A UNIFIED FORMALISM FOR VIDEO DESCRIPTORS

Olivier Kihl, David Picard, Philippe-Henri Gosselin

A UNIFIED FORMALISM FOR VIDEO DESCRIPTORS

Olivier Kihl[⋆] *David Picard*[⋆] *Philippe-Henri Gosselin*[†]

[⋆] ETIS/ ENSEA - Université Cergy-Pontoise, CNRS, UMR 8051, France
[†]INRIA Rennes Bretagne Atlantique, France

ABSTRACT

In this paper, we propose a unified formalism for video descriptors. This formalism is based on the descriptors decomposition in three levels: primitive, scattering and projection. With this framework, we are able to rewrite easily all the usual descriptors in the literature such as HOG, HOF, SURF. Then, we propose a new projection method based on approximation with a finite expansion of orthogonal polynomials. Using our framework, we extend all usual descriptors by switching the projection step. The experiments are carried out on the well known KTH dataset and on the more challenging Hollywood2 action classification dataset and show state of the art results.

Index Terms— video analysis,video retrieval, classification, action analysis, local descriptors

1. INTRODUCTION

A popular way of comparing videos is done in three steps: extract a set of local descriptors from the video; find a transform that maps the set of descriptors into a single vector; compute the similarity between obtained vectors. Local feature descriptors have become essential tools in video action classification [1, 2, 3]. The main goal of such descriptors is to extract local properties of the signal. These properties are chosen to represent discriminative characteristic atoms of action. The descriptors are then aggregated into a signature which is used to train an action recognition classifier. Since local descriptors are the ground layer of action recognition systems, efficient descriptors are necessary to achieve good accuracies.

In this paper, we propose a unified formalism for descriptors that includes all the usual descriptors on the literature such as HOG, HOF, SURF. This formalism is based on the decomposition of the descriptor in three levels: primitive, separation and projection. From this formalism, we also propose a new family of projection. From this new projection and by combining primitive, separation and projection, we extend common descriptors. Here, descriptors are applied for actions classification.

The paper is organized as follows. In section 2, we present the most popular descriptors in the literature. Then, in section 3, we present our formalism and rewrite the most popular descriptors. In section 4, we present a new projection approach based on approximation with a finite expansion of orthogonal polynomials. Finally, in section 5, we carry out experiments on two well known action classification datasets for several descriptors and combinations of them.

2. RELATED WORK

In the past ten years, several descriptors have been proposed. The most commonly used are SIFT, Histogram of oriented gradient (HOG) [4], the Histogram of Oriented Flow (HOF) [4], SURF [5] and the Motion Boundary Histogram (MBH) [4]. SIFT and HOG descriptors relie on a histogram of orientation of gradient. Loccally, the orientation of the gradient is computed and associated to an orientation histogram bin (typically 8 or 9 bins). A HOG (or a SIFT) descriptor is composed of a grid of $M \times N$ histogram cells for a given spatial window. In the same way, Dalal et al. also propose the Histogram of Oriented Flow (HOF) [4] which is the same as HOG but applied to optical flow instead of the gradient. They also propose the Motion Boundary Histogram (MBH) that models the spatial derivatives of each component of the optical flow vector field with a HOG. Similarly to SIFT, SURF is composed of a point detector and a local descriptor. Here, we are interested only in the descriptor. The descriptor is composed with a grid of $M \times N$ cell, each composed of a four-component vector, computed by summing the horizontal (dx) or the vertical (dy) Haar responses in the cell and the absolute value of dx and dy. A similar idea has been proposed by Efros et al. [6] to model motion. They decomposed the horizontal ($\mathcal{U}$) and vertical ($\mathcal{V}$) components of a vector field (usually obtained by optical flow approaches) with a technique of half-wave rectification :

$$\mathcal{U}^+(\vec{x}) = \begin{cases} \mathcal{U}(\vec{x}) & \text{if } \mathcal{U}(\vec{x}) > 0 \\ 0 & \text{else} \end{cases} \quad (1)$$

$$\mathcal{U}^-(\vec{x}) = \begin{cases} \mathcal{U}(\vec{x}) & \text{if } \mathcal{U}(\vec{x}) < 0 \\ 0 & \text{else} \end{cases} \quad (2)$$

Similarly, from $\mathcal{V}$, they obtain $\mathcal{V}^+$ and $\mathcal{V}^-$.

Recently, Wang et al [1] propose to model these usual descriptors along dense trajectories. The time evolution of trajectories, HOG, HOF and MBH is modelled using a space time grid following pixels trajectories. To our knowledge, they obtained state of the art results.

3. PRIMITIVE/SEPARATION/PROJECTION FORMALISM

In this section, we present our formalism for local descriptors. Our formalism splits a descriptor in three levels : primitive, separation and projection.

The primitive extract from the video the information to model. It can be the gradient (HOG), the motion (HOF), or the gradient of motion (MBH). The objective is to extract local properties of the video. Generally, it relies on a high frequency filtering, linear for gradient or non-linear in the case of motion (optical flow), filters banks such as Haar (SURF), easy extention of popular filters [7], or non-linear operators.

The separation transform corresponds to a non-linear mapping of the primitive to a higher dimensional space. The objective is to improve the projection step by grouping together the primitive properties that are similar. In the literature, the primitives are separated into orientation bins (HOG, HOF and MBH) or the rectified (or double rectified) components (SURF).

Primitive	Separation	Projection
gradient	raw	cells
motion	rectified	polynomial basis
Haar	abs	sine basis
motion gradient	orientation	wavelets
$\vdots$	$\vdots$	$\vdots$

Table 1: A new formalism for actions descriptors

Name	Primitive	Separation	Projection
HOG	gradient	orientations	cells
HOF	motion	orientations	cells
MBH	motion gradient	orientations	cells
SURF	Haar	abs	cells
Efros	motion	rectified	cells

Table 2: Rewriting of the usual descriptors

Finally, the projection is used to model the separated primitives. Currently, the descriptors of literature (HOG, HOF, MBH, SURF) use a grid of $N \times N$ cells. In this paper, we propose a new projection based on polynomials, but other basis (Sine for example) can be considered. Table 1 summarizes the above proposals. In Table 2, we write the usual descriptors of the literature with our formalism. Currently, all the usual descriptors use cells as projection.

4. POLYNOMIAL BASED PROJECTION

Based on our new formalism, we propose to model the separated primitive by a finite expansion of orthogonal polynomials. Let us define the family of polynomial functions with two real variables as follows:

$$P_{K,L}(x_1, x_2) = \sum_{k=0}^{K} \sum_{l=0}^{L} a_{k,l}\, x_1^k\, x_2^l \tag{3}$$

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:

$$\begin{aligned}\mathbb{B} = \quad & \{B_{0,0}, B_{0,1}, \cdots, B_{0,L}, B_{1,0}, \cdots \\ & \cdots, B_{1,L-1}, \cdots, B_{K-1,0}, B_{K-1,1}, B_{K,0}\}\end{aligned} \tag{4}$$

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

$$\begin{cases} B_{-1,l}(\mathbf{x}) = 0 \\ B_{k,-1}(\mathbf{x}) = 0 \\ B_{0,0}(\mathbf{x}) = 1 \\ B_{k+1,l}(\mathbf{x}) = (x_1 - \lambda_{k+1,l})B_{k,l}(\mathbf{x}) - \mu_{k+1,1}B_{k-1,l}(\mathbf{x}) \\ B_{k,l+1}(\mathbf{x}) = (x_2 - \lambda_{k,l+1})B_{k,l}(\mathbf{x}) - \mu_{k,l+1}B_{k,l-1}(\mathbf{x}) \end{cases} \tag{5}$$

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

$$\begin{aligned} \lambda_{k+1,l} &= \frac{\langle x_1 B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l}(\mathbf{x})\|^2} \quad & \lambda_{k,l+1} &= \frac{\langle x_2 B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l}(\mathbf{x})\|^2} \\ \mu_{k+1,l} &= \frac{\langle B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k-1,l}(\mathbf{x})\|^2} \quad & \mu_{k,l+1} &= \frac{\langle B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l-1}(\mathbf{x})\|^2} \end{aligned} \tag{6}$$

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

$$\langle B_1 \mid B_2 \rangle = \iint_\Omega B_1(\mathbf{x}) B_2(\mathbf{x}) w(\mathbf{x}) d\mathbf{x} \tag{7}$$

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 a decomposed primitive component $\mathcal{P}$ is:

$$\tilde{\mathcal{P}} = \sum_{k=0}^{D} \sum_{l=0}^{D-k} \tilde{u}_{k,l} \frac{B_{k,l}(\mathbf{x})}{\|B_{k,l}(\mathbf{x})\|} \tag{8}$$

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

$$\tilde{p}_{k,l} = \frac{\langle \mathcal{P} \mid B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l}(\mathbf{x})\|} \tag{9}$$

Since an action is performed along more than one frame, we propose to model information along time axis. For HOG, HOF and MBH, the space grid is extended to a space-time grid. Hence we propose to model spatial polynomial coefficients with a temporal basis of polynomials.

Here, we use Legendre polynomial basis of degree d defined by

$$\begin{cases} B_{-1}(t) = 0 \\ B_0(t) = 1 \\ T_n(t) = (t - \langle tB_{n-1}(t) | B_{n-1}(t)\rangle)B_{n-1}(t) - B_{n-2}(t) \\ B_n(t) = \dfrac{T_n(t)}{|T_n|} \end{cases} \tag{10}$$

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

$$\tilde{\mathbf{p}}_{k,l}(i, j, t) = \sum_{n=0}^{d} \tilde{p}_{k,l,n}(i, j, t) \frac{B_n(t)}{\|B_n(t)\|} \tag{11}$$

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

$$\tilde{p}_{k,l,n}(i, j, t) = \frac{\langle \mathbf{p}_{k,l}(i, j, t) \mid B_n(t)\rangle}{\|B_n(t)\|} \tag{12}$$

The time evolution of a given coefficient $\tilde{p}_{k,l}(i, j)$ is given by the vector $\mathbf{m}_{l,k}(i, j, t_0)$ as defined in equation (13)

$$\begin{aligned}\mathbf{m}_{l,k}(i, j, t_0) = [&\tilde{p}_{k,l,0}(i, j, t_0), \tilde{p}_{k,l,1}(i, j, t_0), \\ &\ldots, \tilde{p}_{k,l,d}(i, j, t_0)]\end{aligned} \tag{13}$$

Finally, the descriptor is the concatenation of all the D $\mathbf{m}_{l,k}(i, j, t_0)$ vectors. The size of this descriptor is $\frac{(D+1)\times(D+2)}{2} \times (d + 1) \times np$ with np the number of primitives.

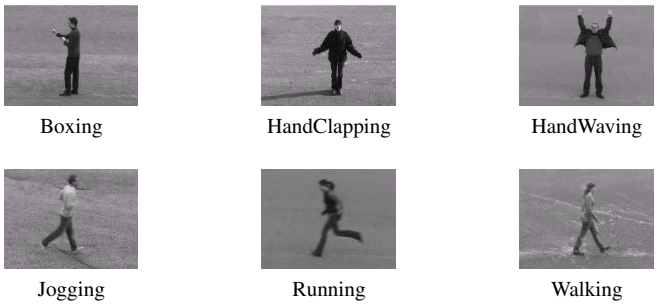

Fig. 1: Example of videos from KTH

5. EXPERIMENTS

Using our framework, we compare several combination of primitives, separation and projections in order to evaluate action descriptors. We compare two primitives (gradient and motion), three separation methods (raw, rectified and orientations) and two projection methods (cells and polynomials). We carry out experiments on two well known human action recognition datasets : KTH dataset [3] and Hollywood2 Human Actions dataset [2].

For motion estimation, we use a Horn and Schunk optical flow algorithm [8] with 25 iteration and the regularization λ parameter is set to 0.1. We extract the gradient with the simple one order approximation difference method. We extract the gradient and motion fields at 1 scales for KTH and 7 scales for Hollywood2, where the scale factor is set to 0.8.

For the experiments, we obtain signatures from our descriptors by using the VLAT indexing method [9] 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 statistic deviations from a given visual codebook. 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. We train a linear SVM for classification.

5.1. KTH dataset

The KTH dataset [3] contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping (Figure 1). 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 [3, 1], 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.

5.2. Hollywood dataset

The Hollywood2 [2] dataset consists of a collection of video clips and extracts from 69 films in 12 classes of human actions (Figure 2). It accounts for approximately 20 hours of video and contains about 150 video samples per actions. It contains a variety of spatial 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.

Fig. 2: Example of videos from Hollywood2 dataset

5.3. Experimental results

First, we present in Table 3 the results obtained with several descriptors based on the gradient or motion primitives on KTH dataset. The results show that the classification accuracy increases with the size of the descriptors. We remark the cell projection provides better results than polynomial projection for all the separation methods proposed here when the primitive is the gradient. Table 3 is vertically split in three parts, for highlighting small, medium and large dimensional descriptors. One can see that the gradient primitive needs relatively larger descriptors than the motion primitive which in turn provides good accuracy even with small descriptors.

Then, we present in Table 4 the results of several combinations of Gradient-Motion descriptors on KTH. We show the best descriptor results of our study on KTH dataset, and compare them to recent results from the literature. Let us note that our approach uses linear classifiers, and thus leads to better efficiency both for training classifiers and classifying video shots, as opposed to methods [1] and [10]. We obtain good results even with only one descriptor. When using $A + B$ combination we obtain 94.2% multiclass accuracy, which is near state of the art performance while still using a linear classifier and combining less descriptors.

Then, we select the best setup according to gradient primitive associated with cells and polynomials projections and the best setup according to Motion primitive associated with cells and polynomials projections (c.f. Table 3). These setups are evaluated on the Hollywood2 dataset and results are reported in Table 5. One can see the results presented here are equivalent to state of the art for single descriptor setups when comparing to HOG (gradient primitive) and to HOF (motion primitive). Note that we do not use the dense trajectories as in [1] to obtain these results. On this more challenging dataset, the results obtained for Gradient primitive are better for polynomials projections than cells projections. Finally, by combining two primitives, we obtain results close to the state of the art.

The results obtained on the challenging Hollywood2 dataset with the combination of several descriptors, including the new family we proposed in section 4, highlight the soundness of our framework.

dim	dec	Gradient		Flow		SP	TP
		Cell	Poly	Cell	Poly		
24	raw	82.5		88.3		2	3
30	raw		76.2		87.3	4	0
32	raw	80.4		87.0		4	1
36	raw		81.0		**89.8**	2	2
40	raw	**82.8**		89.6		2	5
48	rect	84.8		90.7		2	3
60	rect		83.2		90.7	4	0
64	raw		84.5		90.4	2	4
64	rect	**86.5**		90.4		4	1
72	rect		84.5		90.5	2	2
80	raw		83.1		**91.1**	3	3
80	rect	87.2		91.4		2	5
96	ori	92.4		89.2		2	3
120	ori		92.6		90.6	4	0
128	rect		88.0		91.7	2	4
128	ori	**93.4**		91.8		4	1
144	ori		92.8		91.1	2	2
144	rect	88.5		**92.0**		3	4

Table 3: Results for combination of primitives, separation and projections ; dim means the dimension of the descriptor ; dec represent the separation method (raw, rectified or orientation) ; SP means the number of spatial cells for Cells projections and the degree D of spatial polynomials for Polynomials projections ; TP means the number of temporal cells for Cells projection and the degree d of temporal polynomials for Polynomials projections

Method	ND	NL	Results
Wang [1]	4	X	94.2%
Gilbert [10]	$\simeq 3^{\star}$	X	94.5%
A = Gradient + ori + Cell (4,1)	1		93.4%
B = Flow + rect + Cell (3,4)	1		92.0%
A+B	2		94.2%

Table 4: Classification accuracy on the KTH dataset ; ND means the number of descriptors used ; NL stands for non-linear classifiers ; * In [10], the same feature is iteratively combined with itself 3 times

6. CONCLUSION

In this paper, we introduced a new formalism to describe video descriptors. This formalism consists on the decomposition of descriptors in three levels : primitive, separation and projection. Our formalism allows us to easily rewrite all the descriptors of the literature.

We propose a new projection approach based on approximation with a finite expansion of orthogonal polynomials, which in turns leads to a new family of descriptors.

We experimented several combination of primitive, separation and projection on two human action recognition datasets. We obtain better or equivalent results for than the usual descriptors of literature. This confirms the validity and relevance of formalism to create new descriptors.

However, combinatorial related to the number of Primitives / Separation / Projections makes impossible the exploration of all these parameters. Our future works will concern introduction of learning processes in the three levels proposed.

Method	ND	NL	Results
Gilbert [10]	$\simeq 3$	X	50.9%
Ullah [11] HOG+HOF	2	X	51.8%
Ullah [11]	$2(\geq 100^{\star})$	X	55.3%
Wang [1] traj	1	X	47.7%
Wang [1] HOG	1	X	41.5%
Wang [1] HOF	1	X	50.8%
Wang [1] MBH	1	X	54.2%
Wang [1] all	4	X	58.3%
A = Grad + Ori + Cell (4,1)	1		45.2%
B = Flow + rect + Cell (3,4)	1		53.5%
C = Grad + Ori + Poly (2,2)	1		50.0%
D = Flow + rect + Poly(2,4)	1		52.8%
A + B	2		57.4%
C + D	2		57.6%

Table 5: 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

7. REFERENCES

[1] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," in *CVPR*. IEEE, 2011, pp. 3169–3176.

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

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *ICPR*. IEEE, 2004, vol. 3, pp. 32–36.

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

[5] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[6] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE ICCV*, 2003, vol. 2, pp. 726–733.

[7] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, pp. 2032–2047, 2009.

[8] B.K.P. Horn and B.G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[9] R. Negrel, D. Picard, and P. Gosselin, "Compact tensor based image representation for similarity search," in *International Conference on Image Processing*, 2012.

[10] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE trans on PAMI*, , no. 99, pp. 1–1, 2011.

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