Unified framework for human behaviour recognition: An approach using 3D Zernike moments

A. Bouziane, Youssef Chahir, M. Molina, F. Jouen

Unified framework for human behaviour recognition: An approach using 3D Zernike moments

A. Bouziane [a,b,*], Y. Chahir [a], M. Molina [c], F. Jouen [d]

[a] Computer Science Department, GREYC-UMR CNRS 6072, University of Caen, France
[b] MSE Laboratory, University of Bordj Bou Arreridj, Algeria
[c] Psychology Department, PALM Laboratory, University of Caen, France
[d] Department of Cognitive Psychology.CHArt Laboratory, Practical School of High Studies (EPHE) Paris, France

Keywords:
3D Zernike moment
Video analysis
Human action
Hand gesture
High-dimensional data
Graph-based learning
Spectral graph embedding
Regularization on graphs
Manifold

A B S T R A C T

In this paper, we present a unified framework for the analysis of video databases by using Markov spatio-temporal random walks on graph. The proposed framework provides an efficient approach for clustering, data organization, dimension reduction and recognition. The aim of our work is to develop a vision-based approach for human behaviour recognition. Our contribution lies in three aspects. First, we employ 3D Zernike moments to encode the object of interest in a video clip. Then, we propose a new method to represent the video database as a weighted undirected graph where each vertex is a video clip. The weight of an edge between two video clips is defined by a Gaussian kernel on their 3D Zernike moments and their respective neighbourhoods in the feature space. Our objective is to obtain a robust low-dimensional space through spectral graph embedding which provides efficient keypoints transcription into an euclidean manifold, and allows to achieve higher classification accuracy through agglomerative categorization. Finally, we describe a variational framework for manifold denoising based on p-Laplacian, thereby lessening the negative impact of outliers, enhancing keypoints classification and thus, boosting the recognition accuracy. The proposed method is tested on the Weizmann and KTH human action datasets and on a hand gesture dataset. The retrieved results using the 3D Zernike moments prove that the proposed method can effectively capture the form of the behaviours with low order moments. Moreover, our framework allows to classify various behaviours and achieves a significant recognition rate.

1. Introduction

In many applications, such as video surveillance, unusual behaviour detection and sign language translation, it is important to recognize the human activity in order to interpret its behaviour. This latter can be defined as a temporal succession of primitive actions being performed by the human subject along the video clip and that may be composed to form a complex activity. The recognition of actions and gestures remains a challenging research problem. Most of the methods proposed to deal with this problem are based on computing either 2D or 3D appearance models from a silhouette, where a typical important task consists in identifying different body parts such as heads, hands, feet and joints. Other methods seek to monitor and interpret the human behaviour using motion estimation and description techniques, such as optical flow. To recognize human behaviour, Ali et al. [1] use person silhouettes to classify a continuous set of actions by extracting skeleton properties from the shape. Star skeleton features have been introduced by Fujiyoshi et al. [2] to extract 2D posture from a silhouette in real time. A feature star distance is defined so that feature vectors could be mapped into symbols by Vector Quantization. Based on this distance, the classification of actions is achieved by Hidden Markov Models (HMM). Another approach presented by Ahmad et al. [3] consists in extracting optical flow and body shape features from multiple viewpoints to recognize actions. The silhouette of a person is represented by its lower-dimensional subspace using PCA, and each action is represented using a set of multi-dimensional discrete HMMs modelled independently for any viewing direction. Lawrence et al. in [4,5] have introduced the Gaussian Process Latent Variable Model (GPLVM) for non-linear dimensionality reduction allowing the visualization of high dimensional data. Unlike PCA, this method performs data mapping from the embedded space (latent space) to the data space. In [6], this model

* Corresponding author at: Computer Science Department, GREYC-UMR CNRS 6072, University of Caen, France.
E-mail addresses: abderraouf.bouziane@unicaen.fr, bouziane.abderraouf@gmail.com (A. Bouziane), youssef.chahir@unicaen.fr (Y. Chahir), Molina@unicaen.fr (M. Molina), Francois.Jouen@ephe.sorbonne.fr (F. Jouen).

is used to estimate the 3D articulated human pose from silhouettes. Inspired by Lawrence and Hyvarinen [5], Wang et al. in [7] have introduced the Gaussian Process Dynamic Model (GPDM) which comprises a low-dimensional latent space with associated dynamics, and a mapping from the latent space to an observation space. Effectively, GPLVM and GPDM can learn a non-linear mapping between the human motion parameter space and a latent space by providing an inverse mapping. They allow the description of the human motion in a low-dimensional latent space. Wang et al. in [7] use the GPDM to learn models of human pose and motion from high-dimensional motion capture data. Urtasun et al. in [8] use GPLVM and GPDM to learn prior models for tracking 3D human walking styles. They obtained good results even in the case of serious occlusion. Efros et al. in [9] compare two actions based on the features extracted from optical flow measurements in the spatio-temporal space. Manor and Irani [10] propose multi-scale distributions of temporal gradient for isolating and clustering events within long continuous video sequences.

To detect video events, Zhu et al. [11] propose a spatio-temporal descriptor composed of low level image attributes, such as image gradients and optical flows, to capture the characteristics of actions in terms of their appearance and motion patterns in a space–time cube. A set of bag-of-words (BoW) features are constructed from this descriptor at multiple spatial pyramid resolution levels. Then, the action category is identified by fusing the classification results of SVM classifiers at all spatial pyramid levels. In [12], Laptev et al. compare two actions by matching points of interest (Harris). Blank et al. [13] use a stack of points, where silhouettes are extracted and evaluated by using the Poisson equation for each point. Bobick and Davis suggest the use of motion energy images (MEI) and those of motion history (MHI) [14] to represent how an action is performed using different levels of intensity based on the time since the silhouette was captured. At another level, the methods proposed for hand gesture recognition are mainly divided into two approaches. The Data-Glove based approaches which use sensor devices for digitizing hand and fingers, and the vision based approaches which require only a camera. This poses a challenging problem for hand postures and gesture recognition regardless of a number of issues including the complicated nature of static and dynamic hand gestures, complex backgrounds, and occlusions. Early methods dedicated to the hand gesture recognition problem consist in detecting the presence of the color of markers on the fingers in order to identify which ones are active in the gesture. A review of existing methods for the interpretation of hand gestures is presented in [15]. Recent methods based on advanced computer vision techniques do not require markers. Concretely, and from a different point of view, these methods could be divided into two further main approaches: on the one hand, the 3D hand model based approaches [16], which require the use of geometrical models (mesh) and animation techniques to capture the hand articulations and motion. On the other hand, appearance based approaches [17], which use image features to model the visual appearance of the hand. In [18], Chang et al. propose the use of curvature space method to perform hand gesture recognition, which involves finding the boundary contours of the hand. Other computer vision tools, used for 2D and 3D hand gesture recognition, include specialized mappings architecture [19] and particle filters [20].

1.1. Outline

In this paper we develop a new framework which allows a vision-based approach for human behaviour recognition. The proposed framework is composed of three phases which may be executed separately but sequentially. First, we use 3D Zernike moments to describe the object of interest (OI) in the video clip. These descriptors are scale, rotation and translation invariants. In addition, they allow the capture of both temporal and spatial information. Extracting OI have its own problems, therefore, we assume that it has been already segmented and that it represents one person. After that, we propose a diffusion framework for dimensionality reduction which provides a sound and efficient framework for embedding coordinates and classifying videos in an euclidean manifold. In light of this we will describe a variational framework for manifold denoising to enhance keypoints classification, thereby lessening the negative impact of outliers onto our variational shape framework. The major novelties of our work lie in making the proposed approach more robust (less susceptible to particularities of the data or to noise) and achieving higher classification accuracy through agglomerative categorization. The proposed framework will be validated through two applications: (1) Human actions categorization and (2) A vision-based approach for object properties identification. The basic idea of our work is to consider all videos as a weighted graph, where its vertices (video clips) are represented by the 3D volumes (3D Zernike descriptors), and its edges represent the similarity between connected vertices. The weight of an edge between two video clips is defined by a Gaussian kernel on their 3D Zernike moments and their respective neighbourhoods in the feature space. The salient points of our Graph-based framework are:

1. The spectral clustering reduces the clustering problem to a graph partitioning problem. A spectral decomposition of this graph is achieved by computing the eigenvalue decomposition of the normalized graph Laplacian. Then, a low-dimensional euclidean manifold embedding is inferred from this decomposition. The proposed algorithm captures and exploits the similarity between patterns, in the complete graph, dynamically without training. As additional information, we use an euclidean distance between vertices in the measure space.
2. To enhance robustness, we proceed to a regularization of the graph that allows denoising and simplifying data. In this stage, we can select "reliable" set of neighbours for each vertex in its non-local neighbourhood. Thus, the clustering is performed in the inferred regularized euclidean manifold.

Note that our framework relies on three hypotheses:

1. Preservation of the distance relationship.
2. Uniformity of the elements sampling.
3. Convexity of the elements.

It is worth mentioning that at each phase, we will use a different graph (i.e.: a graph in the feature space to allow spectral embedding, and a graph in the measure space (the embedded space) to manifold regularization). Tests have been conducted on three video datasets. The obtained performance evaluation results show that our framework allows to classify efficiently the videos and achieve an accurate recognition rate. The rest of this paper is structured as follows: Section 2 presents the background and the preliminaries of the proposed framework. In Section 3, we detail how the 3D Zernike moments are used to describe the OI in the video clip. Section 4 gives an overview of our spectral embedding framework. We explain in Section 5, a discrete regularization on the graph in the embedded space to enhance data robustness. Experimental results are presented and commented in Section 6. In the last section, we conclude our paper and discuss future extensions.

2. Preliminaries

The basic idea behind our framework is to represent a database by a weighted undirected graph $G=(V,E,w)$, where $V=(v_1,v_2,\ldots,v_n)$, is a finite set of vertices representing a finite data set, and $E\subseteq V\times V$, a finite set of edges representing the similarity between connected vertices. Let $f(v)$ be a function defined on each vertex v of the dataset V, in a q-dimensional space, and represented by the tuple: $(f_1,f_2,\ldots f_q)\in\mathbb{R}^q$. Thus, every vertex $v\in V$ is assigned with a local feature vector denoted by $f(v)\in R^q$. Many choices can be considered for $f(v)$. In the simplest case, one can consider $f(v)=v$. There exist several popular methods that transform the set V with a given pairwise similarity measure w into a graph $G=(V,E,w)$. Among the existing methods, we can quote the ε-neighbourhood graph where two points u, $v\in V$ are connected by an edge if $\|f(u)-f(v)\|\le\varepsilon$, $\varepsilon>0$. We denote then by $u\sim v$ the fact that the vertex u belongs to the ε-neighbourhood of v ($u\in\mathcal{N}_\varepsilon(v)$), which is defined by:

$$\mathcal{N}_\varepsilon(v)=\{u\in V, f(u)=(f'_1,\ldots f'_q)\mid\ |f_i-f'_i|\le\varepsilon_i,\ 0<i\le q\} \tag{1}$$

Another important graph is the k-nearest neighbours graph where two points u, $v\in V$ are connected by an edge if u is among the k-nearest neighbours of v. For images, classical graph representations are the grid graph and the region adjacency graph.

Constructing similarity graphs consists in modelling local and non-local neighbourhood relationships between data points. The similarities between these points are estimated by comparing their respective features which depend, generally, on the function f and the set V. Then, let us define a non-local feature vector denoted by $F(v)\in R^p$, and computed from the patch surrounding the vertex v as follows:

$$F(v)=[f(u),u\in\mathcal{B}_s(v)\subseteq\mathcal{N}_\varepsilon(v)]^T \tag{2}$$

$\mathcal{B}_s(v)$ is a bounding box of size s centred at v. Therefore, the weight function w associated to a graph G can incorporate local and/or non-local features according to the topology of the considered graph. It gives a measure of the similarity between a vertex and its neighbours that can incorporate local and non-local features, and is defined as follows:

$$w(u,v)=\begin{cases}\exp\left(-\dfrac{\|f(u)-f(v)\|^2}{h_1^2}\right).\exp\left(-\dfrac{\|F(u)-F(v)\|^2}{h_2^2}\right) & \text{for each } u\sim v\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

The scale parameter h_i can be estimated using the standard deviation depending on the variations of $\|f(u)-f(v)\|$ and $\|F(u)-F(v)\|$ over the graph, respectively.

3. 3D Zernike moments

To interpret the human behaviour within a video clip, the Object of Interest (OI), should be extracted and identified before its characterization. Extracting voxels is itself a difficult task having its own problems and is not studied in this paper. We assume that the OIs have been already extracted by using an adequate method such as the graph cut algorithm [21,22], which is useful to separate objects from the background. In a previous work [23], we have used this algorithm to perform an interactive video object segmentation. That was motivated by the fact that it is allowed for a straightforward incorporation of prior knowledge into its formulation.

Each OI is referred to by a vertex $v\in G(V)$ and can be viewed as a binary volume expressed by $g(x,y,t)$. For each voxel of the OI, let g_0 represent its initial color/intensity and, x,y,t its spatio-temporal coordinates. To represent how an action is performed, we use 3D Zernike moments to describe the spatio-temporal features (x,y,t) of the object of interest.

Let $Z^v_{nlm}(x,y,t)$ be the 3D Zernike functions:

$$Z^v_{nlm}(x,y,t)=R_{nl}(r)\cdot Y_{lm}(\theta,\phi) \tag{4}$$

where $R_{nl}(r)$ is the radial term, and $Y_{lm}(\theta,\phi)$ are the spherical harmonics of the l-th degree orthonormal on the surface of the unit sphere with m ranging from $-l$ to l and $n-l$ being an even non-negative integer ($n-l=2k$). The equality $(n-l)/2=k$ is the orthonormality condition of the 3D Zernike polynomials inside the unit sphere (more details are in [24]). $Y_{lm}(\theta,\phi)$ is the angular term. Z^v_{nlm} can be written in a more compact form as a linear combination of monomials of order up to n:

$$Z^v_{nlm}(x,y,t)=\sum_{p+q+r\le n}\mathcal{X}^{pqr}_{nlm}x^py^qt^r \tag{5}$$

where, for $k=(n-l)/2$:

$$\mathcal{X}^{pqr}_{nlm}=c_{lm}2^{-m}\sum_{s=0}^{k}q_{kls}\sum_{\alpha=0}^{s}\binom{s}{\alpha}\sum_{\beta=0}^{s-\alpha}\binom{s-\alpha}{\beta}\sum_{r=0}^{m}(-1)^{m-r}\binom{m}{r}$$
$$i^r\sum_{\mu=0}^{(l-m)/2}(-1)^\mu 2^{-2\mu}\binom{l}{\mu}\binom{l-\mu}{m+\mu}\sum_{s=0}^{\mu}\binom{\mu}{s} \tag{6}$$

and the normalization factor C_{lm} is given by:

$$c_{lm}=\frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!} \tag{7}$$

and

$$q_{kls}=\frac{(-1)^k}{2^{2k}}\sqrt{\frac{2l+4k+3}{3}}\binom{2k}{k}(-1)^s\frac{\binom{k}{k}\binom{2(k+l+s)+1)}{2k}}{\binom{k+l+s}{k}} \tag{8}$$

Since Z^v_{nlm} forms a complete orthonormal system, it is possible to approximate the original function g by a finite number of 3D Zernike moments Ω^v_{nlm} as follows:

$$g(x,y,t)=\sum_{n=0}^{\infty}\sum_{l=0}^{n}\sum_{m=-l}^{l}\Omega^v_{nlm}Z^v_{nlm}(x,y,t) \tag{9}$$

The 3D Zernike moments are defined by

$$\Omega^v_{nlm}=\frac{3}{4\pi}\sum_{p+q+r\le n}(-1)^m\mathcal{X}^{pqr}_{nlm}m^v_{pqr} \tag{10}$$

m^v_{pqr} denotes the geometrical moments of order $(p+q+r)$ of the binary volume, and is defined by:

$$m^v_{pqr}=\sum_{x=0}^{N_x-1}\sum_{y=0}^{N_y-1}\sum_{t=0}^{N_t-1}x^py^qt^rg(x,y,t) \tag{11}$$

The choice of the maximum order among the 3D Zernike moments is crucial to describe the subject behaviour and consequently, carry more or less details on the video binary volume. It is selected experimentally to form the descriptor vector $f(v)$ of each video v. It is defined by $(2l+1)$ moments as follows:

$$f(v)=\{\mathcal{V}^v_{nl}=\|\Omega^v_{nlm}\|:n\in[0,N],l\in[0,n],m\in[-l,l]\} \tag{12}$$

The distance between two videos represented by u and v, respectively, is computed using their 3D Zernike moments as follows:

$$\|f(u)-f(v)\|=\|\mathcal{V}^u_{nl}-\mathcal{V}^v_{nl}\|=\sqrt{\sum_{n=0}^{N}\sum_{l=0}^{n}(\mathcal{V}^u_{nl}-\mathcal{V}^v_{nl})^2} \tag{13}$$

4. Behaviour subspace embedding

The goal of subspace embedding is to find an optimized low-dimensional space where relevant information is captured, and similarity between 3D Zernike moments, and therefore between video clips, can be easily expressed. So, from the set of the 3D Zernike moments $\mathcal{V} \in \mathbb{R}^p$, obtained in the previous phase of the framework, we develop an appropriate euclidean mapping $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\} \in \mathbb{R}^q$, representing the low-dimensional space ($q \ll p$). To this end, we construct an undirected graph G on $\mathcal{V}$ to learn a kernel matrix that represents the provided side-information as well as the local/nonlocal geometry of the features. Through the eigendecomposition of the matrix associated to the random walks on G, we define a diffusion distance $D_t(y_i, y_j)$. This transformation is now commonly used for data analysis and dimension reduction [25,26].

The transition matrix P on G given by: $P = \{p^{(1)}(u,v) = w(u,v)/\sum_{v \sim u} w(u,v)\}$, explicits all possible *one time step* transitions, and provides therefore, the first order information of the graph structure. Let P^t be the t power of the matrix P that denotes the set of all transition probabilities $p^{(t)}(u,v)$ of going from one vertex to another one in *t-time steps*. This *t-time steps* transition probability satisfy the Chapman–Kolmogorov equation, that for any k such that $0 < k < t$:

$$p^{(t)}(u,v) = \Pr(X_t = v | X_0 = u) = \sum_{y \in V} p^{(k)}(u,y) \cdot p^{(t-k)}(y,v) \tag{14}$$

For clustering purposes, a connection with the spectral decomposition of P^t is established (see for detail [27]) to generate an euclidean coordinates for the low-dimensional representation of the vertices of the graph G at time t, where for each vertex, these coordinates are given by:

$$\Psi_t(u) = (\lambda_1^t \psi^1(u), \lambda_2^t \psi^2(u), \ldots, \lambda_n^t \psi^n(u))^T \tag{15}$$

$\{\lambda_i^t,\ \psi^i(u)\}$ are the eigenvalues and the eigenvectors associated with the normalized graph Laplacian of P^t. They correspond to the non-linear embedding of the vertices of the graph G onto the new euclidean low-dimensional space. Thus, the *diffusion distance*, $D_t^2(u,v)$, between the vertices (3D Zernike moments) of the graph G can be expressed in the embedded space by

$$D_t^2(u,v) = \sum_{i \geq 1} \lambda_i^{t2} (\psi_i(u) - \psi_i(v))^2 = \|\Psi_t(u) - \Psi_t(v)\|^2 \tag{16}$$

We note in particular that this new distance depends on the time parameter t which is considered here as a precision parameter, where for large values, more information on the structure of the graph is captured. The retrieved eigenvalues are unique and ordered so that: $1 = |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$. Consequently, the first largest eigenvalues and eigenvectors carry the relevant information of the initial complex vectors (3D Zernike moments), and are well-suited to define the new euclidean coordinates. Practically, one can use the subjective scree-test of Cattell [28] to determine the most important k-th dimensions that catch the pertinent information. This criterion is based on the analysis of differences between consecutive eigenvalues, where a breakpoint would be located where there is the biggest change in the slope of the curve of eigenvalues. The first k-th eigenvalues correspond, then, to the number of dimensions to retain (see for detail [29]). Another simple way is to consider the first dominant eigenvalues for which their sum is great than a defined threshold (e.g. $\geq 80\%$).

Applying a clustering algorithm using these new coordinates allows to categorize the actions of the video database. Fig. 1 shows a clustering of the KTH dataset actions in the feature space (i.e. by using the 3D Zernike moments vectors directly, (a class of action per colour)). We can easily point out the ambiguities in actions classification, particularly between *jogging*, *walking*, *running and boxing*.

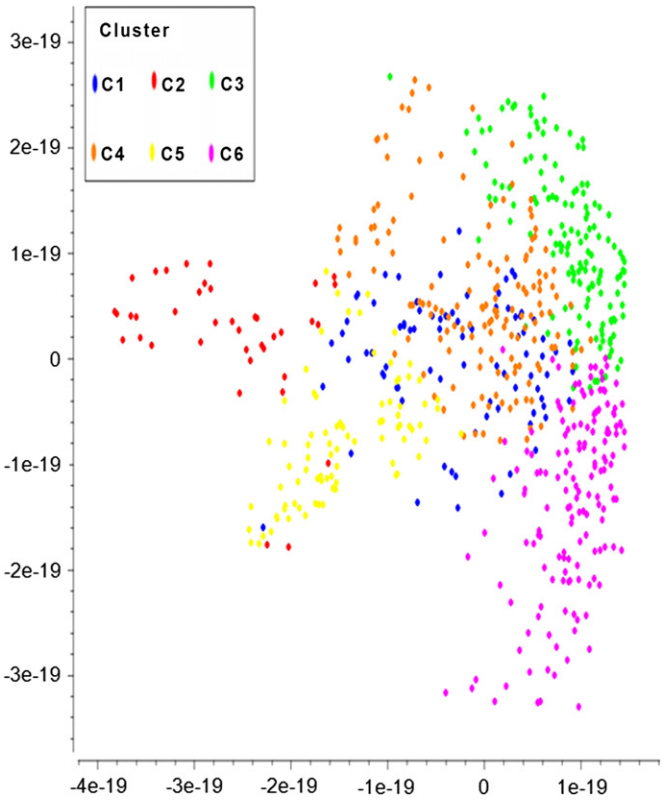

Fig. 1. The KTH actions, projected in the feature space.

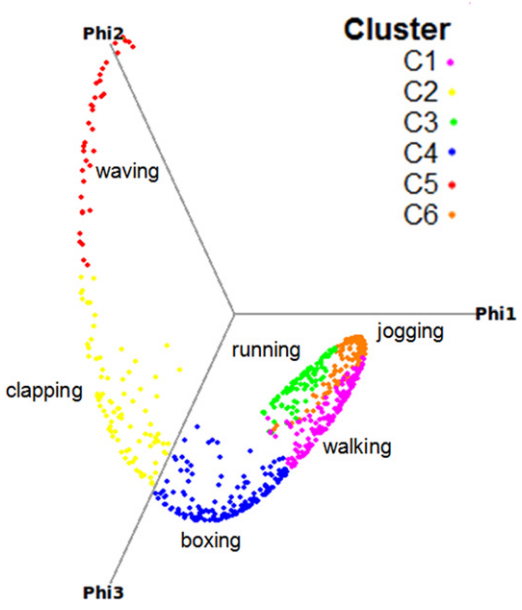

Fig. 2. The KTH actions, projected in the reduced space.

In contrast, Fig. 2 shows the projection of same dataset actions in the reduced space. Each action is represented by the first 10-th coordinates, corresponding to the low space mapping of the 3D Zernike moments. For display convenience, we keep only the first three axes. Here the actions classes are correctly separated and the dataset categorization is clearly visible.

It is worth mentioning that for large databases one can use iterative solvers to determine the eigenvalues of a subset, and

therefore, performs clustering on a restricted set. In [30], a description of the Nyström method is presented. This latter can be used also for incremental clustering, especially in the case where the size of the database is not known in advance, and is filled in as and when capturing videos. This would be very suitable for online video processing.

5. 3D eigenbehaviour regularization

Our motivation for this section is to transcribe the variational methods on a discrete graph. For that purpose, we propose to extend the scope of discrete regularization [31] to high-dimensional data. We have implemented algorithms for regularization on graphs with p-Laplacian with $p \in]0,+\infty[$, for denoising and simplification of data in embedded space. Readers can refer for further details on this formalism to [32].

Recall that the function f^0 is an observation of an original function f affected by noise n: $f^0 = f + n$. The discrete regularization of $f^0 \in \mathcal{H}(V)$ using the weighted Laplacian operator consists to seek a function $f^* \in \mathcal{H}(V)$ which is smooth enough on G, and sufficiently close to f^0. Variational models of regulation can be described by the following minimization problem:

$$f^* = \min_{f \in \mathcal{H}(V)} \left\{ \frac{1}{2} \sum_{v \in V} \|\nabla f_v\|_2^2 + \frac{\lambda}{2} \|f - f^0\|_{\mathcal{H}(V)}^2 \right\} \tag{17}$$

The fidelity parameter λ specifies the trade-off between the first energy term, the smoothness term or regularizer, and the second term, called fitting term. The solution of this regularization problem can be obtained by using the Gauss–Jacobi iterative algorithm presented as follows, where, for all (u,v) in E:

$$\begin{cases} f^{(0)} = f^0 \\ f^{k+1}(v) = \dfrac{1}{\lambda + \sum_{u \sim v} w(u,v)} (\lambda f^0(v) + \sum_{u \sim v} w(u,v) f(u)^k) \end{cases} \tag{18}$$

The new value $f^{(k+1)}(v)$ depends on the original value $f^0(v)$ and a weighted average of the existing values in a neighbourhood of v. Fig. 3 represents the projection of videos dataset over the three principal axis of the embedded space. The graph is built in this space and the new coordinates are classified. As we can see, we observe the difference between Fig. 3-left and Fig. 3-right. More simplification of the graph is obtained, the manifold shape is more clear and the classification process is improved when the manifold is regularized.

6. Experimental validation

To assess the reliability of our approach, we consider two applications: (1) human actions categorization and (2) a vision-based approach for objects properties identification. We conduct our experiments on three videos datasets. Two of them are devoted to human actions categorization, namely: the KTH [12] and Weizmann datasets [13]. The Third dataset, EPHE dataset, contains hand gestures video clips and is the property of the "Ecole Pratique des Hautes Etude, Sorbonne". The KTH and Weizmann datasets are widely studied under various aspects, and multiple results are available. Therefore, they offer an interesting challenging benchmark to evaluate our results. Moreover, the hand gestures dataset allows to test our framework under another aspect, and thus consolidate the obtained results. We recall that our framework is composed of three phases where, first, a binary volume, representing the object of interest (OI) is characterized by using 3D Zernike moments. Secondly, an euclidean low-dimensional mapping of these descriptors is computed through spectral graph embedding and, finally, to boost the classification accuracy, a graph regularization in the inferred embedded space is performed.

6.1. Human actions categorization

The silhouettes of the OI are provided with both the KTH and Weizmann datasets. Thus, one can describe them directly by using the 3D descriptors.

6.1.1. The Weizmann dataset

This dataset contains a total of 90 video clips performed by different individuals. Each video clip contains one person performing an action. There are ten categories of actions involved in the dataset, namely, *walk*, *run*, *skip*, *jack*, *jump*, *jump in place*, *side*, *wave with one hand*, *wave with two hands*, *bend*, and they are

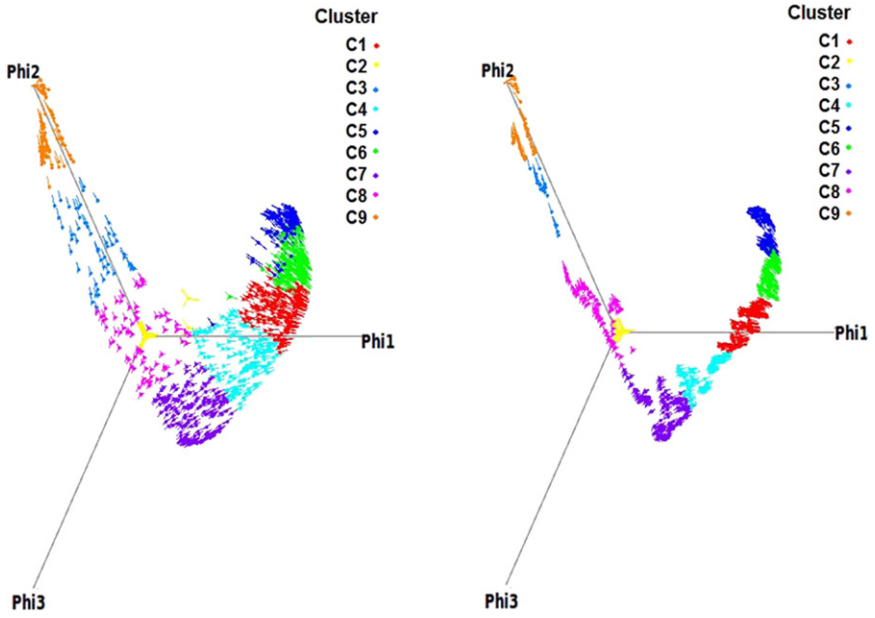

Fig. 3. Discrete regularization of manifold with linear Laplacian.

referred to, hereafter, by (a1, a2, a3, a4, a5, a6, a7, a8, a9, a10), respectively. Fig. 4 shows some clips from this dataset.

The choice of the proper order among the 3D Zernike moments is crucial to describe the subject behaviour and consequently carry more or less details on the video binary volume. A common approach consists in choosing the moment order which allows a better reconstruction of the Object of Interest. In [33], a grey-level image was reconstructed using Zernike moments of increasing order. It was illustrated that using Zernike moments of order 6 resulted in a reconstruction with an error of 10%. This error was of 6% using Zernike moments of order 20. However, this result may vary depending on the case studied, and the optimal order depends on the nature of the objects to reconstruct and therefore cannot be generalized. Moreover, in terms of sensitivity to additive random noise, it has been shown in [34] that in the presence of noise, the image is best reconstructed by using moments of up to a certain optimal order. Reconstructing the image by using moments of orders higher than the optimal order will degrade its quality because higher order moments are more vulnerable to white noise. In this paper, we are primarily interested in human activity recognition and consequently in the overall performance of the proposed framework. To select a proper order of moment which allows a better recognition rate, we have tested different ones. Before performing the regularization phase, Table 1 shows the changes of the recognition rates with different orders of moment. As we can see, a higher recognition rate is obtained by using the moment of order 7, which has practically no difference with that obtained by using the moment of order 6.

Using the moment of order 7 and after the graph regularization in the reduced space, we obtain the confusion matrix for the 10 actions of this dataset shown in Table 2.

It is quite clear that performing discrete regularization in the inferred reduced space enhances significantly the actions recognition rates. Overall, the mean accuracy is **96.33%**, whereas for the same order, without regularization, the mean accuracy is **91.08%**. The results obtained from this experiment are compared with those reported in other works [10,35–38] (see Table 3). From this comparison, it turns out that our method performs competitively with other state-of-the-art methods, and achieves encouraging results compared with some previously published ones.

Fig. 4. Example of actions from Weizmann dataset.

6.1.2. The KTH dataset

We have tested our approach on a second dataset: the KTH human motion dataset. The video clips of this dataset include six types of human actions (i.e., walking, jogging, running, boxing, hand waving, and hand clapping). Each of these actions is performed by a total of 25 individuals in four different settings (i.e., outdoors, outdoors with scale variation, outdoors with different clothes, and indoors). Fig. 5 shows some clips from this dataset.

After characterizing the OIs by the 3D Zernike moments, a low-dimensional mapping is inferred, and new euclidean coordinates are generated for each OI. The projection of the video clips in this reduced space by using these new coordinates is illustrated in Fig. 6.

Though, the video dataset is clearly categorized, still some ambiguities remain. These ambiguities are better separated in Fig. 7, which projects the same video clips in the regularized reduced space.

Table 4 illustrates the confusion matrix for this dataset. We can see that the results obtained with regularization are considerably better. Overall, the mean accuracy is **95.17%**.

To confirm the reliability of our framework, the results obtained for this experiment are, also, compared with those obtained with other state-of-the-art methods [35,37–40] (see Fig. 8 and Table 5). As we can see, the results are compared favourably and our contribution is contrasted.

6.2. A vision-based approach for objects properties identification

In a similar way, we have also tested our approach for hand gesture recognition. The key problem, in this context, is how to make hand gestures understood by computers in order to recognize some object properties using only a camera. In this paper, we aim to recognize the texture and the consistency of an object through the video analysis of the hand actions. Two properties of the object are studied: the texture, which could be either smooth

Table 2
Confusion matrix by using 3D Zernike moments of order 7.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	100	0	0	0	0	0	0	0	0	0
a2	2	98	0	0	0	0	0	0	0	0
a3	0	3	97	0	0	0	0	0	0	9
a4	0	0	0	100	0	0	0	0	0	0
a5	10	0	0	0	83	7	0	0	0	0
a6	0	0	2	0	0	98	0	0	0	0
a7	0	0	0	0	0	0	96.3	1.3	2.4	0
a8	0	0	0	0	2	0	0	96	0	2
a9	0	0	0	0	0	0	0	4	96	0
a10	0	0	0	0	0	0	0	1	0	99

Table 1
Actions recognition rates comparison.

Moment order	Actions recognition rates (%)										Mean accuracy
	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	
Order7	93.1	95	87.2	96	79	93	89	91.3	93.2	94	**91.08**
Order6	92.5	94.8	86.4	95.9	78.2	93	88.8	90.6	92.9	93.7	**90.68**
Order5	91.1	90.7	85.9	92	77.3	84.8	84.5	85.2	87.1	91	**86.96**
Order4	88.7	86	84.7	85	75	76.3	79.3	78	80.7	88.3	**82.2**
Order3	85	80	83	75	70	65	73	68	71	81	**75.1**

Table 3
Performance comparison on Weizmann dataset.

Method	Actions recognition rates (%)										Mean accuracy
	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	
Our approach	100	98	97	100	83	98	96.3	96	96	99	**96.33**
Kellokumpu et al. [37]	100	100	100	100	89	100	100	100	100	100	**98.9**
Xinghua et al. [35]	100	90	90	100	100	100	100	100	100	100	**97.8**
Dhillon et al. [38]	91	93	69	94	92	92	90	91	92	95	**89.9**
Vezzani et al. [36]	100	99	68	87	81	95	57	100	86	94	**86.7**
Zelnik et al. [10]	82.4	34.7	43.5	95.5	29.2	84.9	50.8	29.6	51.9	86.6	**58.91**

Fig. 5. Example of actions from KTH dataset.

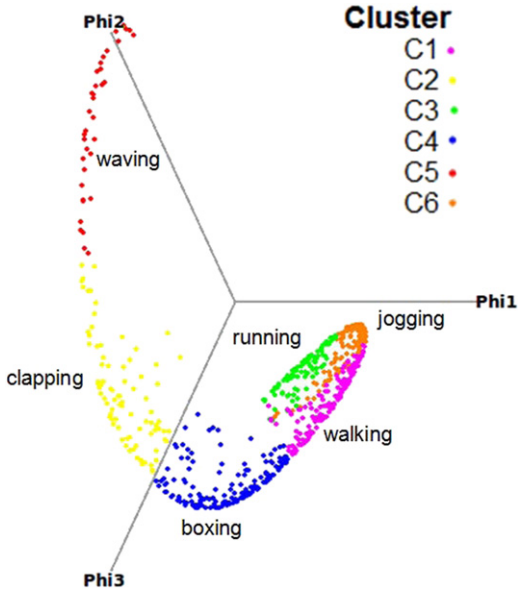

Fig. 6. The KTH actions, projected in the reduced space.

or granular, and the consistency, which could either be hard or soft. We are inspired from the seminal description of Lederman and Klatzky [41], and we studied four exploratory procedures:

1. Lateral Motion for Smooth Object (*LMSO*).
2. Lateral Motion for Granular Object (*LMGO*).
3. Pressure for Soft Object (*PSO*).
4. Pressure for Hard Object (*PHO*).

We conducted experiments on the EPHE corpora composed of 21 gesture video clips representing manipulation of different

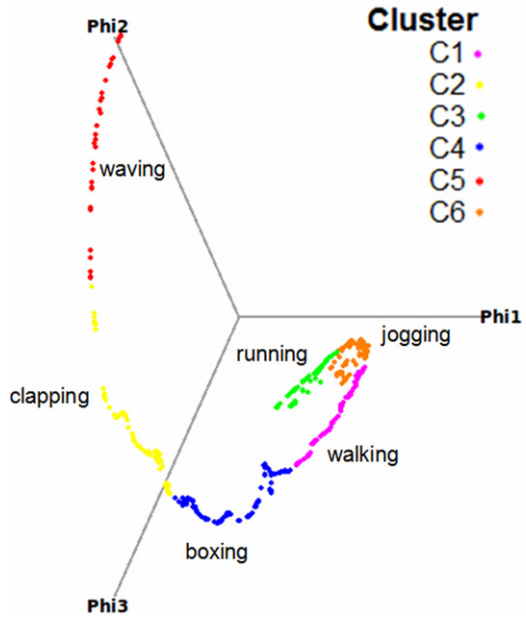

Fig. 7. The KTH actions, projected in the regularized reduced space.

Table 4
Confusion matrix without and with regularization.

	bx	jg	rg	wg	hd-cl	hd-wa
Boxing	95.6/**98**	0	0	3/2	1.4/0	0
Jogging	1.6/0	84.2/**89**	6.2/6	8/5	0	0
Running	0	3/1	95.4/**98**	1.6/1	0	0
Walking	1/0	2/0	0	97/**100**	0	0
Hand-clapping	6/5	0	0	0	89.5/**92**	4.5/3
Hand-waving	5.6/4	0	0	0	5/2	89.4/**94**

objects by both the left and the right hand. Fig. 9 shows the manipulated objects and some frames extracted from the video clips.

To verify the effectiveness of the 3D Zernike moments in our approach, we applied the same process separately on each hand (left/right hand). Fig. 10 shows the result of projecting separate hands by using the Fiedler vector, which allows to categorize left and right hands, respectively. For the classification of the gestures, we opted for the separation of the two hands. Each hand is represented by its 3D moments. The distance between two gestures is calculated by:

$$\frac{\|f(v)^{right}-f(u)^{right}\|+\|f(v)^{left}-f(u)^{left}\|}{2}.$$

The result of the recognition of the left hands and right hands in these videos was 100%. The recognition rates of the gestures in the video clips is summarized in Table 6. Once more, discrete

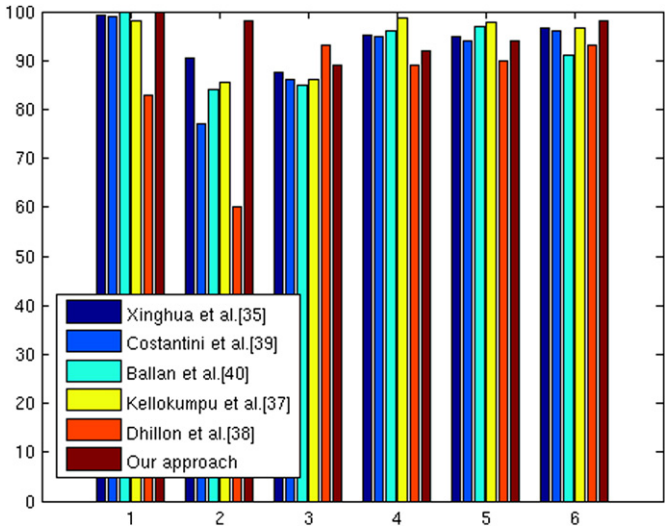

Fig. 8. Actions recognition based on our approach compared with some state of the art methods (KTH dataset).

Table 5
Performance comparison on KTH dataset.

Methods	Rates (%)
Our approach	95.17
Xinghua et al. [35]	94.0
Costantini et al. [39]	91.17
Ballan et al. [40]	92.17
Kellokumpu et al. [37]	93.77
Dhillon et al. [38]	84.67

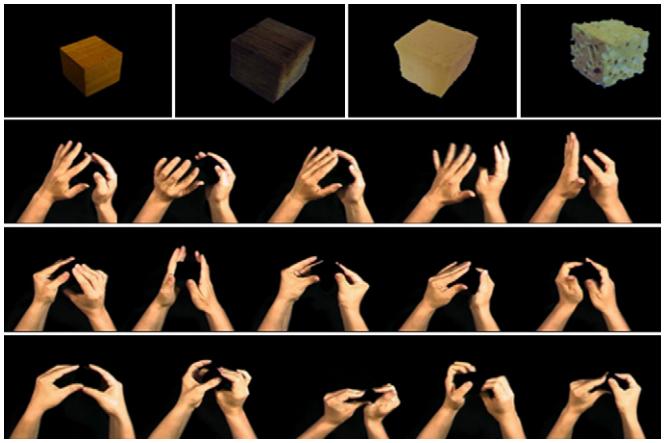

Fig. 9. Manipulated objects and frames extracted from the video clips.

graph regularization in the inferred reduced space allows to achieve better results. This consolidates the results obtained in the first application and, thus, confirms the effectiveness of our approach.

7. Conclusions

In this paper we propose a new method for the recognition of human behaviour based on an extension of the Zernike moments to the spatio-temporal domain in order to characterize human actions in video clips. In fact, 3D Zernike moments have

Fig. 10. Fiedler vector to order left and right hands.

Table 6
Gesture recognition based on 3D Zernike moments without and with regularization.

Recognised gesture	**Recognition rates (%)**	
	Without regularization	With regularization
LMSO	90	93.6
LMGO	85.7	89.2
PSO	71.4	77.4
PHO	80	85.9

interesting properties for describing structural and temporal information of a time varying video sequence. Our approach is composed of three main steps. The first one uses 3D Zernike moments to describe the silhouette and the dynamic of the object of interest (OI) in the video clip. Next, in the second step, we construct a visual similarity network, by computing the pairwise similarity, based on the latter 3D features. The basic idea behind this step is to consider all videos as a weighted graph whose vertices (video clips) are represented by the 3D volumes (3D Zernike descriptors), and edges represent the similarity between connected vertices. Finally, we describe a variational approach for manifold denoising that allows us to exploit the geometry of the data distribution.

We validated this framework through two applications: (1) Human actions categorization and (2) a vision-based approach for objects properties identification. In light of the obtained results for both applications, it turns out that our framework achieves a significant recognition rates with respect to other state-of-the-art methods and that the 3D Zernike moments can effectively capture the form of the behaviours with low order moments, which confirms the effectiveness of the proposed approach.

Our framework has a natural connection to reproducing Kernel Hilbert Space. Under this space, we can consider the problem of learning a relationship between two structured input and output

datasets from labelled and unlabelled samples. Thus, as an extension to the current work, one can consider the use of incremental solvers like the nyström extension to categorize large video databases where a subset has been already classified. Indeed, this can be very useful for many applications such as online video categorization.

Acknowledgments

This paper is supported by the PRETHERM ANR project, co-financed by the French National Research Agency (ANR).

References

[1] A. Anjum, J.K. Aggarwal, Segmentation and recognition of continuous human activity, in: IEEE Workshop on Detection and Recognition of Events in Video'01., IEEE Computer Society, 2001, pp. 28–28.
[2] H. Fujiyoshi, A.J.Lipton, Real-time human motion analysis by image skeletonization, in: Proceedings of IEEE WACV98, IEEE Computer Society, 1998, pp. 15–21.
[3] M. Ahmad, S.-W. Lee, Hmm-based human action recognition using multiview image sequences, in: Proceedings of the 18th International Conference on Pattern Recognition, vol. 01, ICPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 263–266.
[4] L.N, Gaussian process latent variable models for visualization of high dimensional data, in: NIPS, 2003.
[5] N. Lawrence, A. Hyvarinen, Probabilistic non-linear principal component analysis with Gaussian process latent variable models, J. Mach. Learn. Res. 6 (2005) 1783–1816.
[6] C.H. Ek, P.H.S. Torr, N.D. Lawrence, Gaussian process latent variable models for human pose estimation, in: Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction, MLMI'07, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 132–143.
[7] J.M. Wang, D.J. Fleet, A. Hertzmann, Gaussian process dynamical models for human motion, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008) 283–298.
[8] R. Urtasun, 3D people tracking with Gaussian process dynamical models, in: CVPR, CVPR, 2006, pp. 238–245.
[9] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV '03, vol. 2, IEEE Computer Society, Washington, DC, USA, 2003, pp. 726–733.
[10] L. Zelnik-manor, M. Irani, Event-based analysis of video, in: 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kuai, HI, USA, pp. 123–130.
[11] G. Zhu, M. Yang, K. Yu, W. Xu, Y. Gong, Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor, in: Proceedings of the 17th ACM International Conference on Multimedia, MM '09, ACM, New York, NY, USA, 2009, pp. 165–174.
[12] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), ICPR '04, vol. 03, IEEE Computer Society, Washington, DC, USA, 2004, pp. 32–36.
[13] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV '05, IEEE Computer Society, Washington, DC, USA, 2005, pp. 1395–1402.
[14] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 257–267.
[15] V. Pavlovic, R. Sharma, T. Huang, Visual interpretation of hand gestures for human–computer interaction. A review, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 677–695.
[16] Y. Wu, J. Y. Lin, T.S. Huang, Capturing natural hand articulation, in: Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada II, 2001, pp. 426–432.
[17] Q. Chen, N.D. Georganas, E.M. Petriu, Real-time vision based hand gesture recognition using haar-like features, in: Proceeding of the IEEE Instrumentation and Measurement Technology Conference Proceedings, May 1–3, 2007, Warsaw, pp.1–6, http://dx.doi.org/10.1109/IMTC.2007.379068.
[18] C. Chang, I. Chen, Y. Huang, Hand pose recognition using curvature scale space, IEEE International Conference on Pattern Recognition, 2002.
[19] R. Rosales, V. Athitsos, L. Sigal, S. Sclaroff, 3D hand pose reconstruction using specialized mappings, IEEE International Conference on Computer Vision (2001) 378–385.
[20] L. Bretzner, I. Laptev, T. Lindberg, Hand gesture recognition using multi-scale color features, hierarchical models and particle filtering, IEEE International Conference on Automatic Face and Gesture Recognition, 2002.
[21] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 1124–1137.
[22] Y. Boykov, V. Kolmogorov, Computing geodesics and minimal surfaces via graph cuts, in: ICCV, 2003, pp. 26–33.
[23] Y. Chahir, M. Molina, F. Jouen, B. Safadi, Haptic gesture analysis and recognition, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008.
[24] N. Canterakis, 3D zernike moments and zernike affine invariants for 3D image analysis and recognition, in: In 11th Scandinavian Conference on Image Analysis, 1999, pp. 85–93.
[25] N. Boaz, S. Lafon, R. Coifman, Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators, Neural Inf. Process. Syst. 18 (2005).
[26] S. Lafon, Y. Keller, R. Coifman, Data fusion and multicue data matching by diffusion maps, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1784–1797.
[27] B. Nadler, S. Lafon, R. Coifman, I.G. Kevrekidis, Diffusion maps—a probabilistic interpretation for spectral embedding and clustering algorithms, in: A.N. Gorban, B. Kegl, D.C.WunschnA. Zinovyev (Eds.), Principal Manifolds for Data Visualization and Dimension Reduction, Springer, 2007.
[28] R.B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. (1966) 245–276.
[29] G. Raiche, M. Riopel, J.G. Blais, Non graphical solutions for the Cattell's scree test, International Annual Meeting of the Psychometric Society, 2006.
[30] P. Arias, G. Randall, G. Sapiro, Connecting the out-of-sample and pre-image problems in kernel methods, in: Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2007.
[31] M. Ghoniem, Y. Chahir, A. Elmoataz, Nonlocal video denoising, simplification and inpainting using discrete regularization on graphs, J. Signal Process. 90 (2010) 2445–2455.
[32] A. Elmoataz, O. Lezoray, S. Bougleux, Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing, IEEE Trans. Image Process. 17 (2008) 1047–1060.
[33] J. Boyce, W. Hossack, Moment invariants for pattern recognition, Pattern Recogn. Lett. 1 (1983) 451–456.
[34] C.-H. Teh, R.T. Chin, On image analysis by the methods of moments, IEEE Trans. Pattern Anal. Mach. Intell. 10 (1988) 496–513.
[35] S. Xinghua, C. Mingyu, A. Hauptmann, Action recognition via local descriptors and holistic features, Computer Vision and Pattern Recognition Workshop 0 (2009) 58–65.
[36] R. Vezzani, D. Baltieri, R. Cucchiara, Hmm based action recognition with projection histogram features, in: Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos, ICPR'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 286–293.
[37] V. Kellokumpu, G. Zhao, M. Pietikäinen, Recognition of human actions using texture descriptors, Machine Vision and Applications 22 (5) (2011) 767–780. http://dx.doi.org/10.1007/s00138-009-0233-8.
[38] P. Dhillon, S. Nowozin, C. Lampert, Combining appearance and motion for human action classification in videos, Computer Vision and Pattern Recognition Workshop (2009) 22–29.
[39] L. Costantini, L. Seidenari, G. Serra, L. Capodiferro, A. Del Bimbo, Space-time zernike moments and pyramid kernel descriptors for action classification, in: ICIAP (2), Lecture Notes in Computer Science, Springer, 2011, pp. 199–208.
[40] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra, Recognizing human actions by fusing spatio-temporal appearance and motion descriptors, in: Proceedings of the 16th IEEE International Conference on Image Processing, ICIP'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 3533–3536.
[41] S. Lederman, R.R.L. Klatzky, Hand movements: a window into haptic object recognition, Cognitive Psychology 19 (1987) 342–368.

A. Bouziane Associate Professor at the computer science department at BBA university. He is a member of the MSE laboratory. Currently, he is an invited researcher at GREYC Laboratory. His research interest fields include spectral analysis, organization and indexing of highdimensional multimedia data.

Y. Chahir Professor at the computer science department at Lower Normandy university. He is a member of the Image team at the GREYC laboratory. His research interest fields include Image and Video Processing & Analysis, Multimedia data-Mining, Spectral analysis and Restitution & Animation in Virtual Environment.

M. Molina Professor in developmental psychology. She is a director of PALM Laboratory at University of Caen. Personal Interests: Developmental psychology, Developmental psychobiology, Complex systems modeling and simulation.

F. Jouen Professor and Director of study at Ecole Pratique des Hautes Etudes, Sorbonne Paris. He is a director of CHART Laboratory at EPHE. Personal Interests: Complex systems modeling and simulation of brain development.