Spatial image polynomial decomposition with application to video classification

Redouane El Moubtahij, Bertrand Augereau, Hamid Tairi, Christine Fernandez-Maloigne

To cite this version:

Redouane El Moubtahij, Bertrand Augereau, Hamid Tairi, Christine Fernandez-Maloigne. Spatial image polynomial decomposition with application to video classification. Journal of Electronic Imaging, 2015, Special Section on Quality Control by Artificial Vision: Nonconventional Imaging Systems, 24 (Issue 6), 10.1117/1.JEI.24.6.061114 . hal-01228892

HAL Id: hal-01228892
https://hal.science/hal-01228892

Submitted on 17 Nov 2015

A spatial image polynomial decomposition with application to video classification

R. EL MOUBTAHIJ[a,b], B. AUGEREAU[a], H. TAIRI[b] and C. FERNANDEZ-MALOIGNE[a]
[a]XLIM Laboratory, UMR CNRS 7252 Bd Marie et Pierre Curie, 86962 Chasseneuil France
[b]LIIAN laboratory - Faculty of Science Dhar EL Mahraz USMBA, 30003 Fez Morocco

Abstract. This paper addresses the use of orthogonal Polynomial Basis transform in video classification due to its multiple advantages, especially for multiscale and multiresolution analysis similar to the wavelet transform. In our approach, we benefit from these advantages to reduce the resolution of the video by using a multiscale/multiresolution decomposition, to define a new algorithm which decomposes a color image into geometry and texture component by projecting the image on a bivariate polynomial basis and considering the geometry component as the partial reconstruction and the texture component as the remaining part, and finally to model the features (like motion and texture) extracted from reduced image sequences by projecting them into a bivariate polynomial basis in order to construct a hybrid polynomial motion texture video descriptor. To evaluate our approach, we consider two visual recognition tasks, namely the classification of dynamic textures and recognition of human actions. The experimental section shows that the proposed approach achieves a perfect recognition rate in the Weizmann database and highest accuracy in the Dyntex++ database compared to existing methods.

Keywords: Texture extraction, polynomial decomposition,dynamic texture, video classification.

Address all correspondence to: Redouane EL MOUBTAHIJ, University of Poitiers, XLIM Signal Image and Communication Department, Bvd. Marie Et Pierre Curie BP 30179, 86962 France; Tel: +33 5 49 49 74 85;
E-mail: redouane.el.moubtahij@univ-poitiers.fr

1 Introduction

Video actions classification is one of the topics of current research in computer vision and pattern recognition. They can be found in many applications such as human action recognition,[1] anomaly detection,[2] face recognition[3] and dynamic texture.[4] Its ultimate objective is to develop a video model that is efficient, fast and simple to implement against complex motions and an amazing increase in database sizes. To achieve these objectives, we propose a new fast video classification tool that models the motion and texture of the video block after reducing image sizes by using polynomial basis transformations. Since this technique offers a compact and hierarchical representation of images, and offers several advantages like multiscale and multiresolution decomposition, it has become among the best known methods in the analysis and modeling of videos compared with histograms of the orientation of gradient HOG,[5] histograms of the orientations of optical flow HOF[6] or spatio-temporal local binary patterns LBP-TOP.[7]

The remainder of this paper is organized as follows: Section 2, discusses the most recent and popular video classification descriptor in the literature based on modeling of motion and texture features to describe the action in video; Section 3, presents our algorithm based on the polynomial transform and used in a video classification, firstly presenting the Polynomial Transform, then the decomposition step. After that, the use of our image decomposition to construct a video descriptor for human action recognition and dynamic texture classification; and finally Section 4, describes and discusses the proposed approach compared to existing methods in the state of the art.

2 Related Work

The main idea of this paper is to fuse the motion features and the color texture feature after extracting from video database to perform action classification (like human action and dynamical texture), so we review representative works based on modeling of these two different features respectively.

2.1 Motion Modeling Methods

Efros et al[8] proposed a descriptor based on blurred measures optical flow to recognize the actions of the players at a football and tennis match : firstly, Lucas and Kanade optical flow algorithm[9] is used to extract the motion information between two consecutive frames. Then, the half-wave rectification technique is adopted to decompose the vector field U and V to four components. The components are then blurred with a Gaussian smoothing. The vector image with four components forms the motion descriptor. The descriptors obtained in this way are compared by a normalized correlation. Using this measure, a motion similarity matrix is constructed and the classification is performed using a k-nearest neighbour classifier. This method is extended by Danafar and al[10] by using a shape estimation. They used Harris detector to find the area where the silhouette is located. The background and silhouette are separated by using K-means algorithm. The silhouette is then divided into three horizontal parts: head and shoulders, body and arm then legs. The local optical flow for each part of the silhouette is modeled by the method proposed by Efros et al[8] to recognize simple behaviors. Mikolajczyk and Uemura[11] use optical flow information to form a vocabulary forest of local motion-appearance features to recognize actions.

Recently, some frequency transformation approaches,[12] wavelet-based approaches,[13,14] and polynomial transformation approaches[15,16] have been shown to offer good recognition accuracy. For example, in[13] wavelets are used for proposing local descriptors utilizing the capability in compacting and discriminating data, whereas in[14] wavelet processing techniques are applied to solve the problem of real time processing as well as to filter the original signal in order to achieve better classification. For polynomial-based approaches, the descriptors are based on using the orthogonal polynomial basis to describe the motions, and it has been applied very recently for action modeling. The mathematical background to generate a Bivariate Polynomial Basis as following :

First, a *Bivariate Polynomial (BP)* of degree d is a function of $x = (x_1, x_2) \in \mathbb{R}^2$ given by

$$P(x) = \sum_{\substack{(d_1,d_2)\in[0;d]^2 \\ d_1+d_2\leq d}} a_{d_1,d_2}\, x_1^{d_1}\, x_2^{d_2} \tag{1}$$

with any $a_{d_1,d_2} \in \mathbb{R}$.

Considering a finite set of pairs $D = \{(d_1, d_2)\} \subset \mathbb{N}^2$, we represent by $\mathbb{E}_D$ the space of all BP such as $a_{d_1,d_2} \equiv 0$ if $(d_1, d_2) \notin D$ and by $\mathcal{K}_D$ the subset of monomials

$$\mathcal{K}_D = \left\{ K_{d_1,d_2}(x) = x_1^{d_1}\, x_2^{d_2} \right\}_{(d_1,d_2)\in D} \tag{2}$$

Obviously $\mathcal{K}_D$ satisfies the linear independence and spanning conditions and so, $\mathcal{K}_D$ is a basis of $\mathbb{E}_D$, the canonical basis. In our context of color image decomposition, we look for bases with more suitable properties such as orthogonality or normality. So, to construct a *discrete orthonormal BP finite basis* we first have to consider the underlying discrete domain

$$\Omega = \left\{ x_{(u,v)} = \left(x_{1,(u,v)}, x_{2,(u,v)} \right) \right\}_{(u,v)\in D} \tag{3}$$

where D will now represent the set of pairs associated to Ω. Starting from $\mathcal{K}_D$ we intend to construct a new orthonormal basis applying the Gram-Schmidt process. That implies that we need some product and norm for functions defined on Ω. Given two bivariate functions, F and G, their *discrete extended scalar product* is defined by

$$\langle F|G\rangle = \sum_{(u,v)\in D} \omega(x_{(u,v)})\, F(x_{(u,v)})\, G(x_{(u,v)}) \tag{4}$$

with ω a real positive function over Ω [Legendre, Chebichev, Hermite, ...]. Then, the actual construction process of an orthonormal basis

$$\mathcal{B}_{D,\omega} = \{B_{d_1,d_2}\}_{(d_1,d_2)\in D} \tag{5}$$

is a recurrence upon (d_1, d_2)

$$T_{d_1,d_2}(x) = K_{d_1,d_2}(x) - \sum_{(l_1,l_2) \prec_2 (d_1,d_2)} \langle K_{d_1,d_2} | B_{l_1,l_2} \rangle_\omega B_{l_1,l_2}(x) \tag{6}$$

$$B_{d_1,d_2}(x) = \frac{T_{d_1,d_2}(x)}{|T_{d_1,d_2}|_\omega} \tag{7}$$

where $\prec_2$ is the lexicographical order and $|\;|_\omega$ the norm induced by $\langle\;|\;\rangle_\omega$. The resulting set of B polynomials verifies

$$\langle B_{d_1,d_2} | B_{l_1,l_2} \rangle_\omega = \begin{cases} 0 & \text{if } (d_1, d_2) \neq (l_1, l_2) \\ 1 & \text{if } (d_1, d_2) = (l_1, l_2) \end{cases} \tag{8}$$

and so $\mathcal{B}_{D,\omega}$ is effectively an orthonormal basis with respect to a weighting function ω. A special case, later used in this paper, is the *complete base* where D exactly represents the set of pairs associated to Ω, that is

$$D = [0; N_1] \times [0; N_2] \tag{9}$$

The space $\mathbb{E}_D$ being dense in the space of functions over Ω, it allows to well approximate any bivariate function I by an appropriate combination of elements of a $\mathcal{B}_{D,\omega}$ orthonormal basis

$$P_I(x) = \sum_{\{(d_1,d_2)\} \subset D} b_{d_1,d_2} \, B_{d_1,d_2}(x) \tag{10}$$

where b_{d_1,d_2} is the scalar resulting of the projection $b_{d_1,d_2} = \langle I | B_{d_1,d_2} \rangle_\omega$. In fact, with a complete orthonormal basis, the polynomial approximation of I is a first order osculatory polynomial interpolation : for all points of the domain we have $P_I(x) = I(x)$. An other nice property is that the projection on polynomial B_{d_1,d_2} can be considered as an approximation of the partial derivation $\partial_{d_1}\partial_{d_2}$. Finally and in practice, the discrete projection process supposes that both I and B_{d_1,d_2} can be evaluated on the common domain Ω. So, the set of collocation points can be obtain by uniform or non-uniform discretization of given intervals. For example, with $[-1; 1]^2$ and referring to equation (9), the collocation points obtained by uniform discretization are

$$x_{1,(u,v)} = -1 + \frac{2u}{N_1} \qquad x_{2,(u,v)} = -1 + \frac{2v}{N_2} \tag{11}$$

Based on this background, Druon et al[15] in 2009 are used the orthogonal polynomial basis for human action recognition. They introduced a polynomial modeling method based on projection into a bivariate polynomial basis and used the polynomial coefficients to analyze the displacement fields from experimental fluid mechanics. Encouraged by their positive results, O. Khil et al[16] in 2013 are based on this formalism to create a spatio-temporal descriptor for human action recognition. Their descriptor is based on two polynomial transformations: Spatial transformation is defined by projecting of displacement fields on a bivariate orthogonal polynomial basis and the temporal transformation is defined by projecting the temporal evolution of spatial polynomial coefficients on the univariate polynomial dimension basis.

2.2 Texture Modeling Methods

For a dynamic texture (DT) classification, several studies were carried out in the literature in various applications like video indexing,[17] spatial-temporal segmentation,[18] synthesis videos,[19] etc.

They can be defined as a varying spatio-temporal phenomenon and having a spatial and temporal repeatability. Among the recent techniques developed for DT recognition, Xu et al[20] have proposed a model based on a dynamic Fractal analysis to model DT with a multi-slice volumetric and dynamic Fractal spectrum. In,[21] Zaho et al. proposed a method for dynamic texture recognition. Their approach is based on modeling of texture with the LBP[22] in space-time volume. LBP method is also used in[23] to solve the unreliable information problem of LBP features by applying a Principal Histogram Analysis and a super histogram of each of all LBP histograms patch.

In more recent work, few authors were able to develop a descriptor that can classify the global video actions whatsoever human actions or dynamic textures. In[24] Ehsan et al used a directional spatio-temporal oriented gradients over nine symmetric planes for the classification of dynamic textures, human gestures and human actions. They proved the relevance of their descriptor and compared their method with the majority of descriptors existing in the action classification field. Another method was based on non-linear Stationary Subspace Analysis to separate a stationary part of the video from its non-stationary parts as proposed by Mahsa et al in.[25] Their technique is used especially for signals and has adapted to the video action classification.[25]

2.3 Texture Extraction Models

Yves Meyer[26] has proposed a model of image decomposition using the algorithm of Rudin-Osher-Fatemi.[27] According to this model, an image is split in two parts, one containing the structure u, the other one containing the texture u. The result is provided by the minimization of the functional

$$\mathcal{F}(u, v) = \|f\|_F + \lambda \|g\|_G \tag{12}$$

where $f \in F$, $g \in G$ and λ is the parameter of the model. More precisely, F is the space of functions with bounded variations and G the space of oscillating functions with the property that more a function is oscillating, more its standard norm $||g||_G$ will be low. This model can be solved numerically due to the formulation proposed in J-F.Aujol,[28,29] by the introduction of an additional parameter μ corresponding to the maximum norm of textures in the space G. The use of non-linear projectors defined by A.Chambolle[30] provides the decomposition of the image by an iterative algorithm (see[28,29] for more details).

A. Buades[31] has created a method that, as we know, is the fastest and most efficient implementation of the theory given Yves Meyer.[26] It is a fast approximate solution to the original variational problem obtained by applying non-linear filtering to the image. For each image pixel, a decision is made whether it belongs to the geometric part or to the texture part. This decision is made by computing a local total variation of the image around the point, and comparing it to the local total variation after a low pass filter has been applied. In fact, edge points in an image tend to have a slowly varying local total variation when the image is convoluted by a low pass filter while textural points instead show a strong decay of their local total variation. After the selection of the points belonging to the geometrical part, the texture part is considered as the difference between the original image and the geometrical part. (See Figure 1 for image decomposition with Buades[31] method). In fact, there is no unique decomposition and the algorithm relies on an important parameter, the scale parameter which is directly related to the granularity of textures distinguished.

3 Action Recognition Method

In this section, we describe our algorithm based on the polynomial transform and used for classification of videos. Firstly, we define the multiscale/multiresolution decomposition used to reduce

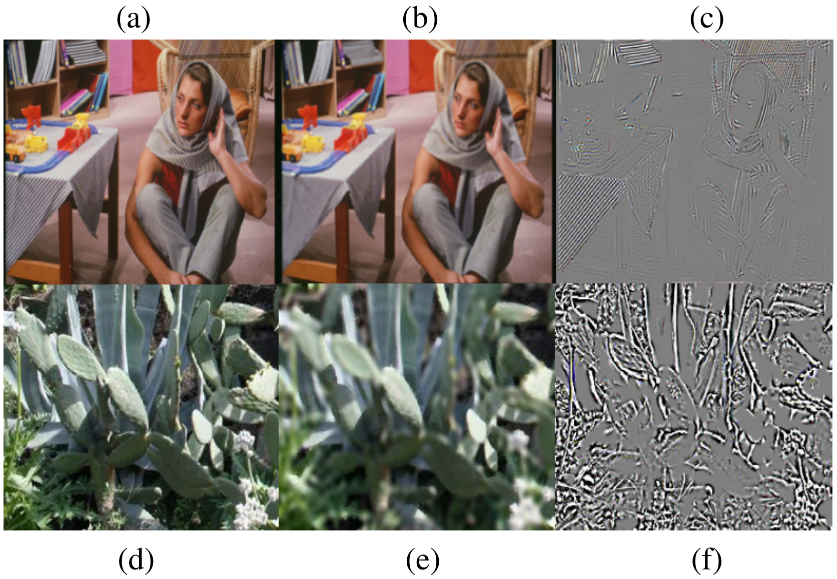


Fig 1 A. Buades method with scale parameter set to 3 : (a and d) original images, (b and e) images of geometry, (c and f) images of texture.

video resolution. Then, we introduce our polynomial image decomposition method[32] to decompose a color image into geometry and texture component by projecting the image on a bivariate polynomial basis and considering the geometry component as the partial reconstruction and the texture component as the remaining part. Finally, we define how to model the features extracted from the reduced image to construct a hybrid polynomial motion texture video descriptor.

3.1 Polynomial Transform

Now we describe the Polynomial Transform algorithm which is founded on piecewise discrete polynomial approximation and the principle of Wavelet Packet. At a given level of this multiresolution transform, lets consider a function U defined on a domain Ω of size $n_1 \times n_2$, and a basis $\mathcal{B}_{M,\omega}$ defined on a support M of size $h_1 \times h_2$, the transform process is defined as follows :

1. definition of a covering set of the discrete domain Ω with sub-domains Ω_M of size $h_1 \times h_2$

2. for each sub-domain Ω_M, projection of the corresponding restriction U_M in the basis $\mathcal{B}_{M,\omega}$

that provides the coefficients $b_{M,d_1,d_2} = \langle U_M | B_{d_1,d_2} \rangle_\omega$

3. for all pair (d_1, d_2) the reordering of the global set of coefficients b_{M,d_1,d_2} into $h_1 \times h_2$ new functions U_{d_1,d_2} defined on domains of size $n_1|_{h_1} \times n_2|_{h_2}$

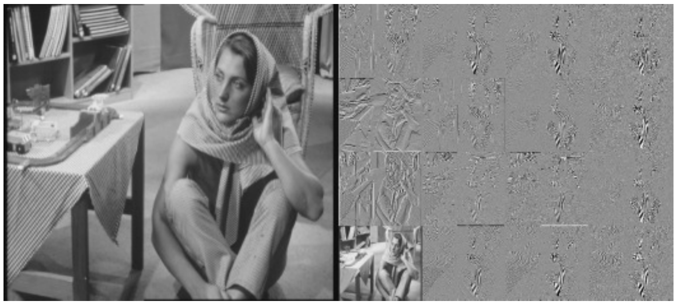

(a) (b)

Fig 2 First level of Polynomial Transform : (a) original image and (b) 4×4 Hermite complete basis transform

This method provides some flexibility especially in the choice of the resolution factors, depending of the sub-domains size and of their offset, the transform can be perform with juxtaposed or overlapped sub-domains. Moreover, the choice of the weighting function ω allows, at the same time, to perform a multi-scale and a multiresolution transformation. As an illustration, Figure 2 shows an example of a first level transformation using a 4×4 Hermite complete basis.

3.2 Image Decomposition by Partial Reconstruction

A Polynomial Transform performed with a complete basis is perfectly reversible. However, it is possible to obtain many kind of approximations by selecting the coefficients during the reconstruction phase, i.e. *partial reconstruction.* This choice of coefficients may follow various strategies, among them we have : (a) brutal restriction to a given subset, for example the polynomials of degree less than a threshold; (b) restriction based on energies, for example by using the normality of the basis to assimilate the absolute value of its coefficients to a part of the energy of a sub-domain, then sort the coefficients and finally retain a fixed number of these coefficients or those satisfying a certain condition (cf. Principle Component Analysis (PCA)).

In our case, to decompose the image into geometric and texture component, we assume that the geometrical part is given by a partial reconstruction $\tilde{I}$ of the original image I in an overlapped Polynomial Transform context. As seen before, this transform is very flexible, so there are many conceivable solutions. In order to get a compromise between quality and computation time, we choose to use an Hermite basis, to set the sub-domain size to 3×3, with an offset of 2, and to select the three dominant coefficients for each sub-domain. The partial reconstruction of a color image $I = (I_j)_{j=1\cdots 3}$, i.e. the construction of the geometrical part, can then be summarize by

$$\tilde{I}_j(x) = \frac{1}{c(x)} \sum_{\{\Omega_M \ni x\}} \left(\Psi(\Omega_M)\, \omega(x_M) \sum_{(d_1,d_2) \in \mathcal{P}_{j,M}} b_{j,d_1,d_2}(I_{j,M})\, B_{d_1,d_2}(x_M) \right) \tag{13}$$

where x is a point referring to the global image domain Ω, x_M is the same point referring to a given sub-domain Ω_M, $I_{j,M}$ the restriction of I_j to sub-domain Ω_M, $\mathcal{P}_{j,M}$ the set of selected polynomials for $I_{j,M}$ approximation, ω the weighting function of the scalar product and b_{j,d_1,d_2} the coefficient of the projection of $I_{j,M}$ on the basis polynomial B_{d_1,d_2}. A degree of anisotropy $\Psi(\Omega_M)$ is assigned to each sub-domain Ω_M and $c(x)$ is the sum of x contributions, $c(x) = \sum_{\{\Omega_M \ni x\}} \Psi(\Omega_M)\, \omega(x_M)$. The degree of anisotropy is evaluated according to

$$\Psi(\Omega_M) = \frac{1}{1 + \lambda^r} \tag{14}$$

where λ is the largest eigenvalue of a color structure tensor composed with the approximations of partial derivatives, projections on the basis polynomials of degree one

$$\mathcal{S} = \begin{pmatrix} \sum_j (b_{j,1,0})^2 & \sum_j b_{j,1,0}\, b_{j,0,1} \\ \sum_j b_{j,0,1}\, b_{j,1,0} & \sum_j (b_{j,0,1})^2 \end{pmatrix} \tag{15}$$

The balance between isotropic and anisotropic reconstruction is adjust by the parameter r that controls the degree of anisotropy in a range of 0.25 for isotropic (gaussian) to 2 for highly isotropic. By doing that, we assure a real color process and avoid marginal treatment deficiencies. Finally, the texture component I^T is simply deduce from the partial reconstruction by considering that it is the residual part of the image

$$I^T = I - \tilde{I} \tag{16}$$

The results of image decomposition into geometry and texture components by partial reconstruction after an Hermite Polynomial Transform as defined in equation(13), with the parameter r of equation (14) set to 0.75, are shown in Figure 3.

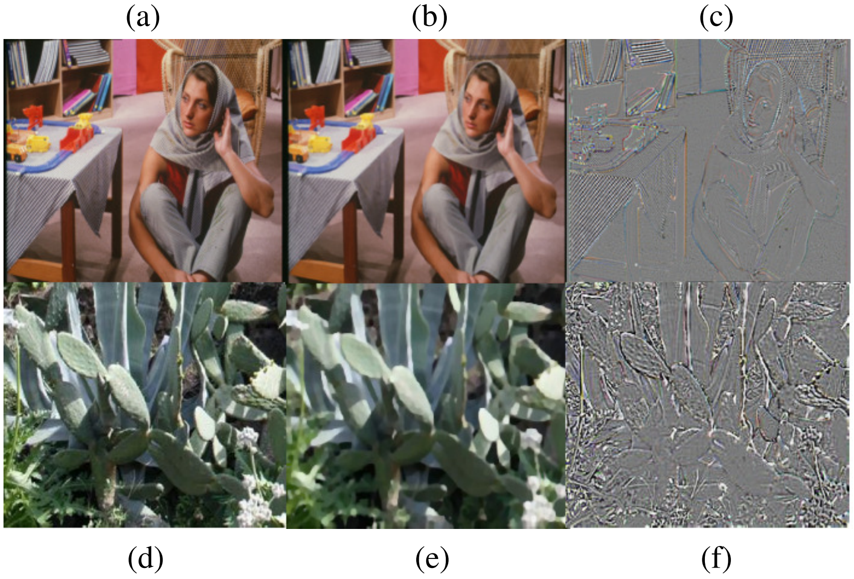


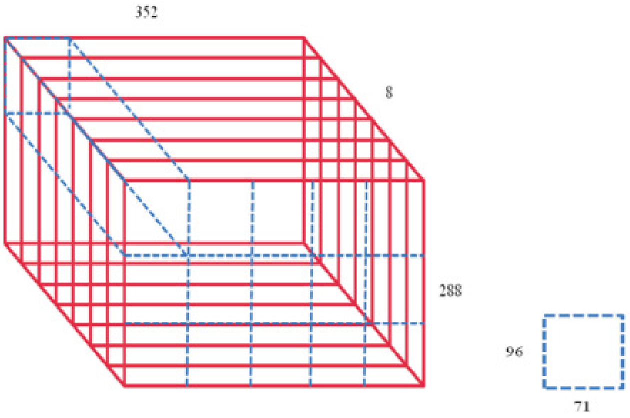
Fig 3 Our decomposition method with $r = 0.75$: (a and d) original images, (b and e) images of geometry, (c and f) images of texture.

3.3 Construction of Video Descriptor

Our video descriptor is based on the modeling of motion and texture information that define the dynamic of the objects from video scene. The motion feature has been chosen because it defines a very important source of visual information. It provides information about the three-dimensional structure of the scene, the trajectory of objects and the activity occurring in the scene. Estimation of all this information requires a precise and relevant measurement of motion in image sequences. To do this, we start by extracting the motion, notably optical flow, with a method based on color structure tensor,[33] that describes the average local information of orientations and preserves the structure of the motion. A structure tensor corresponds to the product of the gradient vector by its transposed vector. Also, if one is located in a three-dimensional space, the corresponding structure tensor is a real symmetric matrix 3×3. The color structure tensor can be constructed by a global displacement tensor from the individual component tensors[34](e.g. displacement tensor for each color component). This method allows to evaluate the co-variance between successive frames by describing the spatio-temporal structure of a given neighborhood. In addition, it provides information relative to the local velocity.

Depending on the good results obtained from the use of Polynomial transform for image decomposition, we intended to develop a classifier for dynamic texture relying on the modelization of texture and motion. To achieve this, we introduced a new algorithm based almost entirely on the projections into polynomial basis in all its stages and summarized as : pretreatment step, primitives extraction step and primitives modeling step.

For a pretreatment step, we have solved the problem of using video database of large sizes thanks to a multiresolution/multi-scale decomposition as defined in the subsection (3.1) extended



(a) (b)

Fig 4 A three-dimensional multiresolution decomposition : (a) A 3D volume with size $288 \times 352 \times 8$, (b) Reduced image after a multiresolution decomposition by projecting (a) into a three-dimensional Hermite basis with degree $d_{\mathcal{B}} = 1$ and support $3 \times 5 \times 8$ (a patch of color blue).

in 3 dimensions. That which allowed us to reduce the video resolution and therefore reduce the computing time to build the descriptive vector so it can be used within a system of dynamic texture recognition. According to these transformations, we obtain an image representation with different scales as shown in Figure 2 for the two dimensional case. In our case, we arrived to represent a video block of $288 \times 352 \times 120$ to images sequences fairly small of size 71×96. This can be achieved by projecting the video block into a trivariate polynomial basis with patch $5 \times 3 \times 8$ as shown in Figure 4, which is why the concept of reducing the image size has been introduced.

For the primitive extraction step, we applied the image decomposition method defined in the subsection (3.2) on the reduced image sequences to extract the texture primitive. The motion is extracted directly on the original reduced image sequences by applying the color structure tensor method. Finally, the extracted texture component as well as the motion fields (u, v) would be projected thereafter on a two-dimensional Hermite basis and concatenated into descriptive vector for each video sequence as shown in Figure 5.

4 Experimental Results

To evaluate the effectiveness and robustness of our approach, we consider two difficult visual tasks of action recognition, namely the classification of dynamic textures and recognition of human actions represented respectively by two databases Dyntex++[35] and Weizmann[36] (see Figure 6). A support vector machine with Libsvm library and a radial basis function (RBF) kernel is trained for classification.[37]

4.1 Human Action Recognition

In order to classify the human actions, we use the Weizmann database outcome of the work of Gorelick and al,[36] which is constituted of videos of a single person, consisting of 10 classes of human actions and performed by 9 different persons. This database is widely used by many authors for the validation of their results. We have chosen this database because it is relatively simple to study as the background of the video is relatively constant on all the videos. In addition, the actions are carried with a static camera (e.g. there is no camera motion or change of illumination).

There are two basic versions of Weizmann databases in the literature: A standard database of basic $10 - classes$ and a smaller version that does not include skipping class (Weizmann database with 9 classes). In our approach, we use these two databases and we apply the procedure shown

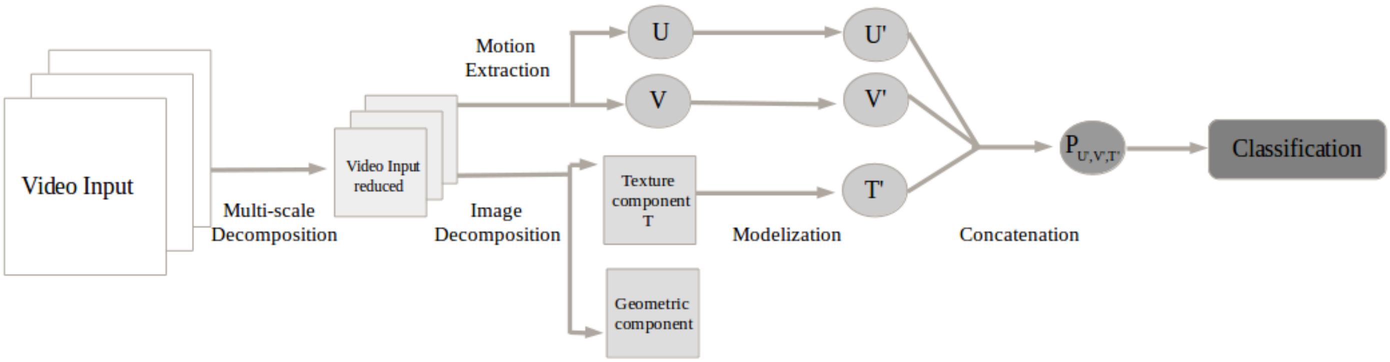


Fig 5 Spatial Polynomial Motion texture descriptor(SPMT)

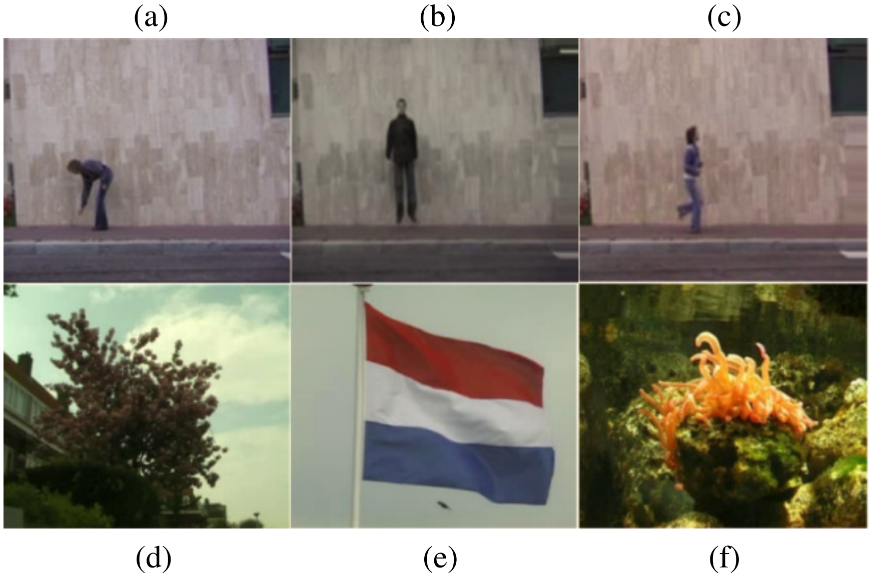


Fig 6 Example of video action databases : (a) bend , (b) jump and (c) run for Weizmann databases. (d) Blossoming tree in the wind, (e) flag and (f) underwater life Dyntex++ databases.

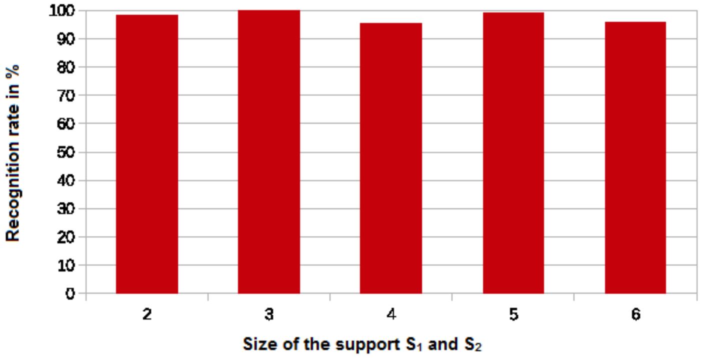


Fig 7 The recognition rate versus the support size of 3D Hermite basis for a multiresolution step in Weizmann dataset.

in Figure 5 for each video sequence. Similar to,[38] all experience is performed using leave-one-out strategy. Figure 7 shows the average recognition rate relative to the size of the patch S_1, S_2 and S_3 for multiresolution decomposition step (where $S_1 = S_2$ = $S_3 = 3$ in this case). The results of the classification are shown in Table 1.

Since we need polynomials coefficients $b_{0,0,0}$, one can use a trivariate basis of degree $d_{\mathcal{B}} = 1$ in order to reduce the resolution size of the images and finally to have the most minimal execution time in multiresolution step. In order to get a compromise between quality and computation time

Table 1 Classification results of the Weizmann database

Class number	Method	Accuracy (%)
10 classes	Fusion (Gabor filter + LBP + HOG)[38]	$97, 8$
	LBP-Top (Local Binary Patterns on Three Orthogonal Planes)[21]	$95, 6$
	SIFT-3D (Scale Invariant Feature Transform)[39]	$82, 6$
	HOG-3D (Histogram of Oriented Gradients 3D)[5]	$84, 3$
	OF+CTM (Optical Flow and Correlated Topic Model)[40]	89.20
	HOG-NSP (Histogram of Oriented Gradients with Nine Symmetry Planes)[24]	$95, 9$
	Our approach	100
9 classes	LBP-Top[21]	$98, 7$
	Fusion[38]	100
	Our approach	100

in the texture extracting step, we choose to use a Hermite basis, to set the sub-domain size to 5×5, with an offset of 2. Subsequently, we model the U and V of motion and T of texture by projecting them into a two-dimensional basis with the following settings:

- $S_1 = (\text{height of image})/3$ and $S_2 = (\text{width of image})/3$.
- $d_{\mathcal{B}_{U,V}} = 6$ and $d_{\mathcal{B}_T} = 4$.

where $d_{\mathcal{B}_{U,V}}$ and $d_{\mathcal{B}_T}$ are the degree of the two-dimensional basis to model U ,V and T respectively.

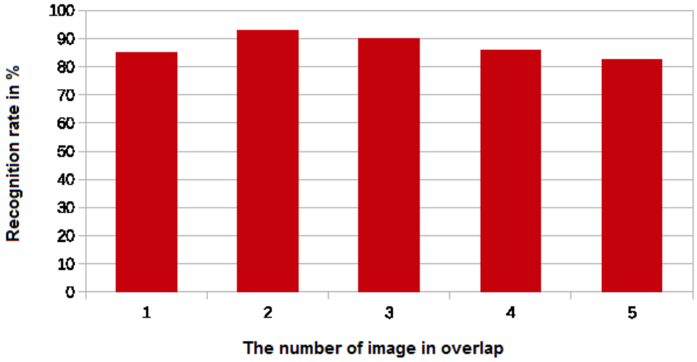


Fig 8 The recognition rate versus the image overlap in multiresolution decomposition step applied to Dyntex++ database.

4.2 Recognition of Dynamic Textures

To illustrate the efficiency of our method, we will use it in a dynamic texture classification scheme which is among challenging research themes in the image analysis field. The database used in our application is Dyntex ++ database[35] which is an improved version of the Dyntex database.[41] Because it is composed of images that represent the textures in motion, it seems appropriate for the present study. Since the videos are not evenly distributed among the 36 classes in this database, we used the same experimental setting as that in[42] in the evaluation (e.g without converting the video sequences in grey level) to have the same sequence number of each class. We finally have 3600 dynamic textures, grouped into 36 classes and each sequence have a size of $50 \times 50 \times 50$ pixels. Thereafter, the data were randomly split into two equal size training and test sets. The random split was repeated 10 times and the average classification accuracy is reported in Table (2).

In the step of multiresolution decomposition, we project each sequence on a trivariate polynomial basis of degree $d_{\mathcal{B}} = 1$ with a patch of size $3 \times 3 \times 8$ with overlapping temporal of 2 images to reduce the images sizes 16×16 instead of 50×50. Then, in the texture extracting step, the best results are obtained by projecting the reduced images into a Hermite bivariate basis of degree $d_{\mathcal{B}} = 2$ and a patch $S_1 = 3$, $S_2 = 3$. Finally, the features extracted such as motion (U and V) and texture (T) are modeled by projecting them into a Hermite bivariate basis with the following parameters:

- $S_1 =$ height of image and $S_2 =$ width of image.
- $d_{\mathcal{B}_{U,V}} = 2$ and $d_{\mathcal{B}_T} = 2$.

The average recognition rate relative to overlap image number and the size of the patch for multiresolution step are shown respectively in Figure 8 and Figure 9.

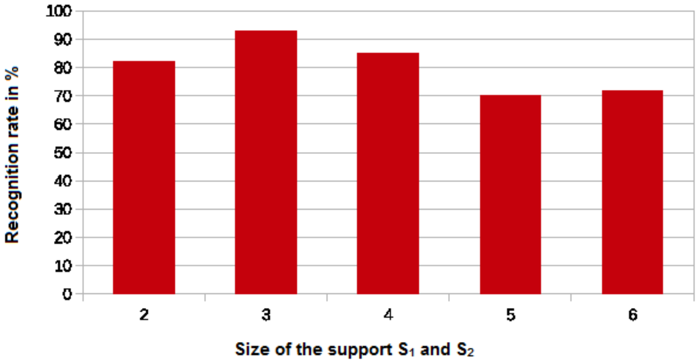


Fig 9 The recognition rate versus the support size of 3D Hermite basis for multiresolution step in Dyntex++ database.

Table 2 Classification results for Dyntex++ database.

VLBP (Volume Local Binary Patterns)[43]	61.1%
SIFT-3D (3 Dimensional Scale Invariant Feature Transform)[39]	63.7%
LBP-TOP(Local Binary Patterns on Three Orthogonal Planes)[21,24]	71.2%
DFS (Dynamic Fractal Spectrum)[20]	$89, 9\%$
HOG-NSP (Histogram of Oriented Gradients with Nine Symmetry Planes)[24]	$90, 1\%$
NLSSA (Non-Linear Stationary Subspace Analysis)[25]	92.4%
Our method[32]	$93, 13\%$

From Table 1, we can see that our descriptor is much better than the methods very well known in the field of recognition of human action. We can also see in Table 2 that the descriptor is robustness with the database Dyntex++ where the recognition rate can be ranked among the best rates found in recent methods dealing with the classification of dynamic textures. As a final evaluation, we compare the computational time of our motion texture descriptor against other state-of-the-art video descriptors such as HOG-NSP,[24] 3D-SIFT[39] and LBP-TOP[21,23] on a set of videos with resolution 160×120. For our descriptor, experiments were developed in C++/Matlab environment and performed on an Intel 2.40GHz, whereas for all other descriptors, their experiences are developed and mentioned in.[23] Table 3 presents the average run-time for all descriptors. The proposed descriptor is more than $5, 5s$ times faster than LBP-TOP, HOG-NSP, 3D-SIFT and Morphological

Table 3 Average run-time (sec) for computing of our descriptor, LBP-TOP[21,23] , HOG-NSP[24] and 3D-SIFT[39] descriptors on a video with 160×120.

Descriptor	Our Descriptor	LBP-TOP	HOG-NSP	3D-SIFT	MCA
Run-time (s)	$5,5$	$7,4$	$8,6$	$74,7$	7200

Component Analysis MCA[44] because all main computations can be realized through convolutions. Besides that, the computational time with MCA is around 2 hours, it serves to decompose an video of the Dyntex database[41] into geometric and texture component with size 648×540 on a classical computer, which is equal to $3,65s$ if we use our polynomial decomposition method.

5 Conclusion

In this paper, we have proposed a new approach for texture extraction from color image sequence, by using Polynomial Transformations. Partial reconstruction and global approximation are used to build descriptors used in a classification process of dynamic textures. In addition to the simplicity of implementation, we provide a computing time which is especially fast compared to most of methods based on the theory of Yves Meyer due to the cost of the minimization of the total variation. The experimental results show that the proposed approach achieves a very good recognition rate for the Dyntex++ database. This shows the relevance of our texture extraction method in the context of classification of dynamic textures. In some future, we will continue to improve our image decomposition method in order to extract the noise coefficients ignored in the partial reconstruction of the image. We will also investigate the abilities of a derived method which only relies on three dimensional transformations in our classification process of dynamic textures.

References

1 D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Computer Vision and Image Understanding* **115**(2),

224 – 241 (2011).

2 M. Bertini, A. D. Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding* **116**(3), 320 – 329 (2012). Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

3 C.-Y. Lu, H. Min, J. Gui, L. Zhu, and Y.-K. Lei, "Face recognition via weighted sparse representation," *Journal of Visual Communication and Image Representation* **24**(2), 111 – 116 (2013). Sparse Representations for Image and Video Analysis.

4 M. Koleini, M. R. Ahmadzadeh, and S. Sadri, "A new efficient method to characterize dynamic textures based on a two-phase texture and dynamism analysis," *Pattern Recognition Letters* **45**(0), 217 – 225 (2014).

5 A. Klser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *In BMVC08*,

6 L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, **2**, II–123–II–130 vol.2 (2001).

7 V. Kellokumpu, G. Zhao, and M. Pietikinen, "Human activity recognition using a dynamic texture based method," in *In BMVC*, (2008).

8 A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 726–733 vol.2 (2003).

9 B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelli-*

gence - Volume 2, *IJCAI'81*, 674–679, Morgan Kaufmann Publishers Inc., (San Francisco, CA, USA) (1981).

10 S. Danafar and N. Gheissari, "Action recognition for surveillance applications using optic flow and svm," in *Computer Vision ACCV 2007*, Y. Yagi, S. Kang, I. Kweon, and H. Zha, Eds., *Lecture Notes in Computer Science* **4844**, 457–466, Springer Berlin Heidelberg (2007).

11 K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8 (2008).

12 H. Imtiaz, U. Mahbub, and M. Ahad, "Action recognition algorithm based on optical flow and ransac in frequency domain," in *SICE Annual Conference (SICE), 2011 Proceedings of*, 1627–1631 (2011).

13 L. Shao and R. Gao, "A wavelet based local descriptor for human action recognition," in *Proceedings of the British Machine Vision Conference*, 72.1–72.10, BMVA Press (2010). doi:10.5244/C.24.72.

14 L. Palafox and H. Hashimoto, "Human action recognition using wavelet signal analysis as an input in 4w1h," in *Industrial Informatics (INDIN), 2010 8th IEEE International Conference on*, 679–684 (2010).

15 M. Druon, *Modélisation du mouvement par polynômes orthogonaux : application à l'étude d'écoulements fluides*. These, Université de Poitiers (2009).

16 O. Kihl, D. Picard, and P.-H. Gosselin, "Local polynomial space-time descriptors for actions classification," in *International Conference on Machine Vision Applications*, –, (Kyoto, Japon) (2013).

17 S. Dubois, R. Péteri, and M. Michel, “Indexation de Textures Dynamiques à l’aide de Décompositions Multi-échelles,” in *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, 978–2–9539515–2–3, (Lyon, France) (2012). Session ”Articles”.

18 J. Chen, G. Zhao, M. Salo, E. Rahtu, and M. Pietikainen, “Automatic dynamic texture segmentation using local descriptors and optical flow,” *Image Processing, IEEE Transactions on* **22**, 326–339 (2013).

19 G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, “Dynamic textures,” *International Journal of Computer Vision* **51**(2), 91–109 (2003).

20 Y. Xu, Y. Quan, H. Ling, and H. Ji, “Dynamic texture classification using dynamic fractal analysis,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1219–1226 (2011).

21 G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **29**, 915–928 (2007).

22 K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii, “Feature extraction of temporal texture based on spatiotemporal motion trajectory,” in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, **2**, 1047–1051 vol.2 (1998).

23 J. Ren, X. Jiang, and J. Yuan, “Dynamic texture recognition using enhanced lbp features,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2400–2404 (2013).

24 E. Norouznezhad, M. Harandi, A. Bigdeli, M. Baktash, A. Postula, and B. Lovell, “Di-

rectional space-time oriented gradients for 3d visual pattern analysis," in *Computer Vision ECCV 2012*, **7574**, 736–749 (2012).

25 M. Baktashmotlagh, M. Harandi, B. Lovell, and M. Salzmann, "Discriminative non-linear stationary subspace analysis for video classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **36**, 2353–2366 (2014).

26 Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations: The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures*, American Mathematical Society, Boston, MA, USA (2001).

27 L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena* **60**(14), 259 – 268 (1992).

28 J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle, "Image decomposition: application to textured images and SAR images," Tech. Rep. RR-4704 (2003).

29 J.-F. Aujol, Gilboa, Guy, Chan, Tony, Osher, and Stanley, "Structure-texture image decompositionmodeling, algorithms, and parameter selection," *International Journal of Computer Vision* **67**(1), 111–136 (2006).

30 Chambolle and Antonin, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision* **20**(1-2), 89–97 (2004).

31 A. Buades, L. Triet, M. Jean-Michel, and V. Luminita, "Cartoon/Texture Image Decomposition ," *Image Processing On Line* **1** (2011).

32 R. El Moubtahij, B. Augereau, C. Fernandez-Maloigne, and H. Tairi, "A polynomial texture extraction with application in dynamic texture classification," *Twelfth International Confer-*

ence on Quality Control by Artificial Vision 2015 - Proc. SPIE **9534**, 953407–953407–7 (2015). [doi:10.1117/12.2182865].

33 B. Augereau, B. Tremblais, and C. Fernandez-Maloigne, "Vectorial Computation of the Optical Flow in Color Image Sequences," in *Thirteenth Color Imaging Conference*, **13**, 130–134, (Scottsdale, États-Unis) (2005).

34 J. Benois-Pineau and B. Augereau, "Motion Estimation in Color Image Sequences," in *Digital Color Imaging*, L. M. Christine Fernandez-Maloigne, Frederic Robert-Ignacio, Ed., *Digital Signal and Image Processing*, 368, Willey (2012).

35 R. Pteri, S. Fazekas, and M. J. Huiskes, "Dyntex: A comprehensive database of dynamic textures," *Pattern Recognition Letters* **31**(12), 1627 – 1632 (2010).

36 M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, **2**, 1395–1402 Vol. 2 (2005).

37 C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.* **2**, 27:1–27:27 (2011).

38 S. Brahnam and L. Nanni, "High performance set of features for human action classification.," in *IPCV*, 980–984 (2009).

39 P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07*, 357–360, ACM, (New York, NY, USA) (2007).

40 L.-m. Tu, Hong-bin. Xia and Z.-w. Wang, "The complex action recognition via the correlated topic model," *The Scientific World Journal* , 10 (2014).

41 R. Péteri, S. Fazekas, and M. J. Huiskes, “DynTex : a Comprehensive Database of Dynamic Textures,” *Pattern Recognition Letters* **doi: 10.1016/j.patrec.2010.05.009**. http://projects.cwi.nl/dyntex/.

42 B. Ghanem and N. Ahuja, “Maximum margin distance learning for dynamic texture recognition,” in *Computer Vision ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., *Lecture Notes in Computer Science* **6312**, 223–236, Springer Berlin Heidelberg (2010).

43 G. Zhao and M. Pietikinen, “Dynamic texture recognition using volume local binary patterns,” in *Dynamical Vision*, *Lecture Notes in Computer Science* **4358**, 165–177, Springer Berlin Heidelberg (2007).

44 S. Dubois, R. Péteri, and M. Ménard, “Analyse de Textures Dynamiques par décompositions spatio-temporelles: application à l’estimation du mouvement global,” in *Actes de la conférence CORESA 2010*, (Lyon, France) (2010).

Redouane EL MOUBTAHIJ received his Master Degree of Imaging, Computer graphics and Computer science from University of Sidi Mohamed Ben Abdellah (USMBA), Fez Morroco in 2011. He is currently a Ph.D student in joint between XLIM-SIC laboratory of University of Poitiers in France and LIIAN laboratory of USMBA in Fez Morocco.

Bertrand AUGEREAU lecturer and doctor of science, and actively involved in research at the University of Poitiers since 1997 in the SIC department of the XLIM, CNRS Research Institute, France. He is working on mathematical models for image and image-sequence as well as the establishment of a template of their analysis and processing application. For several years, this research has developed in the more specific direction of perceptible motion through a sequence of images.

Hamid TAIRI received his PhD degree in 2001 from the University Sidi Mohamed Ben Abdellah (USMBA) Morocco. In 2002 he has done a postdoc in the Image Processing Group of the Laboratory LE2I in France. Since 2003, he has been an associate professor at the USMBA, where he has worked as a professor of computer science. His research interests are in 3D reconstruction of artificial vision, in visual information retrieval and pattern recognition.

Christine FERNANDEZ-MALOIGNE received her PhD in 1989 from the University of Technology of Compiegne, where she was associated Professor in Image processing and she succeeded in her accreditation to supervise researches in the University of Lille, France. Then she moved to the University of Poitiers, to create a new research pole for color image processing and analysis. She is currently deputy director of XLIM laboratory, as well as director of a CNRS research federation.

Biographies and photographs of the other authors are not available.

List of Figures

List of Tables