NONLINEAR TEMPORAL MODELING FOR MOTION-BASED VIDEO OVERVIEWING

Eric Bruno and Stéphane Marchand-Maillet

Computer Vision and Multimedia Laboratory, University of Geneva
25 rue du Général Dufour
1211 Geneva 4, Switzerland
email: bruno, marchand@cui.unige.ch

ABSTRACT

This paper presents a method for video collection overviewing based the dynamic content of the scenes. In an unsupervised context, our approach relies on the nonlinear temporal modeling of wavelet-based motion features directly estimated from the image sequence. Based on SVM-regression, the nonlinear model is able to learn the behavior of the motion descriptors along the temporal dimension and to catch useful informations of the dynamic content. A similarity measure associated to the temporal model is then defined. It allows to compare video segments according to motion descriptors and thus defines a high-dimensional feature space where the video sequences under investigation are projected. The Curvilinear Component Analysis algorithm is finally used to map the feature space onto a 2D space. This operation enables us to display the video collection and gives an overview of the content according to motion features.

1. INTRODUCTION

Video databases are growing so rapidly that most of the information they contain is lost in this mass and is thus inaccessible. A valuable tool for the management of visual records would provide the ability to automatically describe and index the content of video sequences in order to create a relevant overview of the data's variety. Such a facility would allow the users to navigate and to discover the existence of video material contained in large databases.

In an unsupervised context, this issue is handled by extracting from videos[1] low level features which are related to some high level concepts, such as film genre, particular scenes or activities. The feature space spanned by those descriptors is then analyzed in order to give to the users an understandable view of the structure of the data that it represents. Amongst all the video primitives available (color, shape, texture, audio descriptors,...), motion features are well-suited to index video documents according to their dynamic content, for e.g. event characterization [13], human behavior recognition [12] or video summarization [11]. An efficient extraction of such information needs to consider both spatial and temporal properties of the motion-based descriptors. Spatio-temporal models have then to be defined in order that dynamic main features are clearly expressed and create suitable description of video documents. For example, motion parameter trajectory combined to condensation algorithm are considered in [1], temporal Gibbs model of motion-related measures are used in [6] and a 3D Gabor decompositions performs a spatio-temporal video analysis in [4].

Our approach, presented in this paper, relies on previous work which concern global motion estimation between two images using a wavelet-based parametric model [2]. This model can directly be applied over the whole image without any prior and generally unreliable segmentation stage. The estimated motion parameters then provide a robust, global, meaningful and compact description of activity content [3]. The motion descriptors are estimated between any two consecutive frames of the sequence so that the video sequence is characterized by a *sequence of descriptors*. This temporal content is captured by using a temporal model based on the estimation of the nonlinear prediction function of the sequence. This model enables us to define a similarity measure between videos based on the prediction error, which avoids facing the problem of temporal alignment. The similarity measure defines a multidimensional feature space where each video is an element of this space. A 2D representation of this feature space is obtained by using a nonlinear dimensionality reduction algorithm which preserves local topology between elements. In this way, the video database can be displayed on a 2D map where spatial relations between documents correspond to local structures in the feature space. Experiments are carried out on a set of video shots containing sport and news programs and on shots extracted from a TV movie.

The paper is organized as follow. Section 2 describes the wavelet-based motion estimation and the motion descriptors

[1] In this paper videos are considered as consistent units, equivalent to what is normally referred to as "shots"

derived from the motion wavelet coefficients. The nonlinear temporal model and similarity measure are presented in section 3. Section 4 outlines the CCA algorithm. Experiment on real videos are displayed in section 5 and conclusion in section 6 ends this paper.

2. MOTION FEATURE EXTRACTION

The motion descriptors we propose to use are based on the wavelet coefficients of the optical flow directly estimated from the image sequence. These descriptors have the ability to characterize activity according to the motion magnitude, scale and orientation [3].

2.1. Motion wavelet coefficient estimation

In this section, we briefly outline the algorithm that we have developed to estimate motion wavelet coefficients. Further details can be found in [2].

Let us consider an image sequence $I(\boldsymbol{p}_i, t)$ with $\boldsymbol{p}_i = (x_i, y_i) \in \Omega$ the location of each pixel in the image. The *brightness constancy assumption* states that the image brightness $I(\boldsymbol{p}_i, t+1)$ is a simple deformation of the image at time t

$$I(\boldsymbol{p}_i, t) = I(\boldsymbol{p}_i + \mathbf{v}(\boldsymbol{p}_i), t+1), \tag{1}$$

where $\mathbf{v}(\boldsymbol{p}_i, t) = (u, v)$ is the optical flow between $I(\boldsymbol{p_i}, t)$ and $I(\boldsymbol{p}_i, t+1)$. This velocity field can be globally modeled as a coarse-to-fine 2D wavelet series expansion from scale L to l

$$\begin{aligned}\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_i) = &\sum_{k_1,k_2=0}^{2^L-1} \boldsymbol{c}_{L,k_1,k_2}\Phi_{L,k_1,k_2}(\boldsymbol{p_i}) \\ &+ \sum_{j\geq L}^{l}\sum_{k1,k2=0}^{2^j-1}\Big[\boldsymbol{d}^H_{n,k1,k2}\Psi^H_{j,k1,k2}(\boldsymbol{p}_i) \\ &+\boldsymbol{d}^D_{n,k1,k2}\Psi^D_{j,k1,k2}(\boldsymbol{p}_i) + \boldsymbol{d}^V_{n,k1,k2}\Psi^V_{j,k1,k2}(\boldsymbol{p}_i)\Big], \end{aligned} \tag{2}$$

where $\Phi_{L,k_1,k_2}(\boldsymbol{p}_i)$ is the 2D scaling function at scale L, and $\Psi^{H,D,V}_{j,k_1,k_2}(\boldsymbol{p}_i)$ are wavelet functions which respectively represent horizontal, diagonal and vertical variations. These functions are dilated by 2^j and shifted by k_1 and k_2. The coarsest level corresponds to $L = 0$ whereas l defines the finest details that can be fitted by the motion model.

In order to recover a smooth and regular optical flow, we use *B-spline* wavelets, which have maximum regularity and symmetry. The degree of the B-spline determines the approximation accuracy.

The motion parameter vector $\boldsymbol{\theta}$, which contains wavelet coefficients $\boldsymbol{c}_{L,k_1,k_2}$ and $\boldsymbol{d}^{H,D,V}_{j,k_1,k_2}$ for all j, k_1, k_2 is estimated by minimizing an objective function

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{p}_i\in\Omega} \rho\left(I(\boldsymbol{p_i} + \mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_i), t+1) - I(\boldsymbol{p}_i, t)\right), \tag{3}$$

where $\rho(\cdot)$ is a robust norm error (M-estimator). The minimization step is achieved using an incremental and multiresolution estimation method [8].

The wavelet-based motion model enables to estimate for successive frames an accurate optical flow defined by its wavelet coefficients. The finer scale l determines how precise the final estimation is. In the context of video indexing, a fine estimation is not needed, as we only want discriminative descriptors over a wide range of contents. Figures 1.b., c. and d. display the estimated optical flows for various final scale levels. For our experiment, we have used a final scale $l = 3$ which correspond to a motion model configured by 128 wavelet coefficients.

2.2. Activity descriptors

As we can see in Figure 1, the motion parameter vector θ contains an accurate description of the optical flow. For the purpose of video overviewing, we have observed that such an accuracy is rather a shortcoming since large variabilities between descriptors may occur only because of local differences within optical flows. To overcome this problem, we consider a variance measure of the wavelet coefficients in the different subbands of the representation

$$\begin{aligned}&\boldsymbol{\sigma} = \left[\sigma_0, \sigma_1^H, \sigma_1^D, \sigma_1^V, \sigma_2^H, \ldots, \sigma_l^V\right], \\ &\text{with} \\ &\sigma_0 = c_0^2 \\ &\sigma_j^{H,D,V} = \sum_{k_1,k_2=0}^{2^j-1}\left|\boldsymbol{d}^{H,D,V}_{j,k_1,k_2}\right|^2, \forall j \in [1,l]\end{aligned} \tag{4}$$

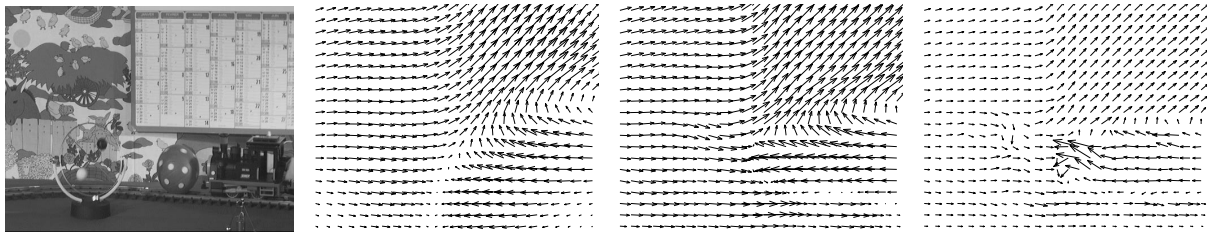
where l is the finest scale level used in (2), meaning that $\boldsymbol{\sigma}$ is a 10-component vector in our case, characterizing optical flow in term of its global magnitude, scale and orientation.

Hence, given an image sequence of N frames, the activity description consists in a sequence of $N-1$ descriptors $\boldsymbol{\sigma}$ computed over all consecutive frames.

3. SVM REGRESSION FOR NONLINEAR TEMPORAL MODELING AND SIMILARITY MEASURE DEFINITION

3.1. Feature temporal modeling as a time series prediction problem

Let S be an image sequence characterized by a set of descriptors $\{\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_N\}$, $\mathbf{X}_t \in \mathbb{R}^D$, with N the length



(a) Frame from *Mobile and Calendar* sequence

(b) Estimated flow field at $j = 2$

(c) Estimated flow field at $j = 3$

(d) Estimated flow field at $j = 4$

Fig. 1. Frame from *Mobile and Calendar* sequence and global motion estimated at level a) 2, b) 3 and c) 4. B-spline of degree 2 were used to model motion.

of the descriptor sequence. The H^{th} order prediction function $\mathbf{F} : \mathbb{R}^{D\times H} \rightarrow \mathbb{R}^D$ of the temporal series $\{\mathbf{X}_t\}_{t=0}^N$ is defined

$$\mathbf{X}_t = \mathbf{F}(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-H}) \; \forall t \in [H, N]. \qquad (5)$$

The multidimensional function $\mathbf{F}$ can be considered as a temporal model of the visual descriptors and is therefore able to characterize the dynamic content of the sequence S. The order H determines the memory of the model since $\mathbf{X}_T$ is a function of the H previous descriptors $\{\mathbf{X}_t\}_{t=T-H}^{T-1}$. The larger H is, the more the model is specific to the sequence and over-fits the dynamic content. On the other hand, the information characterized by the prediction function tends toward zeros as the model memory decreases.

The estimation the multidimensional function $\mathbf{F}$ (eq. 5) is done separately over each dimension, which implies that there is no interaction between dimensions. Let us note x^l the l^{th} component of $\mathbf{X}$, the problem therefore consists in estimating f^l such as

$$x_t^l = f^l(x_{t-1}^l, x_{t-2}^l, \ldots, x_{t-H}^l). \qquad (6)$$

Then

$$\mathbf{F} = [f^1, f^2, \ldots, f^D]^T. \qquad (7)$$

For the sake of simplicity in the notation, we define the $H-$dimensional vector

$$\mathbf{x}_t^l = [x_t^l, x_{t-1}^l, \ldots, x_{t-H}^l] \; \forall t \in [H, N], \qquad (8)$$

in a such way that equation (6) can be written as $x_t^l = f^l(\mathbf{x}_{t-1}^l)$. The main difficulty of this approach is to estimate f^l efficiently. As the descriptor sequence is nonstationary, we have to estimate a nonlinear prediction function from the set of the $N - H$ observations. Many regression techniques can used to solve this problem, but results obtained by using Support Vector Machines in regression show that this kernel-based algorithm is well-suited for such nonlinear estimation [7].

3.2. Support Vector Machines for regression

We present here a short description of SVM for regression. Further details can be found [10, 9], especially for issues related to the robustness of the algorithm. This classical problem of regression consist in approximating an unknown function $g : \mathbb{R}^D \rightarrow \mathbb{R}$ from sampled data $\{\mathbf{x}_i, y_i\}_{i=1}^N$ such as $y_i = g(\mathbf{x}_i) + \eta$, with η some noise. In order to approximate g, the SVM algorithm considers a parametrical model of the form

$$f(\mathbf{x}) = \sum_{i=1}^{L} c_i \phi_i(\mathbf{x}) + b, \qquad (9)$$

where $\{\phi_i\}_{i=1}^L$ are basis functions. Parameters b and $\{c_i\}_{i=1}^L$ are unknown parameters that have to be estimated by minimizing the functional

$$R(f) = \frac{1}{N} \sum_{i=1}^{N} |y_i - f(\mathbf{x}_i)|_\epsilon + \lambda ||\mathbf{c}||^2, \qquad (10)$$

with $\mathbf{c} = [c_1, \ldots, c_N]$ and λ a smoothness constraint applied to the solution space. The error function is defined as follow

$$|x|_\epsilon = \begin{cases} 0 & \text{if } x < \epsilon \\ x & \text{otherwise.} \end{cases} \qquad (11)$$

In [10], Vapnik has shown that the function which minimize the functional (10) has the following form

$$f(\mathbf{x}, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) K(\mathbf{x}, \mathbf{x}_i) + b \qquad (12)$$

with $\alpha_i^* \alpha_i = 0$, $\alpha_i, \alpha_i^* \geq 0 \; i = 1, \ldots, N$ and where $K(\mathbf{x}, \mathbf{y})$ is the so-called kernel function that describes the inner product in the D-dimensional feature space defined by the functions ϕ_i

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{L} \phi_i(\mathbf{x}) \phi_i(\mathbf{y}). \qquad (13)$$

The main interest of SVM technique is that only the kernel K has to be known and the feature space spanned by ϕ_i never need to be explicitly computed. This allows to uses several type of basis functions, including infinite sets, which give a wide choice of nonlinear models to approximate the unknown function.

Concerning sequences of visual descriptors, where no priors about the solution form are known, we use the radial gaussian kernel $K(\mathbf{x},\mathbf{y}) = exp(-\gamma||\mathbf{x}-\mathbf{y}||^2)$ which can fit a large range of complex functions. The scale parameter γ determines distances between observations $\{\mathbf{x}_t\}_{t=K}^{N}$ and thereby the smoothness of the solution in the observation space. We set it as follow

$$\gamma = \frac{1}{2}\left(\frac{1}{N-1}\sum_t ||\Delta_t \mathbf{x}_t||^2\right)^{-1}, \qquad (14)$$

where $\Delta_t \mathbf{x}_t = \mathbf{x}_{t-1} - \mathbf{x}_t$. This setting ensures that, on average, distance between temporal neighbors is small enough to obtain a smooth prediction function and to avoid overfitting effects.

3.3. Similarity measure as a prediction error

Let $\mathbf{F}$ and $\mathbf{G}$ be the prediction functions (or temporal model) respectively estimated on time series of descriptors $\{\mathbf{X}_t\}_{t=0}^{N}$ and $\{\mathbf{Y}_t\}_{t=0}^{M}$ related to the image sequences S_1 et S_2. From these prediction functions, we can built two new time series by crossing models and descriptors

$$\begin{aligned}\tilde{\mathbf{X}}_t &= \mathbf{G}(\mathbf{X}_{t-1},\dots\mathbf{X}_{t-H}), \forall t \in [H,N]\\ \tilde{\mathbf{Y}}_t &= \mathbf{F}(\mathbf{Y}_{t-1},\dots\mathbf{Y}_{t-H}), \forall t \in [H,M],\end{aligned} \qquad (15)$$

and then define the distance

$$D(X,Y) = \frac{1}{2}\left[d\left(\{\tilde{\mathbf{X}}_t\}_t, \{\mathbf{X}_t\}_t\right) + d\left(\{\tilde{\mathbf{Y}}_t\}_t, \{\mathbf{Y}_t\}_t\right)\right) \qquad (16)$$

with $d(\cdot,\cdot)$ the quadratic error between the predicted and the original time series. If the sequence $\{\mathbf{Y}_t\}_t$ is closed to the sequence $\{\mathbf{X}_t\}_t$, the prediction function $\mathbf{F}$ and $\mathbf{G}$ will be also similar. In this case, the error of prediction $d\left(\{\tilde{\mathbf{X}}_t\}_t, \{\mathbf{X}_t\}_t\right)$ will be low. On the other hand, dissimilar sequences will produce models unable to predict both of them, and then the error of prediction will be high.

As an illustration of the efficiency of the SVM-based similarity measure for motion descriptors, we have applied our approach on video representing two class of human activity, which consist of five different persons coming toward and going away the video camera (Fig. 2). The test set contains thus ten videos of length comprised between 30 and 40 frames. A 15-order prediction function is used ($H = 15$).

For each image sequences, motion descriptors are estimated and a dissimilarity matrix $\mathbf{D}$ is computed between each sequence of descriptors according to the similarity measure (16) (Fig. 3.a). As a comparison, a second dissimilarity matrix $\mathbf{D}'$ is computed by considering the Euclidean distance between the centroid of each sequence of descriptors, meaning removing the temporal information of the descriptors (Fig. 3.b). To quantify the benefit of the temporal model, an agglomerative clustering is applied on these two matrix. The classification rate is 100% for $\mathbf{D}$, whereas it is only 60% for $\mathbf{D}'$. This result shows the importance of taking into account temporal variations of the descriptors and highlights the relevance of the proposed temporal modeling based on the prediction function.

4. 2D REPRESENTATION OF THE FEATURE SPACE

The dissimilarity matrix computed from a set of videos gives distances between videos in a high-dimensional feature space. It is clearly impossible to directly display this feature space, and a 2D or 3D representation has to be derived in order to give an overview of the structure of the data set.

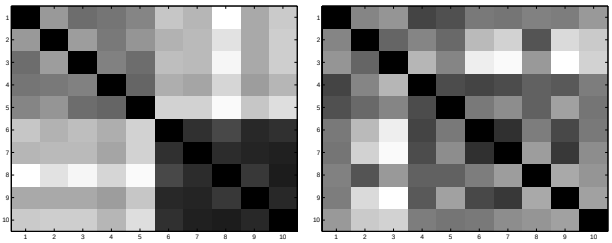
This issue is resolved by the use of the Curvilinear Component Analysis (CCA) [5], which is a nonlinear mapping algorithm allowing dimensionality reduction and preserving nonlinear structure. As an input, CCA just needs the dissimilarity matrix, and provides, as an output, the projection of the data onto a lower dimensional space (2D space in our case). During unfolding the feature space, CCA tries to preserve local topology. This fact means that local structures in the projected space are strongly related to those in the feature space. On the other hand, long range structures have no meaning, and should not be considered for any interpretation.

5. EXPERIMENTAL RESULTS

Our approach has been tested with a collection of 41 video shots representing TV news program (anchor scene) and sport videos (basket-ball, football, windsurf, trial). These videos have been chosen in such a way that dynamic content is relevant to organize the collection.

As detailed above, motion feature and temporal model are estimated (with prediction function of order $H = 20$), allowing to compute a dissimilarity matrix for the whole video set. The feature space corresponding to this matrix is then mapped on 2D space. The result of this mapping is displayed in figure 4. Video shots are represented by their median frame. The position of the center of the frames are the coordinates of the feature space elements into the 2D space. When reviewing the results, one has to keep in mind that no spatial criterion is explicitly accounted for (only the motion estimation is used). That makes the reviewing or

Fig. 2. Two class of activity "*come*" and "*go*". The set of videos contains 5 sequences with "*come*" activity and 5 sequences with "*go*".



(a) Matrix $\mathbf{D}$ computed from the temporal models

(b) Matrix $\mathbf{D}'$ computed from the descriptor's centroid

Fig. 3. Dissimilarity matrix computed for the 10 videos. Line entries 1 to 5 correspond to "*come*" activity, 6 to 10 to "*go*" activity.

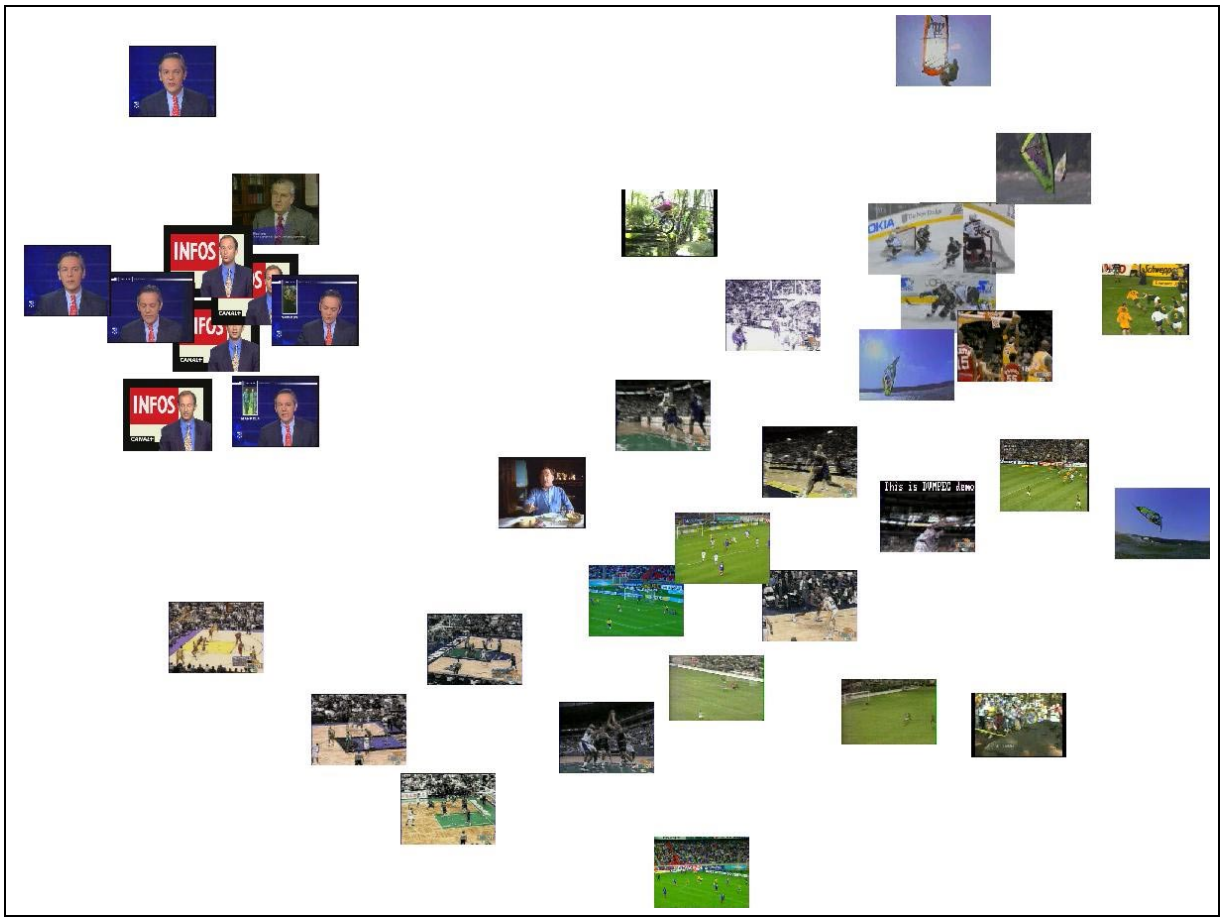


Fig. 4. Sport & News video collection. Nonlinear projection of the video feature space into a 2D space. Spatial video relationships exhibit the diversity of the dynamic properties contained in the documents.

Fig. 5. 2D representation of the *The Avengers* video collection.

results slightly misleading as one should inspect the local motions to check on the validity of the results.

The overview provided by the CCA of the feature space highlights the diversity of the collection content. We can observe that, as expected, a clear partition has been achieved between TV news and Sport programs. Among sport videos, the spatial organization of the documents seems determined by motion content, and more precisely, motion scale. Indeed, sport videos on the top-right corner of the representation exhibit close up views of the scene, whereas they correspond to broader plan when going to the bottom left corner.

A second experimentation result is displayed on figure 5. The video collection contains the 28 first shots of *The Avengers* TV movie[2]. In this example, as the videos are less characterized by typical motion or activity content, the 2D representation does not display a clear partition between different documents. However, one can observe that that scene containing close-up views on actors are concentrated on the center of the map, whereas broad plan and higher activity scenes are spreaded on the periphery.

6. CONCLUSION

We have proposed a method for overviewing the content of video collections according to scene motion. This method is unsupervised and allows to deal with generic video documents. The motion features are derived from a wavelet-based motion estimation algorithm and provide robust and stable informations on the optical flow. The temporal behavior of the descriptors is captured by a nonlinear model. This model consists in a prediction function estimated over the sequence of descriptors. An SVM algorithm is used to deal with the nonstationarity of the descriptors. The prediction error associated to a temporal model and a sequence of descriptors is defined as a similarity measure between videos and then allows to build a high-dimensional feature space where videos are projected. A CCA algorithm is used to reduce the dimensionality of it and to obtain a 2D representation of the collection of videos. Experiments on generic videos have shown the efficiency of the approach.

In the framework of the nonlinear temporal modeling, future research will focus on adding more low-level descriptors (such as color, texture, shape, audio). Indeed, the last result presented in the above section highlight the problem of organizing data according to various visual primitives and modalities in order to obtain a meaningfull representation of the collection. The main challenge will be then to define an interactive scheme for weighting the different information sources which allows the end-user to define his view point of the meaningfull represenation.

[2]This video is extracted from the AIM corpus developed within the French inter-laboratory research group ISIS and the French National Institute of Audiovisuel (INA)

7. REFERENCES

[1] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In H. Burkhardt and B. Neumann, editors, *European Conf. on Computer Vision, ECCV-98*, volume 1406 of *LNCS-Series*, pages 909–924, Freiburg, Germany, 1998. Springer-Verlag.

[2] E. Bruno and D. Pellerin. Global motion model based on B-spline wavelets : application to motion estimation and video indexing. In *Proc. of the 2nd Int. Symposium. on Image and Signal Processing and Analysis, ISPA'01*, June 2001.

[3] E. Bruno and D. Pellerin. Video structuring, indexing and retrieval based on global motion wavelet coefficients. In *Proceedings of International Conference of Pattern Recognition (ICPR)*, Quebec City, Canada, August 2002.

[4] O. Chomat and J. Crowley. Probabilistic recognition of activity using local appearance. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'99*, pages 104–109, June 1999.

[5] P. Demartines and J. Herault. Curvilinear component analysis: A self-organising neural network for non linear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, 1997.

[6] R. Fablet, P. Bouthemy, and P. Perez. Non parametric motion characterization using temporal gibbs models for content-based video indexing and retrieval. *IEEE Transactions on Image Processing*, 11(4):393–407, April 2002.

[7] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Proceeding of IEEE Neural Networks for Signal Processing, NNSP'97*, pages 24–26, September 1997.

[8] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.

[9] A. Smola and B. Schlkopf. A tutorial on support vector regression. Neurocolt2 technical report nc2-tr-1998-030, 1998.

[10] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[11] N. Vasconcelos and A. Lippman. Spatiotemporal motion model for video summarization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'97*, Santa Barbara, CA, 1997.

[12] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, 2(73):232–247, 1999.

[13] L. Zelni-Manor and M. Irani. Event-based analysis of video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'01*, volume 2, pages 123–130, Kauai Mariott, Hawai, December 2001.