a mAP of 55.6%, only second to the method proposed by [12]. However, contrarily to [12], we use only two descriptors and a linear classifier. When compared to the method proposed by Ullah et al., we obtain about the same results. However, we use only 2 signatures in contrast to over 100 BoW signatures accumulated over different regions in [11]. These regions where furthermore obtain from several detectors (*e.g.* Person, Action, Motion) trained on external data sets.

Table 2: Mean Average Precision on the Hollywood2 dataset ; ND : number of descriptors ; NL : non-linear classifiers ; * In [11] HOG/HOF descriptors are accumulated on over 100 spatio-temporal regions each one leading to a different BoW signature

Method	ND	NL	Results
Gilbert [2]	$\simeq 3$	X	50.9%
Ullah [11] HOG+HOF	2	X	51.8%
Ullah [11]	$2(\geq 100^\star)$	X	55.3%
Wang [12] traj	1	X	47.7%
Wang [12] HOG	1	X	41.5%
Wang [12] HOF	1	X	50.8%
Wang [12] MBH	1	X	54.2%
Wang [12] all	4	X	58.3%
HOG	**1**		**48.4%**
SoPAF	**1**		**52.2%**
SoPAF+HOG	**2**		**55.6%**

5 Conclusion

In this paper, we introduced a novel family of local motion descriptors using polynomial approximations of the optical flow and time evolution modeling.

For a given spatial window, after projecting the components of the optical flow on an orthogonal bivariate polynomial basis, we model the temporal evolution of spatial coefficients with one dimension polynomial basis. In order to model homogenous motion patterns, our space-time volumes follows trajectories of associated image patches.

We carry out experiments on the well known KTH and Hollywood2 datasets, using recent signatures method from image classification techniques. We obtain improved results over popular descriptors such as HOG and HOF, which highlight the soundness of the approach.

Further improvement would be to use this framework to model gradient field of images or optical flow as in HOG and MBH.

References

[1] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, pages 428–441, 2006.

[2] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *IEEE trans on PAMI*, (99):1–1, 2011.

[3] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1):185–203, 1981.

[4] O. Kihl, B. Tremblais, B. Augereau, and M. Khoudeir. Human activities discrimination with motion approximation in polynomial bases. In *ICIP*, pages 2469–2472. IEEE, 2010.

[5] T. KOGA. Motion-compensated interframe coding for video conferencing. *Proc. NTC, New Orleans*, 1981.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*. IEEE, 2008.

[7] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[8] D. Picard R. Negrel and P.H. Gosselin. Using spatial pyramids with compacted vlat for image categorization. In *ICPR*, 2012.

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

[10] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.

[11] M.M. Ullah, S. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, volume 2, page 7, 2010.

[12] H. Wang, A. Klaser, C. Schmid, and C.L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176. IEEE, 2011.