Unsupervised Event Discrimination Based on Nonlinear Temporal Modelling of Activity

Bruno, Eric; Moenne-Loccoz, Nicolas; Marchand-Maillet, Stéphane

Pattern Analysis and Applications manuscript No.
(will be inserted by the editor)

Unsupervised Event Discrimination Based on Nonlinear Temporal Modeling of Activity Content

Eric Bruno, Nicolas Moenne-Loccoz and Stéphane Marchand-Maillet

Computer Vision and Multimedia Laboratory, University of Geneva

25 rue du Général Dufour

1211 Geneva 4, Switzerland

email: eric.bruno@cui.unige.ch

Abstract This paper deals with the problem of event discrimination in generic video documents. We propose an investigation on the design of an activity-based similarity measure derived from motion analysis. In an unsupervised context, our approach relies on the nonlinear temporal modeling of wavelet-based motion features directly estimated from the video frame. Based on SVM-regression, this nonlinear model is able to learn the behavior of the motion descriptors along the temporal dimension and to capture useful information about the dynamic content of the shot. A similarity measure

associated with our temporal model is then defined. This measure defines a metric between video segments according to spatial and temporal properties of the movements and provides a theoretic framework to compare, sort and classify videos. Experiments on a large annotated video database and a comparison with a similarity measure based on motion histograms shows that our approach is effective in discriminating between video events without any prior knowledge.

Keywords: Video analysis, activity recognition, motion estimation, SVM learning, video similarity measure.

1 Introduction

The problem of content-based management and manipulation of large video collections is subject to the diversity of the contents to handle. In specific domains such as video surveillance or medical videos, semantic content is generally well-defined and this allows to design systems based on learning machines that produce accurate classification of video content. In the case of generic videos (*i.e.* not restricted to a particular domain), semantic content is unbounded and unpredictable, making supervised techniques of little use. To tackle this problem, various systems have been proposed in the recent literature, such as VIBE [20], VideoQ [4], VIRAGE [9] or CueVideo [19]. These systems consider keyframes and/or video shots as basic entities from which audio-visual feature attributes are extracted. Video modeling, retrieval or browsing is then performed by exploring the feature space (*e.g.*

using retrieval by example, clustering, ...) so as to discover structures related to the semantic content. All of these data analysis operations are based on similarity measures associated to features. Defining such metrics is thus a key issue since they provide conceptual spaces to the system [8] from where significant and meaningful audio-visual contents would be extracted.

An other key issue is the problem of defining which information is to be considered to create relevant indices on videos. Video documents are information-rich spatio-temporal media on which many possible descriptions associated to various feature spaces exist. In [16], Roach *et al* have proposed a taxonomy that defines a compact structure to store and manage generic video content. This taxonomy provides a decomposition of the content within several generic properties, such as *genre*, *event*, *object* or *editing effects*. Automatically filling these fields would then help to build more complex representations of the knowledge and to reduce the semantic gap between low-level information and real content of documents. Among all properties that can characterize a video document, event information occupies an important place, since it provides a way to structure the stream, infer video genre or retrieve particular actions and stories.

The aim of the work reported here is to determine a distance measure able to effectively discriminate between generic events contained in video shots. More precisely, we investigate and evaluate how spatio-temporal motion information is effective in characterizing dynamic content and retrieving generic events. We assume that shots are related to one particular event

that the author wishes to highlight. The events are then defined as the main action of the shot (for example a goal in a soccer game, a close-view on a person speaking, etc...). While this assumption is not always true, it is generally verified and provides us a simple but realistic event segmentation. The problem then becomes that of defining an event-related similarity measure between two video shots of different length.

Motion is a natural feature to characterize events since it is related to dynamic content [7,25,26]. An efficient extraction of such an information owes to consider both the spatial and temporal properties of the motion-based descriptors. Spatio-temporal models have to be defined so that dynamic features are clearly expressed and create suitable descriptions for video documents. Spatio-temporal histograms of optical flow or motion vector of MPEG macroblocks are frequently used [11,20,23], but more sophisticated models have been proposed: Motion parameter trajectory combined to condensation algorithm are considered in [1], temporal Gibbs model of motion-related measures are used in [7] and a 3D Gabor decomposition performs a spatio-temporal video analysis in [5].

Our approach relies on our previous work on 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 (generally unreliable) segmentation stage. The estimated motion parameters then provide a robust, global, meaningful and compact description of activity content [3]. In this paper, a new temporal model based on *time series forecasting*

is introduced is order to capture temporal information from trajectories of the motion descriptors. In our framework, motion descriptors are estimated between any two consecutive frames of the shot so that the video sequence is characterized by a *temporal sequence of descriptors*. As video shots are considered as basic entities, the problem is to define a function able to compare two feature sequences of different length, (the length of each shot respectively) that correlates well with the similarity of events present within the shots. The problem is tackled by modeling sequences of descriptors using nonlinear prediction functions. The estimation of such models is done by using *Kernel Support Vector Machines in Regression* [18]. This operation can be viewed as a learning process of the temporal behavior of descriptors where trained functions are the prediction functions. An event-based similarity measure is then defined as the quadratic error between predicted and original descriptors. This measure avoids facing the problem of temporal alignment and shot length differences. The evaluation has been a important part of our work since about 900 video shots have been manually annotated in order to assess objectively the results provided by the similarity measure. The evaluation is reinforced by a comparison with a state-of-the-arts approach based on MPEG motion vector histograms, highlighting the benefits of our wavelet-based approach.

The paper is organized as follows. Section 2 describes the wavelet-based motion estimation and the motion descriptors derived from the motion wavelet coefficients. The nonlinear temporal model and similarity measure

are presented in section 3. Section 4 describes the video database and the event annotations proposed to evaluate the similarity measure efficiency. Experiments on this database are presented in section 5 and conclusion are drawn in section 6.

2 Motion feature extraction

The role of motion features is to define an unambiguous signature of the motion pattern induced by moving objects and camera displacements. Moreover, the signature has to be compact in order to present some good generalization properties. Many descriptors are possible such as eigenvectors [17], histograms [6] or affine parameters [22]. The motion descriptors we used here are based on the wavelet coefficients of the optical flow. This choice presents several advantages, including that a compact signature gives access to multiscale and orientation properties of the motion patterns [3].

Different strategies to estimate wavelet coefficients are also possible i.e. from MPEG motion vectors, from a dense flow field or directly from the image sequence. In this study, we concentrate on the latter solution. While motion vectors extracted from MPEG videos do not require extra computation, they are noisy and provide less reliable descriptors (in addition, our approach remains valid if videos are not MPEG encoded). An alternative solution would be to obtain coefficients from a wavelet decomposition of flow field estimated by other techniques. This solution does not offer any advantages since it has been shown that it does not improve accuracy [2,

24] and makes the computation more complex (estimation following by a decomposition).

.

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].

Consider an image sequence $I(\boldsymbol{p}_i, t)$ with $\boldsymbol{p}_i = (x_i, y_i)$ the location of each pixel in the image. The *brightness constancy assumption* [10] states that the image brightness $I(\boldsymbol{p}_i, t+1)$ at time $t+1$ 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 given scales 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 $L = 0$ and l defines the finest level of detail that can be fitted by the motion model.

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

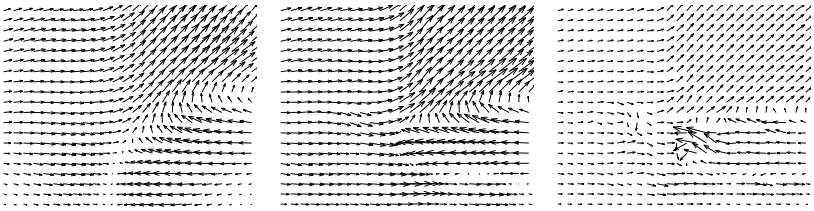
The motion parameter vector $\boldsymbol{\theta}$, containing wavelet coefficients $\boldsymbol{c}_{L,k_1,k_2}$ and $\boldsymbol{d}^{H,D,V}_{j,k_1,k_2}$ for all j, k_1, k_2 is estimated by minimizing an objective function

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

where $\rho(\cdot)$ is a robust norm error (M-estimator). The minimization step is achieved using an incremental and multiresolution estimation method [15].

The wavelet-based motion model enables one to estimate from 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 our work, 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 experiments, we have used a final scale $l = 3$ which corresponds to a motion model configured by 128 wavelet coefficients. Furthermore, using this low-resolution model speeds up the process allowing to process videos at a frame rate of about 2 fps on a standard 2GHz PC.

(a) Frame from *Mobile and Calendar* sequence



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

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

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

Fig. 1 Frame from *Mobile and Calendar* sequence and global motion estimated at various scale levels. B-spline of degree 2 were used to model motion.

2.2 Activity descriptors

As shown in Figure 1, the motion parameter vector $\boldsymbol{\theta}$ contains an accurate description of the optical flow. For the purpose of comparing video content, 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

$$\boldsymbol{\sigma} = \left[\sigma_0, \sigma_1^H, \sigma_1^D, \sigma_1^V, \sigma_2^H, \ldots, \sigma_l^V\right],$$

with

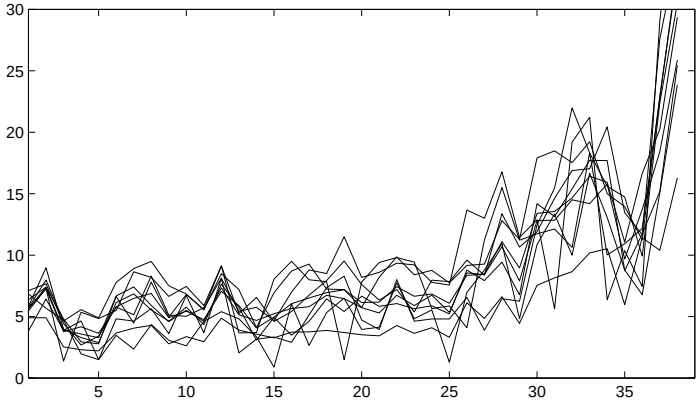
$$\sigma_0 = c_0^2 \tag{4}$$

$$\sigma_j^{H,D,V} = \sum_{k_1,k_2=0}^{2^j-1} \left|\boldsymbol{d}_{j,k_1,k_2}^{H,D,V}\right|^2, \forall j \in [1,l]$$

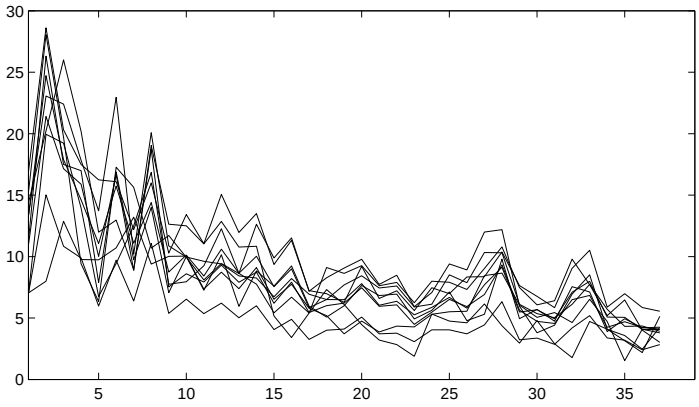
where l is the finest scale level used in (2), meaning that $\boldsymbol{\sigma}$ is a 10-component vector in our case, characterizing optical flow in terms of its global magnitude, scale and orientation.

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. Figures 3 display the temporal behavior of the descriptor $\boldsymbol{\sigma}$ estimated on two videos representing a person arriving toward and leaving away from the video camera (Fig. 2). As we can observe, the temporal evolution of descriptors is clearly related to the nature the activity. Hence, given such time series, the problem is now to define a similarity measure that takes into account both descriptor values and their temporal behavior.

Fig. 2 Two image sequences corresponding to activities "*arrive*" and "*leave*".



(a)



(b)

Fig. 3 Temporal sequence of the activity-based descriptor $\boldsymbol{\sigma}$ for activity: a) *arrive* and b) *leave*

3 SVM regression for nonlinear temporal modeling and similarity measure definition

3.1 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 $N+1$ the length 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 as

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

The multidimensional function $\mathbf{F}$ can be considered as a temporal model of the descriptors and therefore captures 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 may over-fits its dynamic content. On the other hand, the information characterized by the prediction function tends toward zeros as the model memory decreases.

In our case, the wavelet-based descriptor components are by definition supposed to be uncorrelated. Hence, the estimation of the multidimensional function $\mathbf{F}$ (eq. 5) can be achieve separately over each component. Let us note x^l the l^{th} component of $\mathbf{X}$, the problem therefore consists in estimating f^l such as

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

Then

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

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

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

so way that equation (6) can be written as $x_t^l = f^l(\mathbf{x}_{t-1}^l)$. The main difficulty in this approach is to estimate f^l efficiently. As the descriptor sequence is non-stationary, we have to estimate a nonlinear prediction function from the set of observations. Many regression techniques can used to solve this problem, and results obtained by using Support Vector Machines in regression show that this kernel-based algorithm is well-suited for such nonlinear estimation [14].

3.2 Support Vector Machines for regression

We present here a short description of SVM for regression. Further details can be found in [18,21], especially for issues related to the robustness of the algorithm. This classical problem of regression consists in approximating an unknown function $g: \mathbb{R}^D \to \mathbb{R}$ 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

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

where $\{\phi_i\}_{i=1}^{B}$ are basis functions. Parameters b and $\{c_i\}_{i=1}^{B}$ are unknown parameters that have to be estimated from the set of training pairs $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ by minimizing the functional

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

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

$$|x|_\epsilon = \begin{cases} 0 \text{ if } x < \epsilon \\ \\ x \text{ otherwise.} \end{cases} \tag{11}$$

In [21], Vapnik has shown that the function which minimize the functional (10) has the following form

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

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

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

The main interest of SVM techniques is that only the kernel K has to be known and the feature space spanned by the basis ϕ_i never need to be explicitly computed. This allows to use several types of basis functions, including infinite sets, providing modelers with a wide range of nonlinear models to approximate the unknown function g.

Here, we face sequences of visual descriptors where no prior info about the form of the solution is available. Hence, 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 inter-distances between observations $\{\mathbf{x}_t\}_{t=H}^{N}$ and thereby the smoothness of the solution in the observation space. We set it as follow

$$\gamma = \frac{1}{2}\left(\frac{1}{N-H}\sum_{t=H}^{N}||\Delta\mathbf{x}_t||^2\right)^{-1}, \tag{14}$$

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

3.3 Similarity measure as a prediction error

Let $\mathbf{F} = [f^1,\ldots,f^D]^T$ and $\mathbf{G} = [g^1,\ldots,g^D]^T$ be the prediction functions estimated on time series of descriptors $\{\mathbf{X}_t\}_{t=0}^{N}$ and $\{\mathbf{Y}_t\}_{t=0}^{M}$ respectively. From these prediction functions, we can built two new time series by *crossing models and descriptors*

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

and then define the symmetric similarity measure between sequences $\{\mathbf{X}_t\}_{t=0}^{N}$ and $\{\mathbf{Y}_t\}_{t=0}^{M}$

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

with $d(\cdot,\cdot)$ a function to measure the error between the original and the predicted time series. Note that here, since $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are by construction

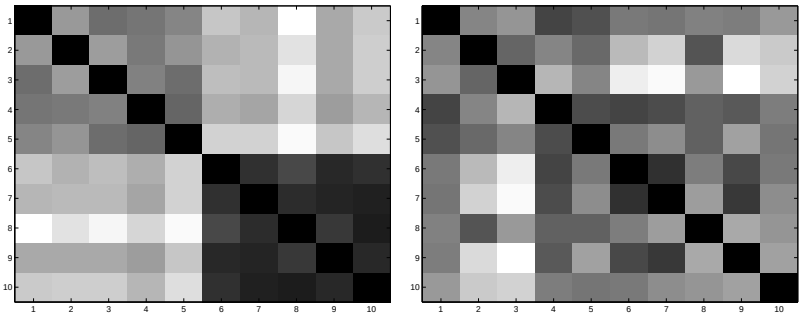
temporally aligned, the comparison between the two sequences may be done point-wise and any proper distance can be used. As the quadratic norm is a standard metric to measure such error, we have chosen to use it. We add a normalizing term in order to make the distance invariant to the magnitude of the descriptors

$$d\left(\{\tilde{\mathbf{X}}_t\}_t, \{\mathbf{X}_t\}_t\right) = \frac{1}{N-H} \frac{\sum_{t=H}^{N} ||\mathbf{X}_t - \tilde{\mathbf{X}}_t||}{\sum_{t=H}^{N} ||\mathbf{X}_t||}. \tag{17}$$

If the sequence $\{\mathbf{Y}_t\}_t$ exhibits a similar behavior to the sequence $\{\mathbf{X}_t\}_t$, the prediction function $\mathbf{G}$ will be able to give a good prediction of $\{\mathbf{X}_t\}_t$. 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 cross-predict each other, leading to high values of the similarity measure.

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

For each image sequence, motion descriptors are estimated and a dissimilarity matrix $\mathbf{D}$ is computed between each sequence of descriptors according to the similarity measure (16) (Fig. 4.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, which corresponds to remove the temporal information carried by the descriptors (Fig. 4.b). To quantify



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

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

Fig. 4 Dissimilarity matrix computed for a set of 10 videos containing 5 sequences with "*come*" activity and 5 sequences with "*go*". Line entries 1 to 5 correspond to "*come*" activity, 6 to 10 to "*go*" activity.

the benefit of the temporal model, an agglomerative clustering is applied on these two matrices. The final 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 to capture activity information from descriptor sequences.

4 The video database

In order to evaluate the proposed similarity measure, we have created a video database containing 830 video shots. These shots have been automatically extracted from the MPEG7 test video corpus using the shot detector proposed in [12]. They contain various genres, including sport (foot-

ball, basket-ball, wind-surf), sitcom series, variety program (involving many dance sequences), TV news and documentaries. This corpus illustrates typical TV broadcast by the variety of its contents.

Each shot has been manually annotated using one of the three following event labels:

- *Action* corresponds to high activity events, such as sport and dance sequences.
- *Human moving* corresponds to events representing human or crowd walking or doing large gestures.
- *Talking head* corresponds to close-up view on talking people, such as anchor scenes in news, dialog scenes in sitcom.

We chose these labels so that they meet three requirements: They should not be too specific in order to cover as much situations as possible, they are meaningful and they are exclusive one to each other. We note that the proposed labels do not totally satisfy the exclusive requirement since *Human moving* could also stand for *Action* as well as for *Talking head*. However, we have stated during the annotation stage that *Human moving* corresponds to situations involving human activities which do not fall neither in *Talking head* nor in *Action* events (it can be viewed as a default label). Other video shots that do not contain any of these three events (around 30% of the DB) have not been annotated but are still present in the database.

Activity-based descriptors and nonlinear temporal models have been estimated from each shot of the database, allowing to compute the similarity

matrices for the whole database. The parameter setting is as follows: Motion is modeled using a 3-level hierarchical B-spline decomposition of degree 2 providing a 10-dimension feature vector. Temporal models of feature vectors are estimated by prediction functions of order 20.

Video shots, annotations, descriptors and similarity matrices are contained within a SQL database associated with access tools which provides us with an effective framework to perform automatic evaluation of descriptors and similarity measures [13].

5 Experimental results

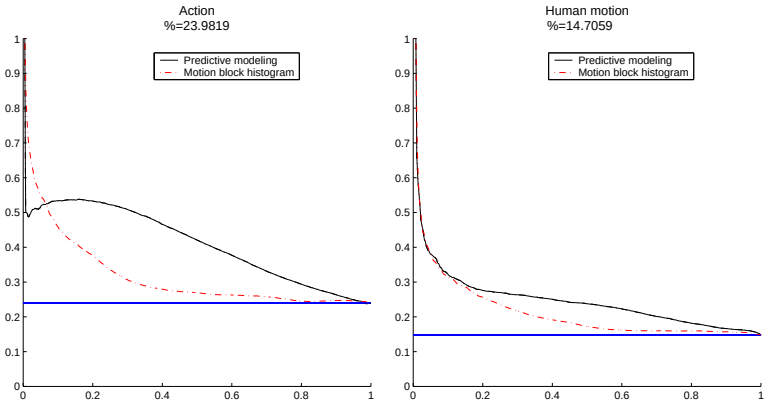
The quantitative evaluation of our method is given by Precision-Recall graphs computed on the annotated video database. Our approach has also been compared to a similarity measure based on motion histograms computed on MPEG motion vectors. Details of the implementation may be found in [11].

Figure 5 displays average Precision and Recall computed on all video shots for each event label and for the two algorithms (see legend). Horizontal lines in the graphs represent the statistical mean value of Precision when documents are randomly selected (which is equal to the percentage of labels in the database). The fact that P-R curves are above these lines simply means that the retrieval operation performs better than a random selection. We can observe that for the three events, P-R curves are largely above the "random case", which validates the ability of the similarity measure to

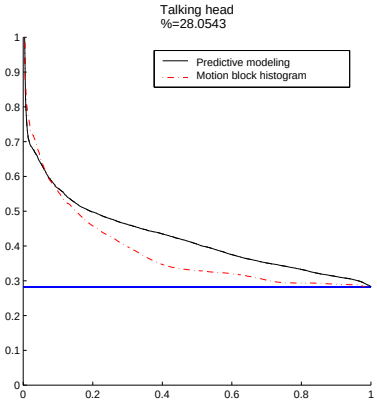
sort documents according to their dynamic content. However, it can be also noticed that performances for our *Human moving* class retrieval are poor compared to the two other classes of events. This result can be explained by the fact that this event is quite ill-defined as it has been used to label heterogeneous dynamic contents. In addition, there is an non-negligible overlap between labels since it is not obvious in some cases to decide whether a particular event should be annotated as *Human moving* or *Talking head* and *Human moving* or *Action*. This is a recurring problem in multimedia asset management that cannot simply be solved or ignored.

We can also note that our similarity measure outperforms the motion histogram approach, especially for recall greater than 0.2. These results show that the predictive model has better generalization properties, *i.e.* is able to better extract the underlying characteristics of the features trajectories, and thus permits a wider retrieval of documents belonging to the same class of activity.

Figure 6 shows examples of query results for the three class of events. Since characterization is done based on motion, one should keep in mind that the temporal aspect is crucial when reviewing the results. These examples confirm the results presented above, where videos retrieved from a *Human moving* query (Fig. 6.b) exhibit more false detections than for *Action* and *Talking head* cases (respectively in Fig. 6.a and 6.c). However, as it can be seen in figure 6.b, we have observed that misclassified videos generally correspond to documents with ambiguous annotation but which present



(a) (b)



(c)

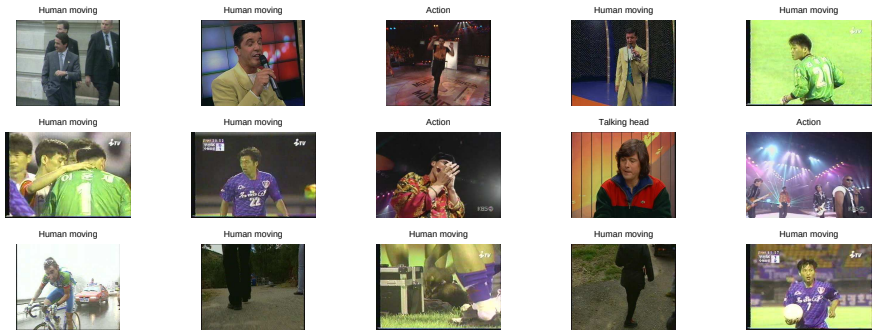
Fig. 5 Precision-Recall graphs for a) Action events, b) Human moving events and c) Talking head events. Horizontal lines represent the percentage of each label in the database (numerical values are given in titles). The sum is not equal to 100% because of the non-annotated shots.

similarities with the query document. A solution would be to refine our annotations to avoid ambiguities on event classification.

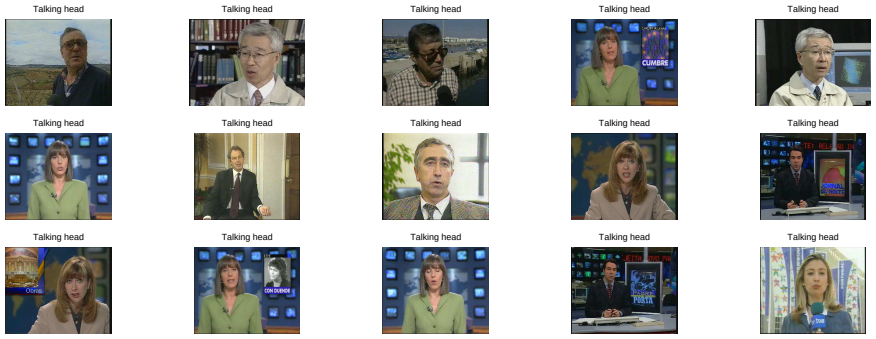
6 Conclusion

The event-based similarity measure presented in this paper offers an unsupervised tool to compare video shots according to their dynamic content. This measure is based on the motion analysis of the image sequences. The motion features are derived from a wavelet-based motion estimation algorithm which provides a multiscale, robust and stable information on the optical flow. Then, the temporal behavior of the descriptors is captured by nonlinear models consisting in prediction functions estimated over the sequence of descriptors. An SVM algorithm is used to deal with the highly non-stationary nature of the descriptors. The prediction error computed between temporal models and sequences of descriptors defines a similarity measure that is related to dynamic content over the whole shots. Using our original approach, we emancipate from the costly process of temporally aligning sequences of different lengths, while preserving most of their temporal information. Such information is typically lost in other current techniques such as that based on histograms. Experiments on a large annotated video database gave encouraging results, also when compared to histogram-based measure. The similarity measure is indeed able to discriminate generic events related to human activities from a set of TV broadcast video. More generally, since the approach is unsupervised and hence not

(a)



(b)



(c)

Fig. 6 First 14 samples of retrieved video shots for a query belonging to a) Action, b) Human moving and c) Talking head. The first image on top left corner corresponds to the query and retrieved shots are ordered from left to right and top to down.

restricted to particular events, the above results show the ability of the proposed metric to map videos in a continuous representation space closely related to our perception of events.

In the framework of nonlinear temporal modeling, future research will focus on adding more low-level descriptors (such as color, texture, audio) to describe video shots. Indeed, we would like to enhance the proposed approach so as to be able to compare video content according to various information sources. Beyond this objective, our aim is to define an interactive scheme for weighting these different information sources so that end-users will be able to define their own video metric to browse and search video documents.

7 Acknowledgments

This work is funded by the swiss Interactive Multimodal Information Management (IM2) and the EU IST Multimodal Meeting Manager (M4) projects.

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. S.F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong. A fully automated content-based video search engine supporting spatio-temporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):602–615, September 1998.

5. 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.

6. Z. Duric, E. Rivlin, and A. Rosenfeld. Qualitative description of camera motion from histograms of normal flow. In *ICPR00*, volume III, 2000.

7. 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.

8. P. Gardenfors. *Conceptual spaces as a basis for cognitive semantics*. Philosophy and Cognitive Science, A. Clark et al. Kluwer, Dordrecht, 1996.

9. A. Hampapur, A. Gupta, B. Horowitz, C. Shu, C. Fuller, J. Bach, M. Gorkani, and R. Jain. Virage video engine. In *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases*, volume 3022, pages 188–197, San-Jose, CA, February 1997.

10. B.K.P. Horn and B.G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–204, 1981.

11. A.K. Jain, A. Vailaya, and X. Wei. Query by video clip. *Multimedia Syst.*, 7(5):369–384, 1999.

12. Bruno Janvier, Eric Bruno, Stéphane Marchand-Maillet, and Thierry Pun. Information-theoretic framework for the joint temporal partioning and representation of video data. In *Proceedings of the European Conference on Content-based Multimedia Indexing, CBMI'03*, September 2003.

13. Nicolas Moënne-Loccoz, Bruno Janvier, Stéphane Marchand-Maillet, and Eric Bruno. Managing video collections at large. In *Proceedings of the First Workshop on Computer Vision Meets Databases CVDB'04*, Paris, France, 2004.

14. 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.

15. 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.

16. M. Roach, J. Mason, L-Q. Xu, and F. W. M. Stentiford. Recent trends in video analysis: A taxonomy of video classification problems. In *Proceedings of the 6th IASTED Int. Conf. on Internet and Multimedia Systems and Applications*, August 2002.

17. Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'00*, volume 1, pages 111–118, Hilton Head, SC, June 2000.

18. A. Smola and B. Scholkopf. A tutorial on support vector regression. Neurocolt2 technical report nc2-tr-1998-030, 1998.

19. S. Srinivasan, D. Ponceleon, A. Amir, and D. Petkovic. What is in that video anyway?: In search of better browsing. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pages 388–392, Florence, Italy, June 1999.

20. C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C.A. Bouman, and E.J. Delp. Vibe: A compressed video database structured for active browsing and search. *IEEE Transactions on Multimedia*, 1(6):103–118, February 2004.

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

22. 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.

23. V.V. Vinod. Activity based video shot retrieval and ranking. In *ICPR 98*, pages 682–684, 1998.

24. Y.T. Wu, T. Kanade, C.C. Li, and J. Cohn. Image registration using wavelet-based motion model. *International Journal of Computer Vision*, 38(2):129–152, July 2000.

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

26. 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.