Local polynomial space–time descriptors for action classification

Olivier Kihl, David Picard, Philippe-Henri Gosselin

Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Local polynomial space-time descriptors for actions classification

Olivier Kihl · David Picard · Philippe-Henri Gosselin

Received: date / Accepted: date

Abstract In this paper we propose to tackle human actions indexing by introducing a new local motion descriptor based on a model of the optical flow. We propose to apply a coding step to vector field before the modeling. We use 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 UCF11 dataset and on the more challenging Hollywood2 action classification dataset and show promising results.

Keywords Action classification · Visual descriptors · Motion

1 Introduction

Human action recognition has become an important research area in computer vision since it concerns several key applications like video indexing, video surveillance or human computer interaction. The typical setup for this task involves the extraction of highly discriminant features localized in both space and time. A wide variety of such descriptors have been introduced recently [38,39,23,32], and have become essential tools of the action classification framework. These descriptors are then aggregated into a single vector using the extension to video of the well known *"Bag of Words"* image signature approaches [35]. To further improve the results, most action classification systems use the combination of several complementary descriptors.

This paper is a revised and extended version of earlier work presented in [19]. The proposed descriptor is localized spatially and temporally in a space-time tube, in order to capture characteristic atoms of motion. The Serie of Polynomial Approximation of Flow (SoPAF) space-time motion descriptor is based on polynomial decomposition of the optical flow [19].

We propose to extend this descriptor by coding the vector field with the half-wave rectification proposed by Efros *et al.* [9]. Moreover, we study two different functions basis (polynomial and sine) for modeling the temporal evolution of spatial polynomial coefficients.

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 the SoPAF descriptor and our extension. 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

Olivier Kihl and David Picard
ETIS/ ENSEA - Université Cergy-Pontoise
CNRS, UMR 8051
Tel.: +33 1 30 73 62 96
Fax: +33 1 30 73 66 27
E-mail: olivier.kihl@ensea.fr
E-mail: david.picard@ensea.fr

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

from video, to find a transformation that maps the set of descriptors into a single vector, and then to measure the similarity between the obtained vectors.

We first present several works related to descriptors extraction, and then we present the most popular signature approaches.

2.1 Video 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 [8] 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. Kellokumpu *et al.* use histograms of *Local Binary Patterns* (LBP) to model the MHI and MEI images [18]. In [17], they propose an extension of the LBP directly applied on the image pixels with successful results. Wang and Suter [41] 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.* study the silhouettes as space-time volumes [4,12]. 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 [4] or the KTH dataset [32]. Moreover, they tend to fail on more realistic data-sets such as UCF11 [24] or Hollywood2 [23].

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 encodes the displacement of pixels from two consecutive frames. The result can be represented by vector fields with two components $\mathcal{U}$ and $\mathcal{V}$. 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 [30]. 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.* propose a descriptor computed on a figure-centric spatio-temporal volume for each person in a video [9]. The vector field representing the motion between two consecutive frames of the volume is computed with the Lucas and Kanade optical flow algorithm [26]. 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 [10, 7].

Tran *et al.* have proposed the motion context descriptor [36]. 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 first compute many kinematic features on the field, and then compute kinematic modes with a spatio-temporal principal component analysis to create a figure-centric descriptor [1].

Figure-centric descriptors are dependent on the person detector associated with them. Moreover, they don't take into account the context in the video that can add relevant information to action recognition. Consequently, these methods tend to fail on more realistic data-sets such as UCF11 [24] or Hollywood2 [23] datasets.

An other approach is proposed in [33,34] that allows to compute the similarity between motions of videos segments without computing motion fields. This method do not have to use video background subtraction. However, this method requires a set of training video centered on the action to recognize.

Finally, the descriptors that have emerged in recent years are the extention to video of still image descriptors [38]. The most commonly used are SIFT [25],

SURF [3] and Histogram of oriented gradient (HOG) [6]. 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. 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.

Recently, Wang *et al.* propose to model these usual descriptors along dense trajectories [38]. 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 increases the performances of popular descriptors (HOG, HOF and MBH).

2.2 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 [31], [39] and [38]). The most common method for such embeddings is inspired by the text retrieval community and is called the "Bag of Words" (BoW) approach [35]. 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 [40] 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 [42] or locality [40] 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 [2], 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 pooling is simply the averaging over all codewords.

In [15], 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 [43], *"Fisher Vectors"* in [16] or *"VLAT"* in [29,28]. Fisher Vectors are known to achieve state of the art performances in image classification challenges [5].

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 [27]. 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 $|c|$ being the number of descriptors ν_{rci} of video i in cluster c, for all videos in the training set. 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:

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

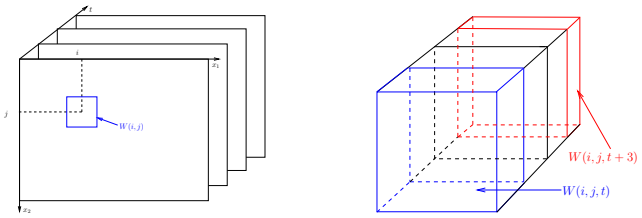
$\tau_{i,c}$ are then flattened into vectors $\mathbf{v}_{i,c}$. The complete VLAT signature $\mathbf{x}_i$ of video i is obtained by concatenation of $\mathbf{v}_{i,c}$ for all clusters c:

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

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

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

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



(a) (b)

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

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 Series of Polynomial Approximation of Flow (SoPAF)

We propose to extend the SoPAF descriptor [19]. SoPAF descriptor models the vector field of motion between two frames using projection on an orthogonal basis of polynomials. This polynomial model is used in [21] to recognize movements in a video. The modeling is applied to the entire field and each frame is processed separately. In another context, this polynomial model is locally used to detect singularities such as vortex or saddle point in fluid motion [20]. 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, the spatial vector field is modeled with two dimensional polynomial basis. Then, time evolution of spatial coefficients are modeled with a one dimensional basis. We propsose to extend the descriptor using the half-wave rectification technique proposed by Efros *et al.* [9]. Moreover, we propose to evaluate sine functions basis in addition to of polynomial functions.

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 Fig.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:

$$P_{K,L}(x_1,x_2) = \sum_{k=0}^{K}\sum_{l=0}^{L} a_{k,l}\ x_1^k\ x_2^l \tag{8}$$

where $k \in \{0..K\}, l \in \{0..L\}$, $K \in \mathbb{N}^+$ and $L \in \mathbb{N}^+$ are respectively the maximum degree of the variables (x_1,x_2) and $\{a_{k,l}\}$ 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:

$$\begin{aligned}\mathcal{B} = \quad & \{P_{0,0}, P_{0,1}, \cdots, P_{0,L}, P_{1,0}, \cdots \\ & \cdots, P_{1,L-1}, \cdots, P_{K-1,0}, P_{K-1,1}, P_{K,0}\}\end{aligned} \tag{9}$$

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

$$\begin{cases} P_{-1,l}(\mathbf{x}) = 0 \\ P_{k,-1}(\mathbf{x}) = 0 \\ P_{0,0}(\mathbf{x}) = 1 \\ P_{k+1,l}(\mathbf{x}) = (x_1 - \lambda_{k+1,l})P_{k,l}(\mathbf{x}) - \mu_{k+1,1}P_{k-1,l}(\mathbf{x}) \\ P_{k,l+1}(\mathbf{x}) = (x_2 - \lambda_{k,l+1})P_{k,l}(\mathbf{x}) - \mu_{k,l+1}P_{k,l-1}(\mathbf{x}) \end{cases} \tag{10}$$

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

$$\begin{aligned} \lambda_{k+1,l} &= \frac{\langle x_1 P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l}(\mathbf{x})\|^2} & \lambda_{k,l+1} &= \frac{\langle x_2 P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l}(\mathbf{x})\|^2} \\ \mu_{k+1,l} &= \frac{\langle P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k-1,l}(\mathbf{x})\|^2} & \mu_{k,l+1} &= \frac{\langle P_{k,l}(\mathbf{x}) | P_{k,l}(\mathbf{x})\rangle}{\|P_{k,l-1}(\mathbf{x})\|^2} \end{aligned} \tag{11}$$

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

$$\langle P_1 \mid P_2 \rangle = \iint_{\Omega} P_1(\mathbf{x})P_2(\mathbf{x})w(\mathbf{x})d\mathbf{x} \tag{12}$$

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:

$$\tilde{\mathcal{U}} = \sum_{k=0}^{D}\sum_{l=0}^{D-k} \tilde{u}_{k,l}\frac{P_{k,l}(\mathbf{x})}{\|P_{k,l}(\mathbf{x})\|} \tag{13}$$

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

$$\tilde{u}_{k,l} = \frac{\langle \mathcal{U} \mid P_{k,l}(\mathbf{x}) \rangle}{\|P_{k,l}(\mathbf{x})\|} \tag{14}$$

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 Fig.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 Fig.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

$$\mathbf{u}_{k,l}(i,j,t_0) = [\tilde{u}_{k,l}(i,j,t_0), \ldots, \tilde{u}_{k,l}(i,j,t_0+\tau)] \tag{15}$$

We 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. In [19], we use Legendre polynomial basis of degree d defined by

$$\begin{cases} P_{-1}(t) = 0 \\ P_0(t) = 1 \\ T_n(t) = (t - \langle tP_{n-1}(t) | P_{n-1}(t) \rangle) P_{n-1}(t) - P_{n-2}(t) \\ P_n(t) = \dfrac{T_n(t)}{|T_n|} \end{cases} \tag{16}$$

In this work, we also use Sine basis for time evolution modeling. Using such basis (polynomial or sine) with degree d, the approximation of $\mathbf{u}_{k,l}(i,j,t)$ is:

$$\tilde{\mathbf{u}}_{k,l}(i,j,t) = \sum_{n=0}^{d} \tilde{u}_{k,l,n}(i,j,t) \frac{P_n(t)}{\|P_n(t)\|} \tag{17}$$

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

$$\tilde{u}_{k,l,n}(i,j,t) = \frac{\langle \mathbf{u}_{k,l}(i,j,t) \mid P_n(t) \rangle}{\|P_n(t)\|} \tag{18}$$

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 (19)

$$\begin{aligned} \mathbf{m}_{l,k}(i,j,t_0) = [&\tilde{u}_{k,l,0}(i,j,t_0), \tilde{u}_{k,l,1}(i,j,t_0), \\ &\ldots, \tilde{u}_{k,l,d}(i,j,t_0)] \end{aligned} \tag{19}$$

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

$$\begin{aligned} \nu(i,j,t_0) = [&\mathbf{m}_{0,0}, \mathbf{m}_{0,1}, \cdots, \mathbf{m}_{0,L}, \mathbf{m}_{1,0}, \cdots, \mathbf{m}_{1,L-1}, \cdots \\ &\cdots, \mathbf{m}_{K-1,0}, \mathbf{m}_{K-1,1}, \mathbf{m}_{K,0}, \mathbf{n}_{0,0}, \mathbf{n}_{0,1}, \cdots \\ &\cdots, \mathbf{n}_{0,L}, \mathbf{n}_{1,0}, \cdots, \mathbf{n}_{1,L-1}, \cdots \\ &\mathbf{n}_{K-1,0}, \mathbf{n}_{K-1,1}, \mathbf{n}_{K,0}] \end{aligned} \tag{20}$$

Here, $\mathbf{m}_{k,l}(i,j,t_0)$ and $\mathbf{n}_{k,l}(i,j,t_0)$ are written as $\mathbf{m}_{k,l}$ and $\mathbf{n}_{k,l}$ for clarity reasons. The size of the descriptor $\nu_(i,j,t_0)$ is $(D+1) \times (D+2) \times d$.

We name Series of Polynomial approximation of Flow the descriptor as it is defined in [19]. If Sine basis is used to model the motion vector field evolution, we name the descriptor SoPAF+Sine. Note, for the spatial modeling of the vector field, only polynomial basis are used.

3.3 Series of local Polynomial Approximation of Rectified Flow

We propose an extension of the SoPAF descriptor by using the half-wave rectification coding proposed by Efros *et al.* in [9] and used in several works. The half-wave rectification coding produces a four dimension code from the horizontal component $\mathcal{U}$ and the horizontal component $\mathcal{V}$ of the vector field. The code is defined as:

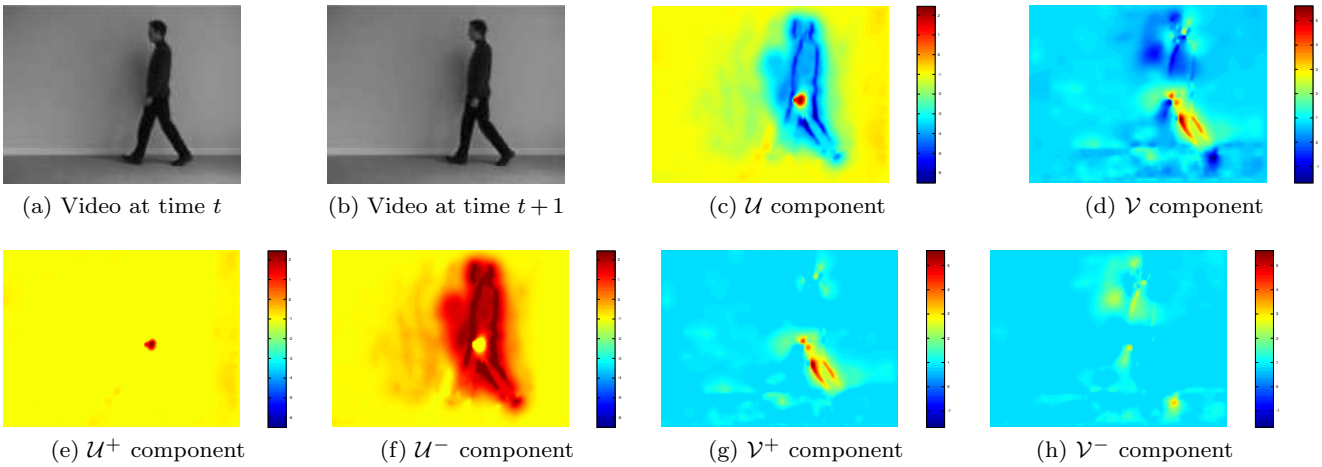
$$\mathcal{U}^+(\mathbf{x}) = \begin{cases} \mathcal{U}(\mathbf{x}) & \text{if } \mathcal{U}(\mathbf{x}) > 0 \\ 0 & \text{else} \end{cases} \tag{21}$$

$$\mathcal{U}^-(\mathbf{x}) = \begin{cases} \mathcal{U}(\mathbf{x}) & \text{if } \mathcal{U}(\mathbf{x}) < 0 \\ 0 & \text{else} \end{cases} \tag{22}$$

$$\mathcal{V}^+(\mathbf{x}) = \begin{cases} \mathcal{V}(\mathbf{x}) & \text{if } \mathcal{V}(\mathbf{x}) > 0 \\ 0 & \text{else} \end{cases} \tag{23}$$

$$\mathcal{V}^-(\mathbf{x}) = \begin{cases} \mathcal{V}(\mathbf{x}) & \text{if } \mathcal{V}(\mathbf{x}) < 0 \\ 0 & \text{else} \end{cases} \tag{24}$$

This coding is applied to motion vector field before the modeling steps of SoPAF descriptor. This preprocessing doubles the dimensions of the obtained descriptor. We



(a) Video at time t (b) Video at time $t+1$ (c) $\mathcal{U}$ component (d) $\mathcal{V}$ component

(e) $\mathcal{U}^+$ component (f) $\mathcal{U}^-$ component (g) $\mathcal{V}^+$ component (h) $\mathcal{V}^-$ component

Fig. 2: Half-wave rectification

show in Fig 2 an example of of half-wave rectification coding. In case we use the half-wave rectification coding step, we name the descriptor Serie of Polynomial Approximation of Rectified Flow (SoPARF). If Sine basis is used to model the motion vector field evolution, we name the descriptor SoPARF+Sine.

3.4 Trajectories

As proposed in [38], 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 [22]. The temporal evolution of spatial coefficients is thus modeled on tubes instead of volumes.

4 Experiments

We carry out experiments on two well known human action recognition datasets. The first one is the UCF11 dataset [24], and the second one is the Hollywood2 Human Actions dataset [23].

In this section, we first introduce the two datasets. Second, we evaluate parameters of our descriptor on the UCF11 dataset. Third, we compare our descriptor to literature results on UCF11 and Hollywood2 datasets. For the parametrization, we use the best results obtained on UCF11 evaluation.

We use a Horn and Schunk optical flow algorithm [13] for motion extraction with 25 iterations and the regularization parameter λ is set to 0.1. We extract the motion fields at 5 scales for UCF11 and 7 for Hollywood2, the scale factor is set to 0.8.

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

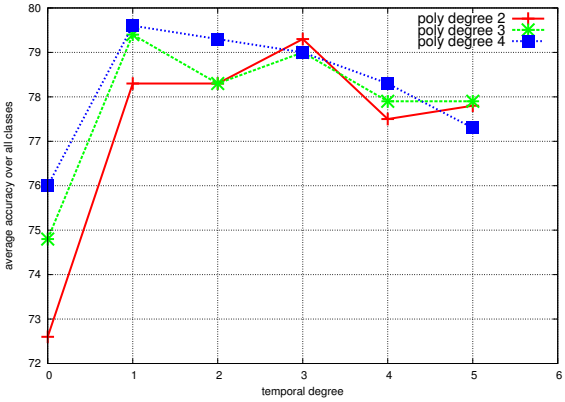
4.1 Datasets

4.1.1 UCF11 dataset

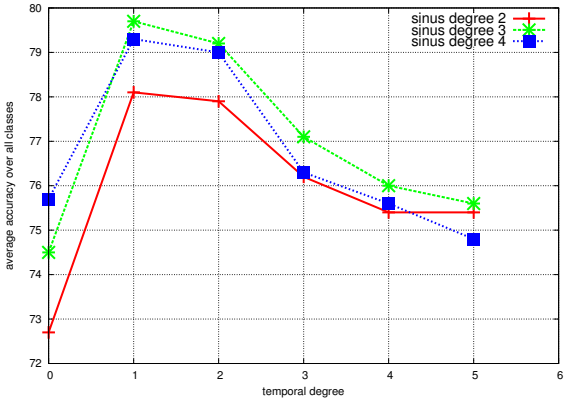
The UCF11 [24] dataset is an action recognition data set with 11 action categories, consisting of realistic videos taken from youtube (Fig. 3). The data set is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination 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.

4.1.2 Hollywood dataset

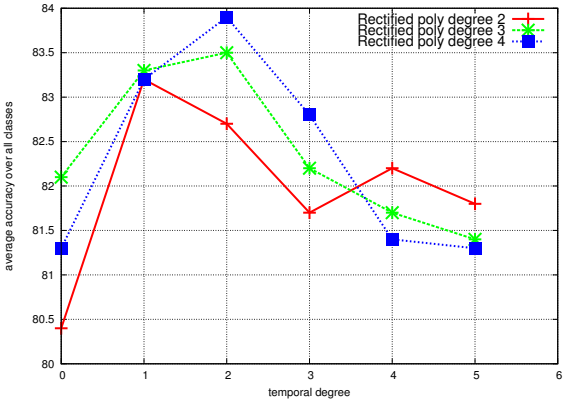
The Hollywood2 [23] dataset consists of a collection of video clips and extracts from 69 films in 12 classes of human actions (Fig.4). It accounts for approximately 20 hours of video and contains about 150 video samples per actions. It contains a variety of spatial scales, zoom camera, deleted scenes 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.



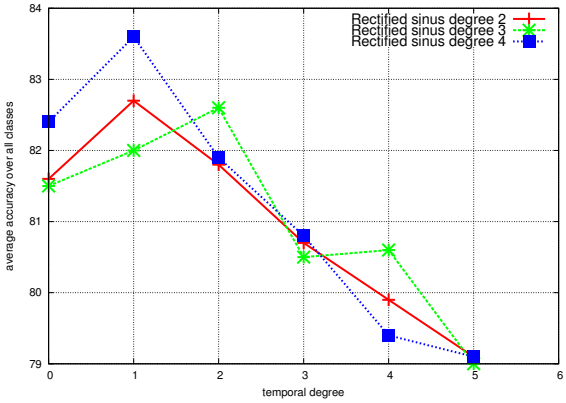
(a) Series of Polynomial Approximation of Flow



(b) Series of Polynomial Approximation of Flow with sine functions approximation along time axis (SoPAF+Sine)



(c) Series of Polynomial Approximation of Rectified Flow



(d) Series of Polynomial Approximation of Rectified Flow with sine functions approximation along time axis (SoPARF+Sine)

Fig. 5: Evaluation of space and time degree for our descriptor on UCF11 dataset; The horizontal axis represents the degree of the temporal functions basis and the vertical axis represents the average accuracy

4.2 Evaluation of our descriptor

In this section, we evaluate our descriptor. The spatial size of space-time volumes are set to 32×32 pixels and the length is set to 15. These parameters are defined according to results of the evaluation of parameters of HOG, HOF and MBH in [38]. The spatial step for dense extraction is set to 10 pixels and the time step is set to 5 frames. In Fig.5, we show the results of our evaluation. In Fig.5(a), we show the results for the SoPAF with spatial degree varied from 2 to 4, and time degree varied from 0 to 5. The best results are obtained for spatial degree 4 and time degree 1. In Fig.5(b), we show the results for SoPAF+Sine with spatial degree varied from 2 to 4, and time degree varied from 0 to 5. The best results is obtained for spatial degree 3 and time degree 1. In Fig.5(c), we show the results for the SoPARF with spatial degree varied from 2 to 4, and time degree varied from 0 to 5. The best results is obtained for spatial degree 4 and time degree 2. This result is clearly better than results of SoPAF and SoPAF+Sine. In Fig.5(d), we show the results for the SoPARF+Sine with spatial degree varied from 2 to 4, and time degree varied from 0 to 5. The best results is obtained for spatial degree 4 and time degree 1. This result is slightly lower than SoPARF but clearly better than those of SoPAF and SoPAF+Sine. We compare now our descriptors with HOF, since it models the same information as ours. In order to compare our descriptor with HOF, we evaluate HOF for space grid from 2×2 to 4×4 cells and time grid from 1 to 4 cells. We show the results of this evaluation in Fig.6. Note we obtain at best 80.4%, which is better than Wang *et al.* in [38], albeit with a different configuration of the HOF descriptor. Our best setup is obtained for a grid of 3×3 cells and a time grid of 2 cells. With SoPARF and SoPARF+Sine descriptors,

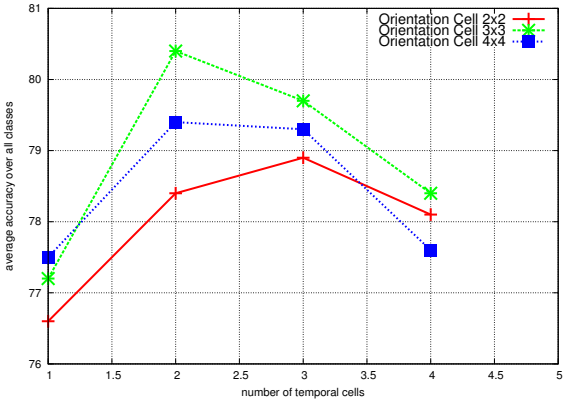


Fig. 6: Evaluation of HOF descriptor on UCF11 dataset ; The horizontal axis represent the number of cells along the time axis and the vertical axis represent the average accuracy

Fig. 3: Example of videos from UCF11

Fig. 4: Example of videos from Hollywood2 dataset

we obtain significantly better results (83.9% and 83.6% resp.) than HOF (80.4%).

4.3 Comparison of descriptors computational time

We compare the computation of our four best setups to the computation time of the best HOF descriptor in the previous evaluation. The computation of descriptors is performed with an Intel(R) Xeon(R) E5-2620 0 @ 2.00GHz processor. We compute all the descriptors of the video called "*v_biking*_01_01" of the UCF11 dataset.

In order to be fair in comparison, we use the same space-time dimensions of descriptors. We use a dimension of 30×30 pixels spatially and 14 pixels temporally. The spatial step between descriptors is set to 10 and the temporal step is set to 5.

The results of computational time by frame (in seconds) are reported in Table 1. As one can see, descriptors that don't use rectification coding are comparable to HOF in computational time. When we use the rectification coding, the computational time clearly increase. However, the best descriptors in our evaluation is the SoPARF and its computational time is still acceptable for real datasets.

4.4 Experimental results

In this section we compare our descriptors to the literature on the two datasets. For each dataset, we show the results with our SoPARF and SoPARF+Sine descriptors alone and with a HOG and MBH descriptors

Table 1: Computational time of the best descriptors from the evaluation presented in section 4.2

descriptor	parameters	parameters	times/frame
HOF	3×3 cells	2 cells	0.68 s
SoPAF	poly degree 4	poly degree 1	0.85 s
SoPAFs	poly degree 3	sine degree 1	0.62 s
SoPARF	poly degree 4	poly degree 2	1.12 s
SoPARFs	poly degree 4	sine degree 1	1.26 s

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 [38] and [11].

On Table 2, we show the results obtained on UCF11 dataset, and compare them to recent results from the literature. We obtain good results only using the proposed SoPARF or SoPARF+Sine descriptors. The SoPARF improves the results of Wang et al. HOF descriptor by 11% and our implementation of HOF by 3%. The SoPARF provides the same results than the MBH of Wang et al. and improve by 4% the SoPAF. When using SoPARF, HOG and MBH combination or SoPARF+Sine, HOG and MBH combination we obtain 86.0% of average accuracy, which is above state of the art performances while using a linear classifier and combining less descriptors.

Table 2: Classification average accuracy on the UCF11 dataset ; ND means the number of descriptors used ; NL stands for non-linear classifiers

Method	ND	NL	Results
Ikizler [14]	6		75.2%
Wang [38](trajectory)	1	X	67.2%
Wang [38](HOG)	1	X	74.5%
Wang [38](HOF)	1	X	72.8%
Wang [38](MBH)	1	X	83.9%
Wang [38](all)	4	X	84.2%
HOG	1		81.1%
HOF	1		80.4%
MBH	1		83.1%
SoPAF	1		79.6%
SoPAF+Sine	1		79.7%
SoPARF	**1**		**83.9%**
SoPARF+Sine	1		83.6%
HOG+HOF+MBH	3		84.7%
SoPARF+HOG+MBH	**3**		**86.0%**
(SoPARF+Sine)+HOG+MBH	**3**		**86.0%**

On Table 3, we show the results obtained on Hollywood2 dataset. With our SoPARF descriptor, we obtain better results than the related HOG, HOF and MBH descriptors of [38] and than our implementation of HOG, HOF and MBH descriptors. Especially, we improve by 4% the HOF of Wang et al. and by 6% our impementation of HOF. The SoPARF improves the SoPAF by 3%, although this comes at the price of slightly increasing the computational time and dimension of the resulting descriptor. When combining SoPARF with HOG and HOF, we obtain a mAP of 58.6% with linear classifier, slightly better than the results obtain by combining 4 descriptors in [38].

Table 3: Mean Average Precision on the Hollywood2 dataset ; ND : number of descriptors ; NL : non-linear classifiers ; * In [37] HOG/HOF descriptors are accumulated on over 100 spatio-temporal regions each one leading to a different BoW signature

Method	ND	NL	Results
Gilbert [11]	$\simeq 3$	X	50.9%
Ullah [37] HOG+HOF	2	X	51.8%
Ullah [37]	$2(\geq 100^\star)$	X	55.3%
Wang [38] traj	1	X	47.7%
Wang [38] HOG	1	X	41.5%
Wang [38] HOF	1	X	50.8%
Wang [38] MBH	1	X	54.2%
Wang [38] all	4	X	58.3%
HOG	1		49.6%
HOF	1		48.4%
MBH	1		53.1%
SoPAF	1		51.3%
SoPARF	**1**		**54.8%**
SoPARF+Sine	1		53.7%
HOG+HOF+MBH	3		56.4%
SoPARF+HOG+MBH	**3**		**58.6%**
SoPARF(+Sine)+HOG+MBH	3		58.5%

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. The use of the half-wave rectification coding improve the results of SoPAF descriptor. Moreover, we show the the possibility of using other basis for modeling the time evolution of spatial coefficients.

We carry out experiments on the well known UCF11 and Hollywood2 datasets, using recent signatures method from image classification techniques. We obtain improved results over popular descriptors such as HOG, HOF and MBH 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 and extending the coding step with other approaches.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. Transactions on Pattern Analysis and Machine Intelligence **32**, 288–303 (2010)
2. Avila, S., Thome, N., Cord, M., Valle, E., de A Araujo, A.: Bossa: Extended bow formalism for image classification. In: ICIP, pp. 2909–2912. IEEE (2011)
3. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. ECCV pp. 404–417 (2006)
4. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, vol. 2, pp. 1395–1402. IEEE (2005)
5. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC, vol. 76, pp. 1–12 (2011)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. ECCV pp. 428–441 (2006)
7. Danafar, S., Gheissari, N.: Action recognition for surveillance applications using optic flow and svm. In: ACCV, vol. 4844, pp. 457–466 (2007)
8. Davis, J., Bobick, A.: The representation and recognition of action using temporal templates. In: Conference on CVPR, pp. 928–934. IEEE (1997)
9. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV, vol. 2, pp. 726–733. IEEE (2003)
10. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: Conference on CVPR, pp. 1–8. IEEE (2008)
11. Gilbert, A., Illingworth, J., Bowden, R.: Action recognition using mined hierarchical compound features. Transactions on Pattern Analysis and Machine Intelligence **33**(5), 883–897 (2011)
12. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2247–2253 (2007)
13. Horn, B., Schunck, B.: Determining optical flow. Artificial intelligence **17**(1), 185–203 (1981)
14. Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition. In: ECCV 2010, pp. 494–507. Springer (2010)
15. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Conference on CVPR, pp. 3304–3311. IEEE (2010)
16. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. Transactions on Pattern Analysis and Machine Intelligence **34**, 1704–1716 (2012)
17. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. In: BMVC, pp. 885–894 (2008)
18. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Texture Based Description of Movements for Activity Analysis. In: VISAPP, vol. 1, pp. 206–213 (2008)
19. Kihl, O., Picard, D., Gosselin, P.H.: Local polynomial space-time descriptors for actions classification. In: IAPR MVA. Kyoto, Japon (2013)
20. Kihl, O., Tremblais, B., Augereau, B.: Multivariate orthogonal polynomials to extract singular points. In: ICIP, pp. 857–860. IEEE (2008)
21. Kihl, O., Tremblais, B., Augereau, B., Khoudeir, M.: Human activities discrimination with motion approximation in polynomial bases. In: ICIP, pp. 2469–2472. IEEE (2010)
22. KOGA, T.: Motion-compensated interframe coding for video conferencing. Proc. NTC, New Orleans (1981)
23. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Conference on CVPR, pp. 1–8. IEEE (2008)
24. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: Conference on CVPR, pp. 1996–2003. IEEE (2009)
25. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
26. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence, vol. 2, pp. 674–679 (1981)
27. Negrel, R., Picard, D., Gosselin, P.: Using spatial pyramids with compacted vlat for image categorization. In: ICPR, pp. 2460–2463 (2012)
28. Picard, D., Gosselin, P.H.: Improving image similarity with vectors of locally aggregated tensors. In: ICIP, pp. 669–672. IEEE (2011)
29. Picard, D., Gosselin, P.H.: Efficient image signatures and similarities using tensor products of local descriptors. CVIU **117**(6), 680–687 (2013)
30. Polana, R., Nelson, R.: Low level recognition of human motion. In: Proc. IEEE Workshop on Nonrigid and Articulate Motion, pp. 77–82 (1994)
31. Sánchez, J., Perronnin, F., Campos, T.d.: Modeling the spatial layout of images beyond spatial pyramids. Pattern Recognition Letters (2012)
32. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR, vol. 3, pp. 32–36. IEEE (2004)
33. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 405–412 (2005)
34. Shechtman, E., Irani, M.: Space-time behavior based correlation or how to tell if two underlying motion fields are similar without computing them? In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **29**(11), 2045–2056 (2007)
35. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV, vol. 2, pp. 1470–1477. IEEE (2003)
36. Tran, D., Sorokin, A.: Human activity recognition with metric learning. ECCV pp. 548–561 (2008)
37. Ullah, M., Parizi, S., Laptev, I.: Improving bag-of-features action recognition with non-local cues. In: BMVC (2010)
38. Wang, H., Klaser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: Conference on CVPR, pp. 3169–3176. IEEE (2011)
39. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009)
40. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Conference on CVPR, pp. 3360–3367. IEEE (2010)

41. Wang, L., Suter, D.: Learning and matching of dynamic shape manifolds for human action recognition. IEEE Transactions on Image Processing **16**(6), 1646 (2007)
42. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Conference on CVPR, pp. 1794–1801. IEEE (2009)
43. Zhou, X., Yu, K., Zhang, T., Huang, T.: Image classification using super-vector coding of local image descriptors. Computer Vision–ECCV 2010 pp. 141–154 (2010)