A Unified framework for local visual descriptors evaluation

Olivier Kihl, David Picard, Philippe-Henri Gosselin

A Unified framework for local visual descriptors

Olivier Kihl[a,*], David Picard[a], Philippe-Henri Gosselin[a,b]

[a] *ETIS/ ENSEA - Université Cergy-Pontoise, CNRS, UMR 8051*
6 avenue du Ponceau, CS 20707 CERGY, F 95014 Cergy-Pontoise Cedex France
Telephone:+33 1 30 73 66 10 and Fax:+33 1 30 73 66 27
[b] *INRIA Rennes Bretagne Atlantique*
Campus de Beaulieu, 35042 Rennes Cedex France

Abstract

Local descriptors are the ground layer of recognition feature based systems for still images and video. We propose a new framework to explain local descriptors. This framework is based on the descriptors decomposition in three levels: primitive extraction, primitive coding and code aggregation. With this framework, we are able to explain most of the popular descriptors in the literature such as HOG, HOF, SURF. We propose two new projection methods based on approximation with oscillating functions basis (sinus and Legendre polynomials). Using our framework, we are able to extend usual descriptors by changing the code aggregation or adding new primitive coding method. The experiments are carried out on images (VOC 2007) and videos datasets (KTH, Hollywood2 and UCF11), and achieve equal or better performances than the literature.

Keywords: Image Processing and Computer Vision, Vision and Scene Understanding, Video analysis, Image/video retrieval retrieval, Object recognition, Feature representation

1. Introduction

Most multimedia retrieval systems compare multimedia documents (image or video) thanks to three main stages: extract a set of local visual descriptors from the multimedia document; learn a mapping of the set of descriptors into a single vector to obtain a signature; compute the similarity between signatures. In this paper, we focus on the computation of visual descriptors. The main goal of local visual descriptors is to extract local properties of the signal. These properties are chosen to represent discriminative characteristic atoms of images or videos. Since local descriptors are the ground layer of recognition systems, efficient descriptors are necessary to achieve good accuracies. Such descriptors have become essential tools in still image classification [1, 2] and video action classification [3, 4, 5].

The main contribution of this paper is a unified framework for visual descriptors that includes all the usual descriptors from the literature such as SIFT (Scale-invariant feature transform) [6], SURF (Speeded Up Robust Features) [7], HOG (Histogram of Oriented Gradient) [8], HOF (Histogram of Oriented Flow) and MBH (Motion Boundary Histogram) [9]. This framework is based on the decomposition of the descriptor in three levels: primitive extraction, primitive coding and code aggregation. Each popular descriptor is composed by a given primitive, a given coding and a given aggregation. Moreover, our framework allows to extend every descriptor by changing one or more of the three levels, e.g. changing the primitive level of HOG (gradient) by the motion primitive produces the HOF descriptor. The second contribution of this paper is the proposal of new coding and aggregation steps, the later being based on oscillating functions (Sinus and Polynomials), leading to new descriptors. Finally, we propose an exploration of the possible combinations of these primitives, coding and aggregation methods provided by the framework, that allows us to design more efficient and complementary descriptors.

The paper is organized as follows. In section 2, we present the most popular descriptors in the literature, for still images and for human action videos. We also present the most common approaches to compute the signature of a multimedia document from a set of descriptors. Then, in section 3, we present our framework, explain the most popular descriptors, and extend them by modifying some of these three steps. Finally, in section 4, we carry out experiments on one still image classification dataset and three action classification datasets for several descriptors and combinations of them.

2. Related work

In this section, we present the most popular descriptors in the literature, first for still image and then for human action video. We also present the most common approaches to compute the signature of a multimedia document from a set of descriptors.

*Corresponding author
Email addresses: olivier.kihl@ensea.fr (Olivier Kihl), picard@ensea.fr (David Picard), philippe-henri.gosselin@ensea.fr (Philippe-Henri Gosselin)

2.1. Still image descriptors

In the past ten years, several descriptors have been proposed for key-points matching and successfully used for still image classification. The most commonly used are SIFT [6], SURF [7] and Histogram of oriented gradient (HOG) [9]. SIFT and SURF are both interest points detector and local image descriptor. In this paper, we only consider the descriptors. SIFT and HOG descriptors rely on a histogram of orientation of gradient. Locally, the orientation of the gradient is quantized in o orientations (typically 8). For a given spatial window, a HOG (or a SIFT) descriptor is computed by decomposing the window with a grid of $N \times N$ cells. Each cell contains the histogram of orientations of the gradient. The descriptor is obtained by the concatenation of the $N \times N$ histograms.

The SURF [7] descriptor has been developed as a faster alternative to SIFT. In the case of SURF, descriptors of each cells are computed using weighted sum of responses to 2D Haar-wavelets along horizontal axis (dx) and vertical axis (dy), the absolute value of dx and the absolute value of dy.

More recently, new descriptors have been proposed with the aim to decrease the computation time without loss of performance. The GLOH [10] is an extension of the SIFT, in which the rectangular grid is replaced by a polar grid. The authors propose to use 3 bins in radial direction and 8 in angular direction. The gradient orientation is quantized in 16 bins inside each cell. To reduce the dimension of the descriptor, a principle components analysis (PCA), computed on several GLOH, is applied. Similarly, Daisy [11] is a SIFT like descriptor designed to be faster to compute in the case of dense matching extraction. The sum in histogram cells are replaced by computing the convolution of the orientation maps with Gaussian kernels. Moreover, the sampling positions of the descriptor are not aligned with a rectangular grid, but in concentric circles at several distances to the descriptor center. For a given distance, the sample points are associated to a particular Gaussian kernel size, increasing with the distance to the center.

2.2. Action descriptors

In the early work on action recognition, silhouette based descriptors, also called motion appearance models were used. These descriptors are computed from the evolution of a silhouette obtained by background subtraction methods or by taking the difference of frames (DOF). From a sequence of binary images, Bobick and Davis [12] propose descriptors called "Motion Energy Image" (MEI) representative of the energy of movement and "Motion History Image" (MHI) providing information about the chronology of motion. These two descriptors are modeled by seven Hu moments. In [13] Kellokumpu et al. use histograms of "Local Binary Patterns" (LBP) [14] to model the MHI and MEI images. In [15], they propose an extension of the LBP directly applied on the image pixels with successful results. Wang and Suter [16] use two other descriptors, namely the "Average Motion Energy" (AME) and the "Mean Motion Shape" (MMS). The AME is a descriptor close to the MHI representing the average image of silhouettes. The MMS is defined from boundary points of the silhouette in complex coordinates with the origin placed at the centroid of the 2D shape. As time is an important information in video, Gorelick et al. [17, 18] study the silhouettes as space-time volumes. Space-time volumes are modeled with Poisson equations. From these, they extract seven spatio-temporal characteristic components.

The main drawback of all these methods is the computation of silhouettes. Indeed, this computation is not very robust, making these methods only relevant in controlled environments such as the Weizmann dataset [17] and the KTH dataset [5]. However, they tend to fail on more realistic data-sets such as UCF11 [19] or Hollywood2 [4] datasets.

Assuming that action recognition is closely linked to the notion of movement, many authors have proposed descriptors based on the modeling of optical flow. The optical flow represents the displacement of pixels from two consecutive frames. The result can be represented by vector field with two components. Here, $\mathcal{U}$ denotes the horizontal component of motion and $\mathcal{V}$ the vertical component. Early works with respect to this approach were proposed by Polana and Nelson [20]. The vector field is first decomposed according to a spatial grid. Then, in each cell of the grid, the magnitude of motion is accumulated. This method can only process periodic actions such as running or walking.

Efros et al. [21] propose a descriptor computed on a figure-centric spatio-temporal volume for each person in a video. The vector field representing the motion between two consecutive frames of the volume is computed with the Lucas and Kanade optical flow algorithm [22]. The two components $\mathcal{U}$ and $\mathcal{V}$ of the vector field are decomposed with a half-wave rectification technique. The resulting four components are blurred using a Gaussian filter and normalized. They are directly used as a descriptor. The obtained descriptors are compared using the normalized correlation measure. This descriptor is used and/or extended by several authors in [23, 24].

Tran et al. propose the motion context descriptor in [25]. It is also a figure-centric descriptor based on the silhouette extraction. They use the vector field and the binary silhouette as three components. The components of the field are blurred with a median filter. Then, the three components are subdivided with a grid of 2×2 cells. Each cell is decomposed in 18 radial bins, each covering 20 degrees. Inside the radial bins, the sum of each component is computed. This provides, for each component, 4 histograms composed with 18 bins. The concatenation of these histograms provides a 216-dimensional vector which is the movement pattern of a given field. From this pattern, the "Motion Context" is created. It is composed of the 216-dimensional vector of the current frame plus the first 10 vectors of the PCA models of the 5 previous frames, the first 50 vectors of the PCA models of 5 current frames and finally the first 10 vectors PCA models of 5 next frames.

Ali and Shah [26] begin by computing many kinematic features on the field and then compute kinematic modes with a spatio-temporal principal component analysis.

Finally, the most successful descriptors developed in recent years are extensions to video of still image descriptors. The most commonly used are the Histogram of Oriented Flow (HOF) [9] and the Motion Boundary Histogram (MBH) [9].

HOF is the same as HOG but is applied to optical flow instead of gradient. The MBH models the spatial derivatives of each component of the optical flow vector field with a HOG.

In this context, several extension of still image descriptors have been proposed. The cuboid [27] is a space-time descriptor, represented by a space-time volume. For a given volume, the gradient is computed on the three directions and the descriptor are the flattening of the gradient in a vector. Consistent with this, Klaser et al. [28] extend HOG to 3DHOG. A 3-dimensional extention of SIFT is proposed in [29]. ESURF [30] is an extention of SURF with 3D Haar-wavelets.

Descriptors based on a polynomial approach for modeling global optical flow are proposed in [31] and [32]. From this preliminary works, a local descriptor for motion named Series of Polynomial approximation of Flow (SoPAF) is proposed in [33]. The descriptor is based on two local modeling, a spatial model and a temporal model. The spatial model is computed by the projection of the optical flow onto bivariate orthogonal polynomials. Then, the time evolution of spatial coefficients is modelled with a one dimension polynomial basis.

Recently, Wang et al. [3] 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. The use of dense trajectories for descriptor extraction tends to increase the performances of popular descriptors (HOG, HOF and MBH).

2.3. Signatures

Once a set of descriptors is obtained from the video, a popular way of comparing images (or videos) is to map the set of descriptors into a single vector and then to measure the similarity between the obtained vectors (for example in [34], [35] and [3]). The most common method for such embeddings is inspired by the text retrieval community and is called the "Bag of Words" (BoW) approach [36]. It consists in computing a dictionary of descriptor prototypes (usually by clustering a large number of descriptors) and then computing the histogram of occurrences of these prototypes (called "Visual Words") within the set.

In still images classification, these approaches have been formalized in [37] by a decomposition of the mapping into two steps. The first step, namely the "coding step", consists in mapping each descriptor into a codeword using the aforementioned dictionary. The second step is to aggregate the codewords into a single vector and is called the "pooling step". Structural constraints such as sparsity [38] or locality [37] can be added to the coding process to ensure most of the information is retained during the pooling step. Common pooling processes include averaging the codewords or retaining the entry-wise maximum among the codewords (max pooling). Extensions of the BoW model have been recently proposed to include more precise statistical information. In [39], the authors propose to model the distribution of distances of descriptors to the clusters centers. In the "coding/pooling" framework, each descriptor is coded by 1 in the bin corresponding to its distance to the cluster's center to which it belongs, and 0 otherwise. The polling is simply the averaging over all codewords.

In [40], the authors proposed a coding process where the deviation between the mean of the descriptors of the set and the center of the cluster to which they belong to is computed. The whole mapping process can be seen as the deviation between a universal model i.e. the dictionary) and a local realization (i.e. the set of descriptors). Using this model deviation approach, higher order statistics have been proposed, like "super-vectors" in [41], "Fisher Vectors" in [42] or "VLAT" in [43, 44]. Fisher Vectors are known to achieve state of the art performances in image classification challenges [2].

To compare the performances of descriptors, in this paper, we consider a compressed version of VLAT which is known to achieve near state of the art performances in still images classification with very large sets of descriptors [45]. In our case, the dense sampling both in spatial and temporal directions leads to highly populated sets, which is consistent with the statistics computed in VLAT signatures. Given a clustering of the descriptors space with C clusters computed on some training set, the first and second order moments μ_c and τ_c are computed for each cluster c:

$$\mu_c = \frac{1}{|c|}\sum_i\sum_r \nu_{rci} \tag{1}$$

$$\tau_c = \frac{1}{|c|}\sum_i\sum_r (\nu_{rci}-\mu_c)(\nu_{rci}-\mu_c)^T \tag{2}$$

with ν_{rci} the descriptor r of the video i which is in cluster c, and $|c|$ being the number of descriptors ν_{rci} of video i in cluster c, for all videos in the training set. To allow a dimension reduction of the signature, the eigen decomposition of the covariance matrix τ_c for each cluster c is then performed:

$$\tau_c = \mathbf{V}_c\mathbf{D}_c\mathbf{V}_c^\top \tag{3}$$

Using this decomposition, descriptors are projected on the subspace generated by the eigenvectors V_c. The compressed VLAT signature $\tau_{i,c}$ of video i is computed for each cluster c with the following equation:

$$\tau_{i,c} = \sum_r (\mathbf{V}_c(\nu_{rci}-\mu_c))(\mathbf{V}_c(\nu_{rci}-\mu_c))^\top - \mathbf{D}_c \tag{4}$$

$\tau_{i,c}$ are then flattened into vectors $\mathbf{v}_{i,c}$. The complete VLAT signature $\mathbf{x}_i$ of video i is obtained by concatenation of $\mathbf{v}_{i,c}$ for all clusters c:

$$\mathbf{v}_i = (v_{i,1}\dots v_{i,C}) \tag{5}$$

It is advisable to perform a normalization step for best performance.

$$\forall j,\quad \mathbf{v}'_i[j] = sign(\mathbf{v}_i[j])|\mathbf{v}_i[j]|^\alpha, \tag{6}$$

$$\mathbf{x}_i = \frac{\mathbf{v}'_i}{\|\mathbf{v}'_i\|} \tag{7}$$

With $\alpha = 0.5$ typically. The size of the compacted VLAT signature depends on the number d_c of eigenvectors retained in each cluster, and is equal to $\sum_c \frac{d_c(d_c+1)}{2}$ (thanks to the matrices $\tau_{i,c}$ being symmetric, only half of the coefficients are kept).

3. Primitive/Coding/Aggregation framework

In this section, we present the main contribution of this paper. We propose a framework providing a formal description of the steps needed to design local visual descriptors. Our framework splits descriptors extraction in three levels: primitive extraction, primitive coding and code aggregation.

3.1. Primitive extraction

At the primitive level, we extract a specific type of low-level information from an image or a video. Such primitives include the gradient (HOG), the responses to 2D Haar-wavelets (SURF), the motion flow (HOF), or the gradient of motion flow (MBH). In Fig. 1, we show three examples of primitive used in literature, the gradient, the motion flow and the gradient of the motion flow. The objective is to extract local properties of the signal. 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 extension of popular filters [46], or non-linear operators. The primitive extraction induces a choice in relevant information and introduces data loss.

3.2. Primitive coding

The primitive coding corresponds to a non-linear mapping of the primitive to a higher dimensional space. The objective is to improve the representation by grouping together the primitive properties that are similar.

In the literature, the most popular primitive coding is the quantization of local vector field orientations [6, 8]. The quantization is usually performed on 8 bins. Let $G_x(\mathbf{x})$,$G_y(\mathbf{x})$ be the horizontal and vertical derivative of an image at position $\mathbf{x}$, the principal orientation bin is computed by:

$$o(\mathbf{x}) = \lfloor \frac{(atan2(G_y(\mathbf{x}), G_x(\mathbf{x})) \mod 2\pi) \times 4}{\pi} \rfloor \quad (8)$$

In order to limit the effect of floor on coding, the distance to the next orientation bin is computed by

$$r(\mathbf{x}) = o(\mathbf{x}) - \left\lfloor \frac{(atan2(G_y(\mathbf{x}), G_x(\mathbf{x})) \bmod 2\pi) \times 4}{\pi} \right\rfloor \quad (9)$$

The value associated to the bin $o(\mathbf{x})$ and $o(\mathbf{x}) + 1$ are

$$\mathcal{O}(\mathbf{x}, o(\mathbf{x})) = \rho(\mathbf{x}) \times (1 - r(\mathbf{x})) \quad (10)$$

$$\mathcal{O}(\mathbf{x}, (o(\mathbf{x}) + 1) \bmod 8) = \rho(\mathbf{x}) \times r(\mathbf{x}) \quad (11)$$

with $\rho(\mathbf{x})$ the magnitude of horizontal and vertical derivative ($\rho(\mathbf{x}) = \sqrt{G_x(\mathbf{x})^2 + G_y(\mathbf{x})^2}$). This primitive coding do not introduce any loss of information or redundancy.

Another primitive coding is proposed in SURF [7]. Here, we call it "absolute coding". In the SURF descriptor, it is applied to the gradient primitive. This is a four dimension code defined as:

$$\mathcal{A}(\mathbf{x}, 0) = G_x(\mathbf{x}) \quad (12)$$

$$\mathcal{A}(\mathbf{x}, 1) = G_y(\mathbf{x}) \quad (13)$$

$$\mathcal{A}(\mathbf{x}, 2) = |G_x(\mathbf{x})| \quad (14)$$

$$\mathcal{A}(\mathbf{x}, 3) = |G_y(\mathbf{x})| \quad (15)$$

This primitive coding introduces redundancy. However, it produces lower dimensions code than orientation coding.

In the context of action recognition and classification, the rectified coding proposed by Efros et al. . [21] has been used by several authors. They decompose 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:

$$\mathcal{R}(\mathbf{x}, 0) = \begin{cases} \mathcal{U}(\mathbf{x}) & \text{if } \mathcal{U}(\mathbf{x}) > 0 \\ 0 & \text{else} \end{cases} \quad (16)$$

$$\mathcal{R}(\mathbf{x}, 1) = \begin{cases} |\mathcal{U}(\mathbf{x})| & \text{if } \mathcal{U}(\mathbf{x}) < 0 \\ 0 & \text{else} \end{cases} \quad (17)$$

$$\mathcal{R}(\mathbf{x}, 2) = \begin{cases} \mathcal{V}(\mathbf{x}) & \text{if } \mathcal{V}(\mathbf{x}) > 0 \\ 0 & \text{else} \end{cases} \quad (18)$$

$$\mathcal{R}(\mathbf{x}, 3) = \begin{cases} |\mathcal{V}(\mathbf{x})| & \text{if } \mathcal{V}(\mathbf{x}) < 0 \\ 0 & \text{else} \end{cases} \quad (19)$$

Orientation coding, absolute coding and rectified coding are the most used in literature.

We also propose a new primitive coding called double rectified coding. This coding corresponds to the 4 components of the rectified coding and the 4 components of the absolute coding. Examples of these primitive coding are shown in Fig. 2.

3.3. Code Aggregation

Finally, the code aggregation is used to model the encoded primitives. The objective of aggregation is to improve the robustness to deformation by allowing inexact matching between deformed image or video patches. Most descriptors from the literature (HOG, HOF, MBH, SURF) use accumulation of each primitive coding (typically with a simple sum). In order to improve robustness, the accumulation is done on the cell of a grid of $N \times N$ cells. In the case of video, the grid could be extended in $N \times N \times T$ cells with T the number of cell bins in time direction. The spatial window could be pondered by a Gaussian to give more importance to the cells which are close to the center, like in SIFT. We show a 4×4 cell aggregation in Fig. 3a.

The regular grid can replaced with concentric circles arranged in a polar manner, as it is proposed in DAISY [11]. The final pattern resembles a flower, and is shown in Fig. 3b. In the following, we name this code aggregation "Flower". The flower aggregation is defined by three parameters R, Q and T. The radius R defines the distance from the center pixel to the outer most grid point. The quantization of the radius Q defines the number of convolved primitives layer associated to different size of Gaussian ($Q = 3$ in Fig. 3b). The parameter T defines the angular quantization of the pattern at each layer ($T = 8$ in Fig. 3b).

The aggregation proposed in [31] is based on the projection of primitive on a two dimensional orthogonal polynomial basis. The family of polynomial functions with two real variables is defined as follows:

$$P_{K,L}(x_1, x_2) = \sum_{k=0}^{K} \sum_{l=0}^{L} a_{k,l}\, x_1^k\, x_2^l \quad (20)$$

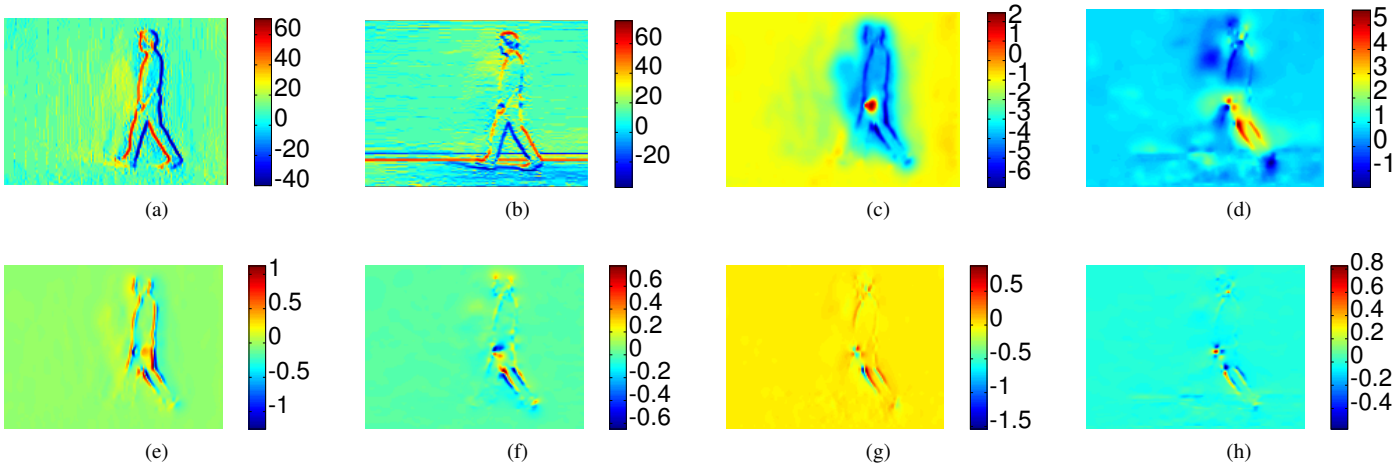


Figure 1: Example of primitive ; (a) Horizontal gradient ; (b) Vertical gradient ; (c) horizontal motion flow ; (d) Vertical motion flow ; (e) Horizontal gradient of horizontal motion flow ; (f) Vertical gradient of horizontal motion flow ; (g) Horizontal gradient of vertical motion flow ; (h) Vertical gradient of vertical motion flow

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:

$$\begin{aligned}\mathbb{B} = \quad & \{B_{0,0}, B_{0,1}, \cdots, B_{0,L}, B_{1,0}, \cdots \\ & \cdots, B_{1,L-1}, \cdots, B_{K-1,0}, B_{K-1,1}, B_{K,0}\}\end{aligned} \tag{21}$$

An orthogonal basis can be created using the following three terms recurrence:

$$\begin{cases} B_{-1,l}(\mathbf{x}){=}0 \\ B_{k,-1}(\mathbf{x}){=}0 \\ B_{0,0}(\mathbf{x}){=}1 \\ B_{k+1,l}(\mathbf{x}){=}(x_1{-}\lambda_{k+1,l})B_{k,l}(\mathbf{x}){-}\mu_{k+1,1}B_{k-1,l}(\mathbf{x}) \\ B_{k,l+1}(\mathbf{x}){=}(x_2{-}\lambda_{k,l+1})B_{k,l}(\mathbf{x}){-}\mu_{k,l+1}B_{k,l-1}(\mathbf{x}) \end{cases} \tag{22}$$

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

$$\begin{aligned} \lambda_{k+1,l} &= \frac{\langle x_1 B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l}(\mathbf{x})\|^2} & \lambda_{k,l+1} &= \frac{\langle x_2 B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l}(\mathbf{x})\|^2} \\ \mu_{k+1,l} &= \frac{\langle B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k-1,l}(\mathbf{x})\|^2} & \mu_{k,l+1} &= \frac{\langle B_{k,l}(\mathbf{x}) | B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l-1}(\mathbf{x})\|^2} \end{aligned} \tag{23}$$

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

$$\langle B_1 \mid B_2 \rangle = \iint_\Omega B_1(\mathbf{x}) B_2(\mathbf{x}) w(\mathbf{x}) d\mathbf{x} \tag{24}$$

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

Using this basis, the approximation of a decomposed primitive component $\mathcal{P}$ is:

$$\tilde{\mathcal{P}} = \sum_{k=0}^{D} \sum_{l=0}^{D-k} \tilde{u}_{k,l} \frac{B_{k,l}(\mathbf{x})}{\|B_{k,l}(\mathbf{x})\|} \tag{25}$$

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

$$\tilde{p}_{k,l} = \frac{\langle \mathcal{P} \mid B_{k,l}(\mathbf{x})\rangle}{\|B_{k,l}(\mathbf{x})\|} \tag{26}$$

We show the polynomials associated to a 4 degree basis in Fig. 3c. The polynomials are defined in a spatial domain of 32×32 pixels. In the case of video classification, space-time aggregation is considered. Kihl et al. . propose [31] to model spatial polynomial coefficients with a d degree temporal basis of Legendre polynomial defined by

$$\begin{cases} B_{-1}(t) = 0 \\ B_0(t) = 1 \\ T_n(t) = (t - \langle tB_{n-1}(t) | B_{n-1}(t)\rangle)B_{n-1}(t) - B_{n-2}(t) \\ B_n(t) = \dfrac{T_n(t)}{|T_n|} \end{cases} \tag{27}$$

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

$$\tilde{\mathbf{p}}_{k,l}(i, j, t) = \sum_{n=0}^{d} \tilde{p}_{k,l,n}(i, j, t) \frac{B_n(t)}{\|B_n(t)\|} \tag{28}$$

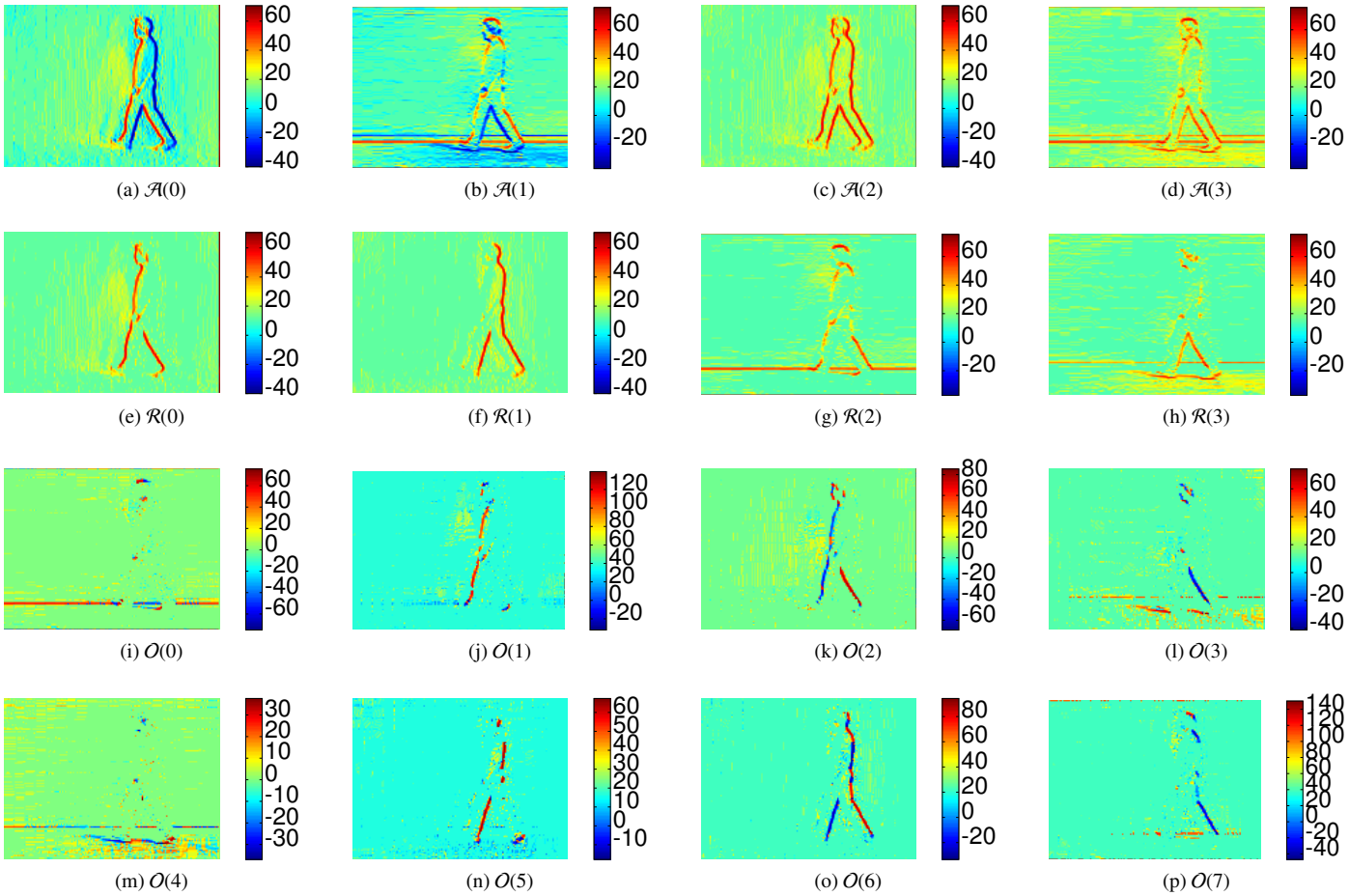
The model has $d + 1$ coefficients $\tilde{\mathbf{p}}_{k,l}(i, j, t)$ given by

$$\tilde{p}_{k,l,n}(i, j, t) = \frac{\langle \mathbf{p}_{k,l}(i, j, t) \mid B_n(t)\rangle}{\|B_n(t)\|} \tag{29}$$

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 (30)

$$\begin{aligned}\mathbf{m}_{l,k}(i, j, t_0) = [&\tilde{p}_{k,l,0}(i, j, t_0), \tilde{p}_{k,l,1}(i, j, t_0), \\ &\ldots, \tilde{p}_{k,l,d}(i, j, t_0)]\end{aligned} \tag{30}$$

Finally, the descriptor is the concatenation of all the $\mathbf{m}_{l,k}(i, j, t_0)$ vectors for each coded primitive. In this paper, we also propose an easy extension of this aggregation using a Sine basis, in place of the Legendre polynomials.



(a) $\mathcal{A}(0)$ (b) $\mathcal{A}(1)$ (c) $\mathcal{A}(2)$ (d) $\mathcal{A}(3)$

(e) $\mathcal{R}(0)$ (f) $\mathcal{R}(1)$ (g) $\mathcal{R}(2)$ (h) $\mathcal{R}(3)$

(i) $\mathcal{O}(0)$ (j) $\mathcal{O}(1)$ (k) $\mathcal{O}(2)$ (l) $\mathcal{O}(3)$

(m) $\mathcal{O}(4)$ (n) $\mathcal{O}(5)$ (o) $\mathcal{O}(6)$ (p) $\mathcal{O}(7)$

Figure 2: Example of coding ; on the first line: Absolute coding of the gradient primitive ($\mathcal{A}(0)$, $\mathcal{A}(1)$, $\mathcal{A}(2)$, $\mathcal{A}(3)$; on the second line: Rectified coding of the gradient primitive ($\mathcal{R}(0)$, $\mathcal{R}(1)$, $\mathcal{R}(2)$, $\mathcal{R}(3)$) ; on the third and fourth lines: Orientation coding of the gradient primitive ($\mathcal{O}(0)$, $\mathcal{O}(1)$, $\mathcal{O}(2)$, $\mathcal{O}(3)$, $\mathcal{O}(4)$, $\mathcal{O}(5)$, $\mathcal{O}(6)$, $\mathcal{O}(7)$) ;

Primitive	Coding	Aggregation
gradient	raw	Regular cells
motion	rectified	Flower
Haar	absolute	polynomial basis
motion gradient	orientation	sine basis
⋮	⋮	⋮

Table 1: A new framework for local descriptors

Name	Primitive	Coding	Aggregation
HOG	gradient	orientations	Regular cells
Daisy	gradient	orientations	Flower
HOF	motion	orientations	Regular cells
MBH	motion gradient	orientations	Regular cells
SURF	Haar	abs	cells
Efros	motion	rectified	Regular cells
SoPAF	motion	raw	polynomial basis

Table 2: Rewriting of the usual descriptors ; raw means the vector field is represented by the horizontal and vertical components

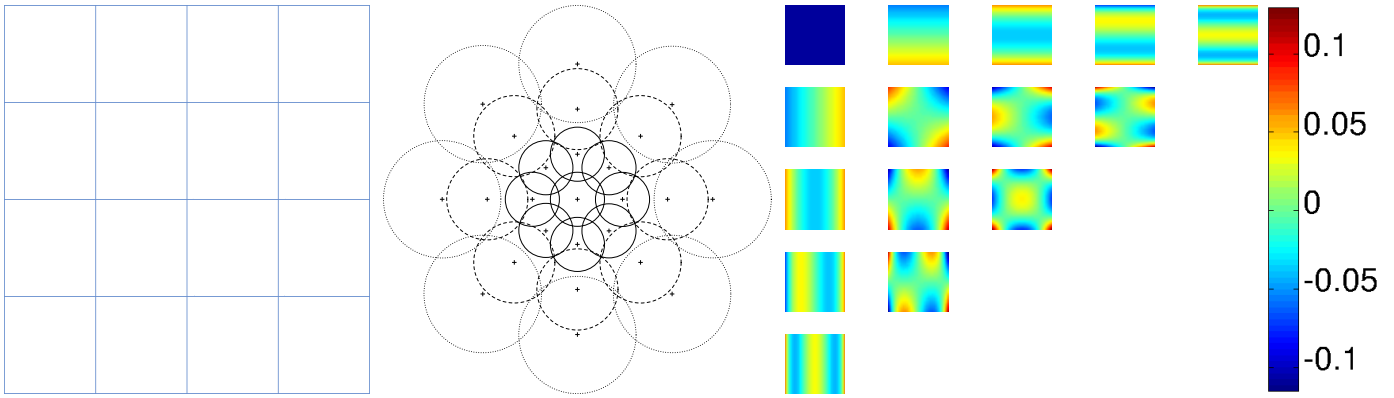
3.4. A unified framework for descriptors

In Table 1, we summarize the different primitives, coding and aggregations currently used for classification. According to specific combinations of primitive, coding and aggregation, we can explain most of the usual descriptors. In Table 2, we explain the usual descriptors of the literature with our framework.

Each new Primitive, Coding or Aggregation defines a new family of descriptors and each new combination of Primitive-Coding-Aggregation defines a new descriptor. Since different primitives correspond to different properties of the signal, we argue that adapted coding and aggregation schemes have to be used to produce efficient descriptors. Indeed, our framework allows to explore and evaluate the possible combinations so as to find the best descriptors.

4. Experiments

In the experiments, we compare several combination of primitives, coding and aggregations provided by our framework in order to evaluate still image descriptors and action descriptors.



(a) (b) (c)

Figure 3: Examples of aggregation ; (a) 4×4 cells aggregation ; (b) Flower aggregation with $Q = 3$ and $T = 8$; (c) Representaion of 4 degree basis spatial polynomials aggregation

As dense sampling outperforms key-point extraction [1, 3] for categories recognition, we use dense sampling in all our experiments. We carry out experiments on an image dataset (VOC2007) and three well known human action recognition datasets (KTH dataset [5], Hollywood2 Human Actions dataset [4] and UCF11 [19]).

For the experiments, we obtain signatures from our descriptors by using the VLAT indexing method [47] as explained in section 2.3.

4.1. Still image classification

We first present results on still image categorization. The gradient is the only primitive considered. The gradient is extracted with the simple one order approximation difference method, at a single resolution.

Pascal VOC 2007 dataset

The PASCAL-VOC 2007 dataset [1] consists in about 10,000 images and 20 categories, and is divided into 3 parts: "train", "val" and "test". We use linear SVM classifier trained on "train" + "val" sets and tested on the "test" set.

We use four primitive coding: absolute, rectified, double rectified and orientation. These coding are combinatorially associated to the following three code aggregations: regular cells, flower and polynomials basis. For the regular grid aggregation, we use 4×4 cells. The cells are evaluated at four scales: 4×4 pixels, 6×6 pixels, 8×8 pixels and 10×10 pixels. For the flower aggregation, the parameter Q is set to 3 and the parameter T is set to 8. We consider the flower aggregation at four scales by setting radius R at 9 pixels, 12 pixels, 15 pixels and 18 pixels. For the polynomial aggregation, we set the basis degree to 4. The polynomial spatial domain is considered at four scales: 16×16 pixels, 24×24 pixels, 32×32 pixels and 40×40 pixels.

For the VLAT signature, the number of projectors in equation (3) is set to 70. We use a dictionary of 256 visual words.

Figure 4: Images from PASCAL Visual Object Classes Challenge 2007

4.1.1. Experimental results on still images

Results for each descriptor are shown in table 3. We remark orientation coding clearly outperforms the other primitive coding for all the code aggregations experimented on this dataset. The GoF, GoC and GoP (c.f. Table 3) provide the best results. We remark that the three features with highest mean average precision are GoC, GoF and GoP for each category of the VOC2007 dataset. Using a simple concatenation of the signatures, we obtain 64.2% of mean average precision.

This result is reported in table 4 and compared with the results from [34]. Note that our approach provides a global image signature which does not include any kind of spatial information like Spatial Pyramidal Matching (SPM) [48] or object detectors [49]. We compare our results to those of Sanchez et al. [34] which gives results without spatial information. In [34]

	mAP	aeroplane	bicycle	bird	boat	bottle	bus
Our method	**64.2**	**83.3**	**73.0**	**59.9**	**73.5**	33.2	**71.2**
SIFT + FV [34]	62.7	80.2	69.1	52.8	72.9	**37.6**	69.5
	car	cat	chair	cow	table	dog	horse
Our method	**84.2**	**65.7**	53.3	**49.5**	58.8	**52.3**	**83.0**
SIFT + FV [34]	81.8	61.8	**54.9**	47.2	**61.5**	50.5	79.1
	bike	person	plant	sheep	sofa	train	tv
Our method	**72.0**	**87.5**	37.2	**47.4**	55.4	**85.5**	58.0
SIFT + FV [34]	67.1	85.8	**37.6**	46.6	**57.0**	82.3	**59.0**

Table 4: Image classification results on Pascal VOC 2007 dataset

name	**Coding**	**Aggregation**	**mAP**	**usual name**
GaC	absolute	regular cells	58,2	SURF
GaF	absolute	flower	56,6	x
GaP	absolute	polynomial basis	57.6	x
GrC	rectified	regular cells	58.1	x
GrF	rectified	flower	57.2	x
GrP	rectified	polynomial basis	57.4	x
GoC	orientation	regular cells	63.2	HOG
GoF	orientation	flower	63.7	DAISY
GoP	orientation	polynomial basis	63.2	x
GdC	double	regular cells	58,2	x
GdF	double	flower	56,9	x
GdP	double	polynomial basis	57,8	x

Table 3: Classification results exprimed by mean average precision for combination of primitives, coding and aggregations on VOC2007 dataset

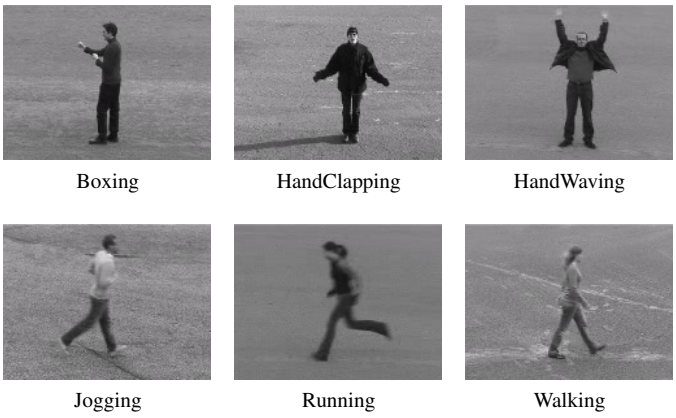


Figure 5: Example of videos from KTH

the SIFT descriptors are highly dense extracted at 7 resolutions and then aggregated with the Fisher Vector signature approach.

We show that our framework allows easy extension of HOG (GoC), for example by changing the codes aggregation from cell to polynomial. According to this new descriptor, we improve the categorization results obtained with only HOG descriptors. Moreover, our framework is compatible with adding spatial information like in [48], which should further improve the results.

4.2. Video actions recognition

In this section, we present results on video actions recognition. First, we evaluate our framework on the KTH [5] dataset. Then, we evaluate our method on two more challenging datasets of the literature, Hollywood2 [4]) and UCF11 [19] and compare our results to that of the literature. We compare three primitive extractions (gradient, motion and gradient of motion), three primitive coding (raw, rectified and orientations) and three code aggregations (cells, polynomials and sinus). We extract the gradient with the simple one order approximation difference method. For motion estimation, we use a Horn and Schunk optical flow algorithm [50] with 25 iteration and the regularization λ parameter is set to 0.1. We extract the primitives at 1 resolution for KTH, 7 resolutions for Hollywood2 and 5 resolutions for UCF11. The resolution factor is set to 0.8 . The resolutions are obtained by down sampling images, we do not use any up sampling in this work. We aggregate the extracted descriptors with the compressed VLAT signature approach as defined in the section 2.3.

4.2.1. Evaluation of our framework on KTH dataset

The KTH dataset [5] contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping (Fig. 5). These actions are performed by 25 different subjects in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, inside. For all experiments, we use the same experimental setup as in [5, 3], where the videos are divided into a training set (8 persons), a validation set (8 persons) and a test set (9 persons). The best hyperparameters are selected through cross-validation using the training and validation sets. The classification accuracy results are obtained on the test set.

We experiment several descriptors according to our framework for several spatial and time modeling. We present on Tables 5, 6 and 7 the main results obtained for each primitive extraction on KTH dataset. We present in Table 5 the results associated with the gradient primitive. As for still image experiments, the orientation coding clearly outperforms the other primitive coding for the three code aggregations. When the orientation coding is associated with the cell aggregation, it produces the best results for the gradient primitive extraction. The best results are obtained for code aggregations with the lower level of modeling along time axis for all code aggregations.

dim	coding	Gradient			SP	TP	Usual name
		Cell	Poly	Sinus			
32	raw	80.4			4	1	x
36	raw		81.0		2	2	x
40	raw	82.8			2	5	x
40	raw			86.8	1	4	x
64	raw		84.5		2	4	x
80	raw		83.1		3	3	x
48	rect			88.5	1	2	x
48	rect	84.8			2	3	x
60	rect		83.2		4	0	x
64	rect	86.5			4	1	x
64	rect			83.3	3	0	x
72	rect		84.5		2	2	x
80	rect	87.2			2	5	x
80	rect			88.5	1	4	x
128	rect		88.0		2	4	x
144	rect	88.5			3	4	x
96	ori	92.4			2	3	HOG
96	ori			91.4	1	2	x
120	ori		**92.6**		4	0	x
128	ori	**93.4**			4	1	HOG
128	ori			**93.3**	3	0	x

Table 5: Results for combination of gradient primitives, coding and aggregation on the KTH dataset ; dim means the dimension of the descriptor ; coding represent the code primitives (raw, rectified or orientation) ; SP means the number of spatial cells or the degree D of spatial polynomials or the spatial degree of the sinus basis ; TP means the number of temporal cells, or the degree d of temporal polynomials or the degree of sinus basis

dim	coding	Flow			SP	TP	Usual name
		Cell	Poly	Sinus			
32	raw	87.0			4	1	x
32	raw			85.1	3	0	x
36	raw		89.8		2	2	SoPAF
40	raw	89.6			2	5	x
40	raw			88.0	1	4	x
64	raw		90.4		2	4	SoPAF
80	raw		91.1		3	3	SoPAF
48	rect	90.7			2	3	x
48	rect			**91.3**	1	2	x
60	rect		90.7		4	0	x
64	rect	90.4			4	1	x
64	rect			87.7	3	0	x
72	rect		90.5		2	2	x
80	rect	91.4			2	5	x
80	rect			91.0	1	4	x
128	rect		**91.7**		2	4	x
144	rect	**92.0**			3	4	x
96	ori	89.2			2	3	HOF
96	ori			90.0	1	2	x
120	ori		90.6		4	0	x
128	ori	91.8			4	1	HOF
128	ori			87.8	3	0	x

Table 6: Results for combination of motion primitive, coding and aggregations on the KTH dataset ; The legend is the same as Table 5

dim	coding	Gradient of Motion			SP	TP	Usual name
		Cell	Poly	Sinus			
48	raw	90.0			2	3	x
48	raw			90.0	1	2	x
60	raw		90.3		4	0	x
64	raw	90.0			4	1	x
72	raw		90.6		2	2	x
80	raw	89.9			2	5	x
80	raw			89.4	1	4	x
128	raw		91.0		2	4	x
32	rect	92.2			2	1	x
32	rect			91.5	1	0	x
48	rect		93.1		2	0	x
96	rect	**94.2**			2	3	x
96	rect			**93.4**	1	2	x
120	rect		**93.7**		4	0	x
64	ori	92.5			2	1	MBH
64	ori			91.5	1	0	x
96	ori		93.6		2	0	x

Table 7: Results for combination of gradient of motion primitive, coding and aggregations on the KTH dataset ; The legend is the same as Table 5

We present in Table 6 the results associated with the motion primitive. In the case of motion primitive, the rectified coding allows to obtain good results for polynomial aggregation and sine aggregation. For the motion primitive, higher time modeling improves the results for a given spatial modeling. For instance, for the rectified coding and the polynomial aggregation with a spatial polynomial basis of degree 2, if the time polynomial basis is of degree 2 the classification accuracy is 90.5% and if the time polynomial basis is of degree 4, the accuracy is 91.7.

We present in Table 7 the results associated with the gradient of motion primitive. The best results, for each code aggregation, are obtained with rectified coding. It is interesting to note that we have only generated descriptors whose size does not exceed 144 dimensions. Note the motion of gradient primitive provides 4 components and the orientation coding decomposes each components in 8 orientation maps. So, the size of descriptors associating Motion of gradient primitive and orientation coding are easily of high dimension.

We present in Table 8 the classification accuracy results of several combinations of descriptors on KTH. We show the best descriptor results of our study for each primitives and codes aggregation, and compare them to recent results from the literature. Every single descriptor presented in Table 8 are comparable to those proposed by Wang et al. in [3]. Moreover, simple concatenation of all our signature (9) outperform the classification accuracy of Wang [3] and Gilbert [51]. 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 of [3] and [51]. Moreover, we do not used dense trajectory to follow descriptors along time axis as in [3].

Method	ND	NL	Results
Wang (HOG+traj) [3]	1	X	86.5%
Wang (HOF+traj) [3]	1	X	93.2%
Wang (MBH+traj) [3]	1	X	**95.0**%
Wang (All) [3]	4	X	94.2%
Gilbert [51]	3*	X	94.5%
A = Gradient + ori + Cell (4,1)	1		93.4%
B = Gradient + ori + Poly (4,0)	1		92.6%
C = Gradient + ori + Sine (3,0)	1		93.3%
D = Motion + ori + Cell (4,1)	1		91.8%
E = Motion + rect + Poly (2,4)	1		91.7%
F = Motion + rect + Sine (1,2)	1		91.3%
G = Grad of Motion + rect + Cell (2,3)	1		94.2%
H = Grad of Motion + rect + Poly (4,0)	1		93.7%
I = Grad of Motion + rect + Sine (1,2)	1		93.4%
A + **D** + **G**	3		94.2%
B + **E** + **H**	3		94.4%
C + **F** + **I**	3		93.5%
A+...+I	9		**94.7**%

Table 8: Classification accuracy on the KTH dataset ; ND means the number of descriptors used ; NL stands for non-linear classifiers ; * In [51], the same feature is iteratively combined with itself 3 times

4.2.2. Comparison to State of the art

For further experiments and comparisons to literature on Hollywood2 and UCF11 dataset, we use the nine best descriptors identified thanks to our experiments on KTH dataset.

Hollywood dataset

The Hollywood2 [4] dataset consists of a collection of video clips and extracts from 69 films in 12 classes of human actions (Fig. 6). 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.

UCF11 dataset

The UCF11 [19] dataset is an action recognition data set with 11 action categories, consisting of realistic videos taken from youtube (Fig. 7). The data set is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions. The videos are grouped into 25 groups, where each group consists of more than 4 action clips. The video clips in the same group may share some common features, such as the same person, similar background or similar viewpoint. The experimental setup is a leave one group out cross validation.

Results

We select the best setup according to gradient primitive associated with cells and polynomials projections (c.f. Table 5), the best setup according to Motion primitive associated with cells and polynomials projections (c.f. Table 6) and the best setup according to Gradient of motion primitive associated with cells and polynomials projections (c.f. Table 7). These setups are evaluated on the Hollywood2 dataset and results are reported in Table 9. The results presented here improve the state of the art for single descriptor setups when comparing to HOG (gradient primitive), to HOF (motion primitive) and to MBH (gradient of motion primitive). Note that, opposed to [3], we do not use the dense trajectories to obtain these results. Our framework allows to increase the number of descriptor for a fixed number of primitives. Finally, by combining nine primitives, we obtain a mean average precision of 60.2%.

We evaluate the same descriptors on the UCF11 dataset and we report our results in Table 10. On UCF11 dataset, for all the primitives extraction, the cell aggregation and polynomial aggregation improve the results of Wang et al. [3] for single descriptor corresponding to that primitive. However, the Sine aggregation produces lower results in the case of Gradient primitive and Gradient of Motion primitive. When combining descriptors, we improve the results of Wang et al. [3] without using dense trajectories. The results obtained on the challenging UCF11 and Hollywood2 datasets with the combination of several descriptors highlight the soundness of our framework.

Figure 6: Example of videos from Hollywood2 dataset

Figure 7: Example of videos from UCF11

5. Conclusion

In this paper, we introduced a new framework to describe local visual descriptors. This framework consists in the decomposition of descriptors in three levels: primitive extraction, primitive coding and code aggregation. Our framework allows us to easily explain popular descriptors of the literature. Moreover, our framework allows us to propose extensions of popular descriptors, for instance by introducing a function based aggregation.

Using our framework, we experimented several combination of primitives extraction, primitive coding and code aggregation, some of them being drawn from the most popular descriptors. We obtain better or equivalent results for than the usual descriptors of literature on a popular still image categorization dataset and on three well known videos recognition datasets. This confirms the validity and relevance of our framework to create new descriptors. We are confident our framework can be used to implement descriptors families not covered in this paper (for example dense trajectories).

Furthermore, it is interesting to compare our framework to the coding/pooling approaches [37] used to compute signatures. Indeed, the two last steps of our framework (primitive coding and code aggregation) are close in their objective to the coding step and the pooling step in signature computing methods. In this case, the primitive extraction can be replaced by the extraction of a set of local descriptors. Future work involves adapting recent signature computation methods to the descriptors using our framework. For example, dictionary based approaches [36, 37] and model deviation approaches [40, 42, 44] can be used for the coding and aggregation steps. Future work also involves the optimization of the primitive step by using machine learning algorithms. For example, the primitive can be an adapted filter bank trained on some training set, in a similar way of the deep learning approaches [53] or the infinite kernel learning approaches [54].

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: BMVC, Vol. 76, 2011, pp. 1–12.

[3] H. Wang, A. Klaser, C. Schmid, C. Liu, Action recognition by dense trajectories, in: Conference on CVPR, IEEE, 2011, pp. 3169–3176.

[4] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Conference on CVPR, IEEE, 2008, pp. 1–8.

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

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

[7] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, ECCV (2006) 404–417.

[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Conference on CVPR, IEEE, 2005, pp. 886–893.

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

[10] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, Transactions on Pattern Analysis and Machine Intelligence 27 (10) (2005) 1615–1630.

[11] E. Tola, V. Lepetit, P. Fua, Daisy: An efficient dense descriptor applied to wide-baseline stereo, Transactions on Pattern Analysis and Machine Intelligence 32 (5) (2010) 815–830.

[12] J. Davis, A. Bobick, The representation and recognition of action using temporal templates, in: Conference on CVPR, IEEE, 1997, pp. 928–934.

Method	ND	NL	Results
Gilbert [51]	3	X	50.9%
Ullah [52] HOG+HOF	2	X	51.8%
Ullah [52]	2⋆	X	55.3%
Wang [3] traj	1	X	47.7%
Wang [3] HOG	1	X	41.5%
Wang [3] HOF	1	X	50.8%
Wang [3] MBH	1	X	54.2%
Wang [3] all	4	X	**58.3**%
A = Gradient + ori + Cell (4,1)	1		44.4%
B = Gradient + ori + Poly (4,0)	1		48.4%
C = Gradient + ori + Sine (3,0)	1		45.0%
D = Motion + rect + Cell (3,4)	1		53.3%
E = Motion + rect + Poly (2,4)	1		52.7%
F = Motion + rect + Sine (1,2)	1		49.5%
G = Grad of Motion + rect + Cell (2,3)	1		56.2%
H = Grad of Motion + rect + Poly (4,0)	1		54.9%
I = Grad of Motion + rect + Sine (1,2)	1		52.0%
A + **D** + **G**	3		59.1%
B + **E** + **H**	3		58.8%
C + **F** + **I**	3		56.4%
A+...+I	9		**60.2**%

Table 9: Mean Average Precision on the Hollywood2 dataset ; ND: number of descriptors ; NL: non-linear classifiers ; ⋆ In [52] HOG/HOF descriptors are accumulated on over 100 spatio-temporal regions each one leading to a different BoW signature

Method	ND	NL	Results
Wang [3] traj	1	X	67.2%
Wang [3] HOG	1	X	74.5%
Wang [3] HOF	1	X	72.8%
Wang [3] MBH	1	X	83.9%
Wang [3] all	4	X	**84.2**%
A = Gradient + ori + Cell (4,1)	1		80.1%
B = Gradient + ori + Poly (4,0)	1		81.8%
C = Gradient + ori + Sine (3,0)	1		73.0%
D = Motion + rect + Cell (3,4)	1		81.0%
E = Motion + rect + Poly (2,4)	1		82.6%
F = Motion + rect + Sine (1,2)	1		75.9%
G = Grad of Motion + rect + Cell (2,3)	1		84.2%
H = Grad of Motion + rect + Poly (4,0)	1		86.0%
I = Grad of Motion + rect + Sine (1,2)	1		79.2%
A + **D** + **G**	3		86.1%
B + **E** + **H**	3		**87.6**%
C + **F** + **I**	3		82.8%
A+...+I	9		86.5%

Table 10: Mean Average Precision on the UCF11 dataset ; ND: number of descriptors ; NL: non-linear classifiers ; ⋆ In [52] HOG/HOF descriptors are accumulated on over 100 spatio-temporal regions each one leading to a different BoW signature

[13] V. Kellokumpu, G. Zhao, M. Pietikäinen, Texture Based Description of Movements for Activity Analysis, in: VISAPP, Vol. 1, 2008, pp. 206–213.

[14] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 971–987.

[15] V. Kellokumpu, G. Zhao, M. Pietikäinen, Human activity recognition using a dynamic texture based method, in: BMVC, 2008, pp. 885–894.

[16] L. Wang, D. Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Transactions on Image Processing 16 (6) (2007) 1646.

[17] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: ICCV, Vol. 2, IEEE, 2005, pp. 1395–1402.

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, Transactions on Pattern Analysis and Machine Intelligence 29 (12) (2007) 2247–2253.

[19] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: Conference on CVPR, IEEE, 2009, pp. 1996–2003.

[20] R. Polana, R. Nelson, Low level recognition of human motion, in: Proc. IEEE Workshop on Nonrigid and Articulate Motion, 1994, pp. 77–82.

[21] A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: ICCV, Vol. 2, IEEE, 2003, pp. 726–733.

[22] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th international joint conference on Artificial intelligence, Vol. 2, 1981, pp. 674–679.

[23] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: Conference on CVPR, IEEE, 2008, pp. 1–8.

[24] S. Danafar, N. Gheissari, Action recognition for surveillance applications using optic flow and svm, in: ACCV, Vol. 4844, 2007, pp. 457–466.

[25] D. Tran, A. Sorokin, Human activity recognition with metric learning, ECCV (2008) 548–561.

[26] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 288–303.

[27] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,, IEEE, 2005, pp. 65–72.

[28] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: BMVC, 2008.

[29] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th international conference on Multimedia, ACM, 2007, pp. 357–360.

[30] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, ECCV (2008) 650–663.

[31] O. Kihl, B. Tremblais, B. Augereau, M. Khoudeir, Human activities discrimination with motion approximation in polynomial bases, in: ICIP, IEEE, 2010, pp. 2469–2472.

[32] V. F. Mota, E. Perez, M. B. Vieira, L. Maciel, F. Precioso, P.-H. Gosselin, A tensor based on optical flow for global description of motion in videos, in: 25th SIBGRAPI Conference on Graphics, Patterns and Images, IEEE, 2012, pp. 298–301.

[33] O. Kihl, D. Picard, P.-H. Gosselin, Local polynomial space-time descriptors for actions classification, in: IAPR MVA, Kyoto, Japon, 2013.

[34] J. Sánchez, F. Perronnin, T. d. Campos, Modeling the spatial layout of images beyond spatial pyramids, Pattern Recognition Letters.

[35] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: BMVC, 2009.

[36] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV, Vol. 2, IEEE, 2003, pp. 1470–1477.

[37] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Conference on CVPR, IEEE, 2010, pp. 3360–3367.

[38] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Conference on CVPR, IEEE, 2009, pp. 1794–1801.

[39] S. Avila, N. Thome, M. Cord, E. Valle, A. de A Araujo, Bossa: Extended bow formalism for image classification, in: ICIP, IEEE, 2011, pp. 2909–2912.

[40] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Conference on CVPR, IEEE, 2010, pp. 3304–3311.

[41] X. Zhou, K. Yu, T. Zhang, T. Huang, Image classification using super-vector coding of local image descriptors, Computer Vision–ECCV 2010

(2010) 141–154.
[42] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, C. Schmid, Aggregating local image descriptors into compact codes, Transactions on Pattern Analysis and Machine Intelligence 34 (2012) 1704–1716.
[43] D. Picard, P.-H. Gosselin, Efficient image signatures and similarities using tensor products of local descriptors, CVIU 117 (6) (2013) 680–687.
[44] D. Picard, P.-H. Gosselin, Improving image similarity with vectors of locally aggregated tensors, IEEE, 2011, pp. 669–672.
[45] R. Negrel, D. Picard, P. Gosselin, Using spatial pyramids with compacted vlat for image categorization, in: ICPR, 2012, pp. 2460–2463.
[46] M. Varma, A. Zisserman, A statistical approach to material classification using image patch exemplars, Transactions on Pattern Analysis and Machine Intelligence 31 (11) (2009) 2032–2047.
[47] R. Negrel, D. Picard, P. Gosselin, Compact tensor based image representation for similarity search, in: ICIP, IEEE, 2012, pp. 2425–2428.
[48] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Conference on CVPR, Vol. 2, IEEE, 2006, pp. 2169–2178.
[49] L.-J. Li, H. Su, E. P. Xing, L. Fei-Fei, Object bank: A high-level image representation for scene classification and semantic feature sparsification, Advances in Neural Information Processing Systems 24.
[50] B. Horn, B. Schunck, Determining optical flow, Artificial intelligence 17 (1) (1981) 185–203.
[51] A. Gilbert, J. Illingworth, R. Bowden, Action recognition using mined hierarchical compound features, Transactions on Pattern Analysis and Machine Intelligence (99) (2011) 883–897.
[52] M. Ullah, S. Parizi, I. Laptev, Improving bag-of-features action recognition with non-local cues, in: BMVC, 2010.
[53] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1915–1929.
[54] A. Rakotomamonjy, R. Flamary, F. Yger, Learning with infinitely many features, Machine Learning 91 (1) (2013) 43–66. doi:10.1007/s10994-012-5324-5.