Non-parametric motion activity analysis for statistical retrieval with partial query

Ronan Fablet, Patrick Bouthemy

JOURNAL OF MATHEMATICAL IMAGING AND VISION, 14(3), 2001

Non-parametric motion activity analysis for statistical retrieval with partial query

RONAN FABLET
IRISA/CNRS
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France
rfablet@irisa.fr

PATRICK BOUTHEMY
IRISA/INRIA
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France
bouthemy@irisa.fr

Received July 10th, 2000; Revised December 8th, 2000

Abstract. We present an original approach for motion-based video retrieval involving partial query. More precisely, we propose a unified statistical framework allowing us to simultaneously extract entities of interest in video shots and supply the associated content-based characterization, which can be used to satisfy partial queries. It relies on the analysis of motion activity in video sequences based on a non-parametric probabilistic modeling of motion information. Areas comprising relevant types of motion activity are extracted from a Markovian region-level labeling applied to the adjacency graph of an initial block-based partition of the image. As a consequence, given a set of videos, we are able to construct a structured base of samples of entities of interest represented by their associated statistical models of motion activity. The retrieval operations is then formulated as a Bayesian inference issue using the MAP criterion. We report different results of extraction of entities of interest in video sequences and examples of retrieval operations performed on a base composed of one hundred video samples.

Keywords: motion information, statistical models, video indexing, video retrieval with query by example, partial query, motion-based image segmentation, motion classification

1. Introduction

Retrieving multimedia documents through large databases is of growing importance in various application fields such as audio-visual archives, multimedia editing, meteorology, traffic surveillance. In particular, to cope with the increasing amount of video data, it is required to define appropriate automatic or semi-automatic schemes for video characterization based on their content, and to associate some measure of content similarity to these description schemes [2, 6].

As far as visual documents are concerned, retrieval schemes can be divided into three main categories: retrieval with textual query, retrieval with query by sketch, retrieval with query by example. Textual query simply consists in using natural language to express a query [33]. It implies to assign a list of key-words to a video, as currently done for audio-visual archives in a manual manner. However, manual annotating cannot cope with the tremendous amount of video data to be analyzed. Automatic semantic characterization of visual content can be performed from the transcription of audio stream as in Informedia project [33]. Nevertheless, the correspondence between the audio stream and the visual content is not always guaranteed, and visual information represents also an important cue not to be neglected for video retrieval. Besides, methods for semantic characterization of visual content have also been proposed [24, 31]. For instance, in [31], movies are classified into romance and action films, and in [24], features are proposed to detect explosions or waterfalls in video sequences. However, the achieved characterization reveals rough or dedicated, and automatic annotating of videos still remains beyond reach for non-specific video bases.

The second category of approaches for video retrieval relies on querying by sketch. To introduce more flexibility in the retrieval process, a sketch drawn by the user to express his/her query can be considered [4, 7]. For instance in [7], the user can draw a sketch representing the global shape of the object of interest and can specify its direction of displacement by an arrow. Thus, for queries such "retrieve cars going to the right" or "retrieve skiers moving to the left of the image", this scheme is efficient. However, this approach does not allow to formulate a wide range of queries. For instance, how to sketch a query such as "retrieve rugby game samples".

The third class of approaches handles video queries formulated as searching for similar examples as given video samples [1, 15, 21, 23, 32]. It benefits from the large research efforts devoted to the automatic extraction of numerical content-based descriptors related to color, texture or motion information. These low-level features are stored to compare videos using some feature-based similarity measure. The main limitation of video retrieval with query by example lies in the requirement for the user to be able to provide the system with an initial video query representative of the kind of samples he/she looks for. However, methods relying on query by example currently appears more suited to deal with the variety of contents present in video bases.

As far as retrieval with query by example is concerned, the proposed solutions mainly consider global queries [1, 11, 12, 15, 21, 23, 32]. These schemes rely on a global characterization of video content. Nevertheless, the content of the submitted video query is seldomly of interest as a whole w.r.t. user point of view. Furthermore, it seems important to offer the opportunity to the user to focus on specific areas of the video. Therefore, handling partial queries appears necessary. This reveals complex since it requires to automatically extract and characterize entities of interest (w.r.t. user expectations) from each sample of the considered video base.

In this paper, we propose an original approach to tackle this issue while focusing on dynamic content analysis. We have designed a unified statistical framework for the extraction and the characterization of entities of interest in video shots. It relies on a non-parametric probabilistic modeling of motion content in terms of motion activity. In addition, our method provides a complete scheme for video retrieval using partial query embedded in a statistical framework. The remainder of this paper is organized as follows. Section 2 outlines the general ideas underlying our work. Section 3 presents the local motion-related measurements we exploit for non-parametric motion activity modeling. In Section 4, the statistical modeling of motion information and the issue of estimating these models is addressed. Section 5 is concerned with image segmentation w.r.t motion activity with a view to automatically extracting entities of interest. Section 6 deals with motion-based video retrieval with partial query. Experiments are reported in Section 7, and Section 8 contains concluding remarks.

2. Problem statement

As stressed previously, handling partial queries primary implies to extract entities likely to be of interest for the user. Even if a simple block-based analysis at different scales could supply a first answer to these requirements, some segmentation of the scene, if available, appears more appropriate. The complexity of the segmentation task lies in the variety of situations to cope with in non-dedicated video bases. In a wide range of contexts, the searched entities of interest are nevertheless closely related to specific dynamic content. Hence, motion information represents an important cue for content-based video indexing and retrieval. As far as motion content is concerned, we have obviously to face different kinds of partial video queries. For instance, as illustrated in Figure 1(a-b), the user might be interested in a given moving element in the scene. Such a situation can be tackled using usual motion detection or segmentation techniques

[14, 17, 20, 22, 28] provided the scene is not too complex to process. Other cases have also to be considered such as areas of interest composed of different entities. In Figure 1(c-d), we display an example of rugby video sequence within which one might focus on a particular area of the playing field where the game actually lies. Our goal is to define an appropriate method able to handle these different kinds of situations. Furthermore, in addition to the extraction of entities of interest in video shots, we aim at performing within the same framework the associated characterization w.r.t. motion content with a view to straightforwardly tackling video retrieval with partial query.

Motion detection or segmentation techniques have been exploited to detect and track entities of interest in the context of video indexing [7, 9, 16]. Motion segmentation [17, 22] is usually expressed as the explicit partitioning of the image into homogeneous regions involving 2D parametric motion models. However, they remain highly computationally expensive and not reliable enough to deal with large video corpus. Besides, in case of complex dynamic scenes such as articulated motions, a single object may be divided into several different regions. Grouping them into a meaningful entity remains a difficult issue. On the other hand, motion detection schemes [14, 20, 22, 28] first require in case of a mobile camera the estimation of the dominant image motion assumed to be due to camera motion and aims at separating moving objects from the background, but no further characterization of the associated areas is straightforwardly available. In both cases, the motion-based description of extracted objects is generally their trajectories and the retrieval process relies on trajectory matching [10].

For retrieval issues, the description of motion information is as important as the extraction of the entities of interest. In addition, characterizing scene motion by object trajectories might appear insufficient and inappropriate. Besides, schemes have been recently proposed for global motion characterization based either on local motion-related measures [12, 15, 26] or on dense optic flow fields [11, 13, 21, 32]. In this paper, we adapt this kind of approaches to deliver a relevant motion characterization in a more flexible way compared to descriptions based on 2D parametric motion models. Motion activity will be described using the probabilistic framework we have introduced in [15]. To solve for the extraction of entities of interest, we resort to a Markovian region-level labeling approach applied to the adjacency graph resulting from an initial spatial partition of the image into blocks. This is achieved by exploiting the statistical similarity measure derived from our probabilistic modeling framework. This means that the latter allows us to produce both the desired motion segmentation and the appropriate motion classification.

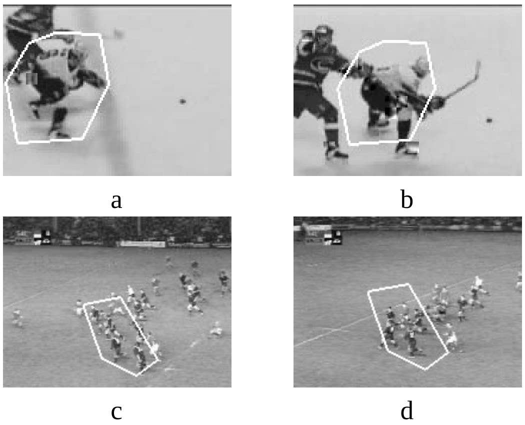


Fig. 1. Two video samples involving different kinds of entities of interest: (a-b) tracking of a given hockey player; (c-d) focus on a particular area of a rugby playing field. The bounding polygon encompassing the area of interest is displayed in white.

To cope with retrieval using partial query, we use the proposed motion activity segmentation scheme to build a base of entities of interest extracted within representatives key-frames of the elementary shots of the processed video set. We store each extracted entity and its associated statistical motion activity model. Given a query (video sample), we similarly extract entities of interest in the submitted example. The user selects one of these entities as the partial query. Since our motion characterization relies on a probabilistic framework, we can formulate the motion-based retrieval process as a Bayesian inference issue [15, 30].

3. Local motion-related measurements

To characterize motion content within a given region, our approach relies on the analysis of the distribution of local motion-related measurements. Two kinds of local motion-related quantities can be exploited. On one hand, one can consider dense optic flow fields [11, 13, 21, 32]. In our context, the computation of such fields reveals time consuming and situations likely to be encountered are too complex to get accurate and reliable estimation. As a consequence, we prefer to consider local motion-related quantities directly derived from the spatio-temporal derivatives of the intensity function [12, 15, 26].

Since our goal is to characterize the actual dynamic content of the scene, we have first to cancel camera motion. To this end, we estimate the dominant image motion between two successive images, and we assume that it is due to camera motion. To cancel camera motion, we then warp the preceding and next images of the selected key-frame in the shot onto the key-frame and process these three images.

3.1. Dominant motion estimation:

To model the global transformation between two successive images, we consider a 2D affine motion model (a 2D quadratic model could also be considered). The velocity $\mathbf{w}_\Theta(p)$, at pixel p, related to the affine motion model parameterized by Θ is given by:

$$\mathbf{w}_\Theta(p) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \tag{1}$$

with $p = (x, y)$ and $\Theta = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6]$. The six affine motion parameters are computed with the robust gradient-based incremental estimation method described in [27]. It comes to solve:

$$\widehat{\Theta} = \arg\min_\Theta \sum_{p \in \mathcal{S}} \rho\left(DFD(p, \Theta)\right) \tag{2}$$

where $DFD(p, \Theta) = I(p + \mathbf{w}_\Theta(p), t + 1) - I(p, t)$, and $I(\cdot, t)$ and $I(\cdot, t+1)$ are the image intensity functions at times t and $t+1$, and $\mathcal{S}$ the support of the estimation, i.e. the image grid. ρ is a hard-redescending M-estimator. In practice, we consider the Tukey's biweight function. This function ρ and its derivative ψ depend on a parameter C and are defined as follows:

$$\rho(x, C) = \begin{cases} \dfrac{x^6}{6} - \dfrac{C^2 x^4}{2} + \dfrac{C^4 x^2}{2} & \text{if } |x| < C \\ \dfrac{C^6}{6} & \text{otherwise} \end{cases} \tag{3}$$

$$\psi(x, C) = \begin{cases} x(x^2 - C^2)^2 & \text{if } |x| < C \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The use of a robust estimator ensures the dominant image motion estimation not to be sensitive to secondary motions due to mobile objects in the scene. The minimization is performed by means of an iterative reweighted least-square technique embedded in a multiresolution framework. More precisely, we operate with an incremental strategy. The initial value Θ^0 is set to zero. A succession of refinements $\Delta\Theta^k$ are then computed and cumulated using first order approximations r_p of the residual $DFD(p, \widehat{\Theta_k})$, where $\widehat{\Theta_k}$ denotes the sum of increments computed before iteration k and $\Delta\Theta^k$ the increment to be estimated at iteration k according to:

$$\widehat{\Delta\Theta^k} = \arg\min_{\Delta\Theta^k} \sum_{p \in \mathcal{S}} \rho(r_p) \tag{5}$$

where r_p is expressed as:

$$\begin{aligned} r_p &= I(p + \mathbf{w}_{\widehat{\Theta^k}}(p), t+1) - I(p, t) \\ &+ \nabla I(p + \mathbf{w}_{\widehat{\Theta^k}}(p), t+1) \cdot \mathbf{w}_{\Delta\Theta^k}(p) \end{aligned} \tag{6}$$

with ∇I denoting the spatial intensity gradient. The estimate $\widehat{\Theta^{k+1}}$ at iteration $k+1$ is updated as follows:

$$\widehat{\Theta^{k+1}} = \widehat{\Theta^k} + \widehat{\Delta\Theta^k} \tag{7}$$

Increments are estimated and cumulated until the stopping criterion on the norm of $\widehat{\Delta\Theta^k}$ is satisfied.

3.2. Local motion-related quantity:

We now process the image sequence generated by compensating for the estimated dominant image motion. To evaluate the residual motion in the compensated image sequence, the following local motion-related quantity is considered:

$$v_{obs}(p) = \frac{\displaystyle\sum_{s \in \mathcal{F}(p)} \|\nabla I^*(s)\| \cdot |I_t^*(s)|}{\max\left(\eta^2, \displaystyle\sum_{s \in \mathcal{F}(p)} \|\nabla I^*(s)\|^2\right)} \tag{8}$$

where $I^*(s)$ is the intensity function at point s in the warped image, $\mathcal{F}(p)$ a 3×3 window centered on p, η^2 a predetermined constant related to the noise level in uniform areas (typically, $\eta = 5$), and I_t^* the temporal derivative of the intensity function I^*. $I_t^*(p)$ is approximated by a simple finite difference. Whereas the normal flow measure $\frac{I_t^*(p)}{\|\nabla I^*(p)\|}$ turns out to be very sensitive to noise attached to the computation of spatio-temporal derivatives of the intensity function, the considered motion-related measurement forms a more reliable quantity, still simply computed from the intensity function and its derivatives. This quantity has already been successively used for motion detection issues [14, 20, 28], and for motion-based video indexing and retrieval [12, 15].

As stressed previously, our approach relies on a statistical modeling of the distribution of the local motion-related measurements. It can be regarded as an extension of texture modeling for grey level images. In our approach, local motion quantities play a role similar to grey levels for texture analysis not withstanding the continuous nature of the exploited motion information. Since the proposed Gibbsian modeling framework is based on cooccurrence measurements, it in fact requires to use motion-related observations defined over a finite set. Besides, to ensure feasible comparison of motion content between different videos, we introduce in practice a quantization of the continuous motion measurements within a predefined bounded interval. Another reason to fix a limit beyond which these local motion measurements are no more regarded as usable is due to the fact that gradient-based motion measurements known to be valid for rather small motion magnitude. Typically, sampling within $[0, 4]$ on 16 levels proves accurate enough in our previous work [12, 15]. We had also investigated logarithmic quantization but it did not prove relevant in our experiments. Let Λ be the discretized range of values for $\{v_{obs}(p)\}$.

4. Statistical modeling of motion activity

4.1. Temporal Gibbs models of motion activity

To characterize motion activity in an area of interest within the key-frame of a given video shot, we exploit the probabilistic framework presented in [15] which involves non parametric statistical models of motion activity. We briefly outline the scheme developed for characterizing global motion content within video shots (further details can be found in [15]) and specify it to the case of a given spatial area.

Let $\{x_k\}$ be the sequence of quantized motion-related quantities for the processed shot and k_0 the frame number of the selected key-frame. Let $\mathcal{R}$ denote the spatial region of interest in the image k_0 and $\{x_k^{\mathcal{R}}\}$ the restriction of sequence $\{x_k\}$ on the spatial support of region R. We assume that the pair $x^{\mathcal{R}} = \{x_{k_0}^{\mathcal{R}}, x_{k_0+1}^{\mathcal{R}}\}$ is the realization of a first-order Markov chain:

$$P_{\mathcal{M}}(x^{\mathcal{R}}) = P_{\mathcal{M}}(x_{k_0}^{\mathcal{R}}) \prod_{p\in\mathcal{R}} P_{\mathcal{M}}(x_{k_0+1}^{\mathcal{R}}(p)|x_{k_0}^{\mathcal{R}}(p)) \quad (9)$$

where $\mathcal{M}$ refers to the motion activity model. $P_{\mathcal{M}}(x_{k_0}^{\mathcal{R}})$ designates the a priori distribution of $X_{k_0}^{\mathcal{R}}$. We will consider in practice a uniform law for $P_{\mathcal{M}}(x_{k_0}^{\mathcal{R}})$.

In this causal modeling framework, we evaluate only temporal interactions, i.e., cooccurrence of two given values at the same grid point at two successive instants. The advantages are two-fold. First, it permits to handle certain kinds of temporal non-stationarity. Second, it enables an exact computation of the conditional likelihood $P_{\mathcal{M}}(x^{\mathcal{R}})$. This is crucial since it allows us to achieve model estimation in an easy way and to define an appropriate statistical similarity measure based on the Kullback-Leibler divergence [3].

In addition, we consider an equivalent Gibbsian formulation of $P_{\mathcal{M}}(x_{k_0+1}^{\mathcal{R}}(p)|x_{k_0}^{\mathcal{R}}(p))$ which comes to the introduction of potentials $\Psi_{\mathcal{M}}\left(x_{k_0+1}^{\mathcal{R}}(p), x_{k_0}^{\mathcal{R}}(p)\right)$ defining the model $\mathcal{M}$ such as:

$$\begin{gathered} P_{\mathcal{M}}\left(x_{k_0+1}^{\mathcal{R}}(p)|x_{k_0}^{\mathcal{R}}(p)\right) = \\ \exp\left[\Psi_{\mathcal{M}}\left(x_{k_0+1}^{\mathcal{R}}(p), x_{k_0}^{\mathcal{R}}(p)\right)\right] \\ \text{with } \forall \nu' \in \Lambda,\ \sum_{\nu\in\Lambda} \exp\left[\Psi_{\mathcal{M}}(\nu, \nu')\right] = 1 \end{gathered} \quad (10)$$

Besides, this Gibbsian setting establishes a correspondence with cooccurrence distributions [15, 19, 34]. In fact, the conditional likelihood $P_{\mathcal{M}}(x^{\mathcal{R}})$ can be expressed according to an exponential formulation involving a dot product :

$$P_{\mathcal{M}}(x^{\mathcal{R}}) = P_{\mathcal{M}}(x_{k_0}^{\mathcal{R}}) \cdot \exp\Big[\Psi_{\mathcal{M}} \bullet \Gamma^{\mathcal{R}}\Big] \quad (11)$$

where $\Gamma^{\mathcal{R}} = \{\Gamma^{\mathcal{R}}(\nu, \nu')\}_{(\nu,\nu')\in\Lambda^2}$ is the cooccurrence distribution defined by:

$$\begin{gathered} \Gamma^{\mathcal{R}}(\nu, \nu') = \\ \sum_{p\in\mathcal{R}} \delta(\nu - x_{k_0+1}^{\mathcal{R}}(p)) \cdot \delta(\nu' - x_{k_0}^{\mathcal{R}}(p)) \end{gathered} \quad (12)$$

where δ is the Kronecker symbol. $\Psi_{\mathcal{M}} \bullet \Gamma^{\mathcal{R}}$ is the dot product between the cooccurrence distribution $\Gamma^{\mathcal{R}}$ and the potentials $\Psi_{\mathcal{M}}$:

$$\Psi_{\mathcal{M}} \bullet \Gamma^{\mathcal{R}} = \sum_{(\nu,\nu')\in\Lambda^2} \Psi_{\mathcal{M}}(\nu, \nu') \cdot \Gamma^{\mathcal{R}}(\nu, \nu') \quad (13)$$

The modeling scheme can be considered as non-parametric in two ways. On one hand, the statistical model $\mathcal{M}$ does not refer to a 2D parametric motion model. Whereas the latter aims at representing the global transformation occurring between successive frames, our approach copes with the description of motion information in terms of motion activity. On the other hand, from a statistical point of view, our approach is also non-parametric in the sense that the distribution $\{P_{\mathcal{M}}(\nu|\nu')\}_{(\nu,\nu')\in\Lambda^2}$ is not assumed to follow a known parametric law (Gaussian,...).

As far as video indexing and retrieval is concerned, the availability of the exponential formulation (11) is interesting for several reasons. First, it makes feasible and simple the computation of the conditional likelihood $P_{\mathcal{M}}(x^{\mathcal{R}})$ for any sequence x and region $\mathcal{R}$. Second, for indexing issues, the storage of the motion-related quantities $x^{\mathcal{R}}$ is not needed. In fact, all motion information are entirely captured by the cooccurrence distribution $\Gamma^{\mathcal{R}}$ and then by the activity model $\mathcal{M}$.

4.2. Maximum likelihood estimation

Given a sequence of motion-related measurements $x^{\mathcal{R}}$ within a region $\mathcal{R}$, we aim at identifying the model $\widehat{\mathcal{M}}^{\mathcal{R}}$, specified by its potentials $\Psi_{\widehat{\mathcal{M}}^{\mathcal{R}}}$, best fitting $x^{\mathcal{R}}$. To this end, we consider the Maximum Likelihood (ML) criterion:

$$\widehat{\mathcal{M}}^{\mathcal{R}} = \arg\max_{\mathcal{M}} P_{\mathcal{M}}(x^{\mathcal{R}}) = \arg\max_{\mathcal{M}} LL_{\mathcal{M}}(x^{\mathcal{R}})$$
$$\text{with } LL_{\mathcal{M}}(x) = \ln P_{\mathcal{M}}(x^{\mathcal{R}}) \tag{14}$$

In fact, the considered temporal Gibbsian modeling framework consists in a product of $|\mathcal{R}|$ independent first-order Markov chains defined at each point of the region $\mathcal{R}$. These Markov chains are characterized by their common transition matrix $\{P_{\widehat{\mathcal{M}}^{\mathcal{R}}}(\nu|\nu')\}_{(\nu,\nu')\in\Lambda^2}$. Therefore, the ML model estimate $\widehat{\mathcal{M}}^{\mathcal{R}}$ is readily determined from the empirical estimation of the transition probability $P_{\widehat{\mathcal{M}}^{\mathcal{R}}}(x^{\mathcal{R}}_{k_0+1}(r)|x^{\mathcal{R}}_{k_0}(r))$ as follows:

$$\Psi_{\widehat{\mathcal{M}}^{\mathcal{R}}}(\nu,\nu') = \ln\left(\frac{\sharp\{x^{\mathcal{R}}_{k_0+1}(p)=\nu, x^{\mathcal{R}}_{k_0}(p)=\nu'\}}{\sharp\{x^{\mathcal{R}}_{k_0}(p)=\nu'\}}\right) \tag{15}$$

Using cooccurrence notation, it comes to:

$$\Psi_{\widehat{\mathcal{M}}^{\mathcal{R}}}(\nu,\nu') = \ln\left(\frac{\Gamma^{\mathcal{R}}(\nu,\nu')}{\sum_{\vartheta\in\Lambda}\Gamma^{\mathcal{R}}(\vartheta,\nu')}\right) \tag{16}$$

4.3. Parzen-Rosenblatt estimation

In Section 5, we will exploit the motion activity modeling framework described above to characterize the dynamic content in a given image block. The use of the ML criterion can become irrelevant if the number of available samples is too small. To overcome this problem occurring for small regions, we consider a regularized solution to the model estimation issue by using a Parzen-Rosenblatt (PR) window estimator [29].

For given sequence x and area $\mathcal{R}$, it comes to substitute in relation (16) a regularized version $\widetilde{\Gamma}^{\mathcal{R}}$ for the cooccurrence distribution $\Gamma^{\mathcal{R}}$. To get $\widetilde{\Gamma}^{\mathcal{R}}$, $\Gamma^{\mathcal{R}}$ is convolved with a Gaussian kernel of variance σ^2:

$$\widetilde{\Gamma}^{\mathcal{R}}(\nu,\nu') = \sum_{(\gamma,\gamma')\in\Lambda^2} \eta(\gamma-\nu)\,\eta(\gamma'-\nu')\cdot\Gamma^{\mathcal{R}}(\gamma,\gamma') \tag{17}$$

where $\eta(\gamma-\nu)$ is the weight of the Gaussian PR window given by:

$$\eta(\gamma-\nu) = \frac{\exp\left[-(\gamma-\nu)^2/2\sigma^2\right]}{\sum_{\upsilon\in\Lambda}\exp\left[-(\upsilon-\nu)^2/2\sigma^2\right]} \tag{18}$$

Therefore, we infer the expression of potentials $\Psi_{\widetilde{\mathcal{M}}^{\mathcal{R}}}$ of the PR model estimate $\widetilde{\mathcal{M}}^{\mathcal{R}}$ as follows:

$$\Psi_{\widetilde{\mathcal{M}}^{\mathcal{R}}}(\nu,\nu') = \ln\left(\frac{\widetilde{\Gamma}^{\mathcal{R}}(\nu,\nu')}{\sum_{\vartheta\in\Lambda}\widetilde{\Gamma}^{\mathcal{R}}(\vartheta,\nu')}\right) \tag{19}$$

In addition, we can perform model complexity reduction in order to supply an informative representations of the motion activity while remaining parsimonious [15]. After PR estimation, we select relevant potentials of estimated model $\widetilde{\mathcal{M}}^{\mathcal{R}}$ by evaluating likelihood ratios as described in [15].

5. Segmentation based on motion activity

Given a video shot, we aim at extracting areas of interest in the representative key-frame of the shot in an automatic way. Here, meaningful entities are supposed to correspond to areas comprising pertinent motion activity. To this end, we exploit the statistical motion activity modeling introduced in the previous section. A prominent advantage of such an approach is to provide within the same framework the extraction and the characterization of particular areas which will then be used to perform video retrieval with partial query.

In the sequel, we assume that a primary partition of the image is available. In practice, we consider a block-based partition (it could also be a texture-based or color-based segmentation). The goal is to build meaningful clusters from this initial set of blocks w.r.t. motion content. Hence, we first define a similarity measure related to motion activity (subsection 5.1). Secondly, this similarity measure is used to achieve the labeling of the block-based partition (subsection 5.2).

As previously defined, k_0 is the frame number of the representative key-frame of the shot. Considering three images in the shot, i.e. the one preceding the key-frame, the key-frame and the one following the key-frame, we determine the pair $\{x_{k_0}, x_{k_0+1}\}$ of maps of motion-related measurements as described in Section 3. Let us consider a partition of the image defined by the set of N_{bl} blocks $\{\mathcal{B}_i\}_{i\in\{0,\ldots,N_{bl}\}}$. We further assume that motion activity within each block $\mathcal{B}_i$ is characterized by means of the associated statistical model $\mathcal{M}^{\mathcal{B}_i}$ issued from the cooccurrence distribution $\Gamma^{\mathcal{B}_i}$ as explained in Section 4. Let us stress that we perform this estimation using the PR estimator (relation (19)).

5.1. Statistical similarity measure related to motion activity

Considering two regions $\mathcal{R}$ and $\mathcal{R}'$, our goal is to define a measure of content similarity between $\mathcal{R}$ and $\mathcal{R}'$ w.r.t. to motion activity. Let us note $\mathcal{M}^{\mathcal{R}}$ and $\mathcal{M}^{\mathcal{R}'}$ the statistical models of motion activity attached respectively to regions $\mathcal{R}$ and $\mathcal{R}'$ for the sequence of motion-related measurements resp. $x^{\mathcal{R}} = \{x^{\mathcal{R}}_{k_0}, x^{\mathcal{R}}_{k_0+1}\}$, $x^{\mathcal{R}'} = \{x^{\mathcal{R}'}_{k_0}, x^{\mathcal{R}'}_{k_0+1}\}$ and for the associated cooccurrence distributions resp. $\Gamma^{\mathcal{R}}$, $\Gamma^{\mathcal{R}'}$.

We have built a similarity measure relying on an approximation of the Kullback-Leibler (KL) divergence [3], which evaluates the similarity of two statistical distributions as the expectation of the log-ratio of these distributions. More precisely, the considered similarity measure $D(\mathcal{M}^{\mathcal{R}}, \mathcal{M}^{\mathcal{R}'})$ is a symmetrical version of the KL divergence:

$$D(\mathcal{M}^{\mathcal{R}}, \mathcal{M}^{\mathcal{R}'}) = \frac{\left[KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'}) + KL(\mathcal{M}^{\mathcal{R}'}\|\mathcal{M}^{\mathcal{R}})\right]}{2} \tag{20}$$

where $KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'})$ is the KL divergence. Based on a Monte-Carlo strategy, the latter quantity is approximated as [15]:

$$KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'}) \approx \frac{1}{|\mathcal{R}|} \cdot \ln\left(\frac{P_{\mathcal{M}^{\mathcal{R}}}(x^{\mathcal{R}})}{P_{\mathcal{M}^{\mathcal{R}'}}(x^{\mathcal{R}})}\right) \tag{21}$$

Using the exponential expression of the laws $P_{\mathcal{M}^{\mathcal{R}}}$ (relation 11), $KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'})$ can be rewritten as:

$$KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'}) \approx \frac{\left[\Psi_{\mathcal{M}^{\mathcal{R}}} \bullet \Gamma^{\mathcal{R}} - \Psi_{\mathcal{M}^{\mathcal{R}'}} \bullet \Gamma^{\mathcal{R}}\right]}{|\mathcal{R}|} \tag{22}$$

Since $\mathcal{M}^{\mathcal{R}}$ is the ML estimate associated to the cooccurrence distribution $\Gamma^{\mathcal{R}}$, $KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'})$ is positive and equals 0 if the two statistical distributions are identical. In fact, this ratio quantifies the loss of information occurring when considering $\mathcal{M}^{\mathcal{R}'}$ instead of $\mathcal{M}^{\mathcal{R}}$ to describe motion activity within area $\mathcal{R}$.

5.2. Region-level graph labeling

We now present the labeling scheme of the block-based partition of the image, $\{\mathcal{B}_i\}_{i\in\{0,\ldots,N_{bl}\}}$, which will be exploited in the next section for image segmentation. It relies on a Markovian region-level labeling framework [14, 17] applied to the adjacency graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is the set of nodes of graph $\mathcal{G}$ and $\mathcal{A}$ the set of its arcs. Each node $n_i \in \mathcal{N}$ holds for block $\mathcal{B}_i$ with $i \in \{0, \ldots, N_{bl}\}$, and $\mathcal{A}$ represents the set of arcs between graph nodes corresponding to connected blocks (in practice, we consider a four-connectivity neighborhood). Over this graph structure $\mathcal{G}$, we define a region-level Markov random field model the sites of which are the nodes of graph $\mathcal{G}$. A two-site clique neighborhood system is deduced from the set of arcs $\mathcal{A}$.

Let us assume that a set of labels $\mathcal{L}$ relative to different motion activity models has been specified (it will be defined in the next section). We further consider that to each label $l \in \mathcal{L}$ is attached a motion activity model $\mathcal{M}^l$ and the associated cooccurrence distribution Γ^l. Let us note $e = \{e_{n_i}\}_{i \in \{0,\ldots,N_{bl}\}}$ the label field with e_{n_i} taking value in $\mathcal{L}$, and $o = \{o_{n_i}\}_{i \in \{0,\ldots,N_{bl}\}}$ the observation field. In our case, at each n_i, o_{n_i} refers to the motion activity characterization attached to the block $\mathcal{B}_i$, i.e. both model $\mathcal{M}^{\mathcal{B}_i}$ and cooccurrence distribution $\Gamma^{\mathcal{B}_i}$. Adopting the Maximum A Posteriori (MAP) criterion and using the equivalence between Markov and Gibbs random fields [18], the labeling scheme comes to solve for:

$$\hat{e} = \arg \min_{e \in \mathcal{L}^{N_{bl}}} U(e,o) \tag{23}$$

where $U(e,o) = U^a(e,o) + U^b(e)$, with U^a the data-driven energy term, and U^b the regularization term. Both energy terms are split in the sum of local potentials V^a and V^b:

$$\left\{ \begin{array}{l} U^a(e,o) = \displaystyle\sum_{n \in \mathcal{N}} V^a(e_n, o_n) \\ \\ U^b(e) = \displaystyle\sum_{(n_i,n_j) \in \mathcal{A}} V^b(e_{n_i}, e_{n_j}) \end{array} \right. \tag{24}$$

The regularization potential V^b tends to favor identical labels for neighboring nodes:

$$V^b(e_{n_i}, e_{n_j}) = \beta \cdot \delta(e_{n_i} - e_{n_j}) \tag{25}$$

with β a parameter tuning the importance of the regularization and δ the Kronecker function. Besides, at node n_i, the data-driven potential $V^a(e_{n_i}, o_{n_i})$ quantifies how relevant is the description of observation o_{n_i} by label e_{n_i} according to motion activity. It involves the similarity measure D, defined by the relation (22), as given by:

$$V^a(e_{n_i}, o_{n_i}) = \exp\left[-D\left(\mathcal{M}^{e_{n_i}}, \mathcal{M}^{\mathcal{B}_i}\right)\right] \tag{26}$$

We introduce an exponential form to get values of potential V^a within the range $[0,1]$, which enables to set more easily the regularization parameter β.

Finally, the minimization issue (23) is tackled using the HCF (Highest Confidence First) algorithm [8] since the number of nodes of the considered graph is relatively small.

5.3. Separation of entities of interest from static background

We want to come to separate the static background from entities of interest. Since we consider warped image sequences using the dominant image motion assumed to be due to the camera motion, this issue reduces to extract regions which do not conform to the dominant image motion [14, 28].

In a first step, we determine a binary labeling of the initial partition of the image in terms of blocks conforming or not to the dominant image motion. As the image segmentation proceeds from motion activity characterization, we have to establish a model corresponding to the static background. Even if we could a priori infer an activity model according to null motion measurements, we prefer to explicitly estimate this model from actual motion quantity distribution at points attached to the static extracted background since camera motion cannot be perfectly cancelled. To achieve this, we exploit a by-product of the robust multiresolution estimation of the affine motion model accounting for the dominant image motion (see Section 3). More precisely, the minimization of criterion (5) is solved using an IRLS (Iterated Re-weighted Least Squares) technique, leading to:

$$\begin{array}{c} \widehat{\Delta\Theta}^{k+1} = \arg \displaystyle\min_{\Delta\Theta^{k+1}} \sum_p \frac{1}{2} \omega_p r_p^2 \\ \text{with } \omega_p = \dfrac{\psi(r_p)}{r_p} \end{array} \tag{27}$$

where ψ is defined in relation (4) and r_p is given by expression (6) where $\Delta\Theta^k$ is given by $\widehat{\Delta\Theta}^k$. At the final step of the estimation of the dominant motion, the weight value ω_p indicates if the point p is likely or not to belong to the part of the image undergoing the dominant image motion. By definition, ω_p belongs to the interval $[0,1]$. In addition, the closer ω_p to 1, the more the point p conforms to the dominant image motion.

We can deduce by thresholding map ω a rough support associated to the dominant image motion. Points p satisfying $\omega_p > \mu$ are stated as belonging to support $\mathcal{S}_d$ associated to the dominant image motion, and they then form the static background. Using our statistical motion activity modeling framework, we can estimate model $\mathcal{M}^{\mathcal{S}_d}$ attached to $\mathcal{S}_d$ using the PR window estimator (relation (19)). If $\overline{\mathcal{S}_d}$ designates the complementary of $\mathcal{S}_d$ (corresponding to the outlier map), we can evaluate in the same way the associated motion activity model $\mathcal{M}^{\overline{\mathcal{S}}_d}$.

At this stage, we achieve a Markovian block-based labeling as described in Subsection 5.2 while considering only two labels referring to statistical models attached to regions $\mathcal{S}_d$ and $\overline{\mathcal{S}_d}$ (i.e. label set $\mathcal{L}$ contains only two labels in that case). The obtained binary segmentation supplies a new estimate of the support of regions $\mathcal{S}_d$ and $\overline{\mathcal{S}_d}$, and their associated models $\mathcal{M}^{\mathcal{S}_d}$ and $\mathcal{M}^{\overline{\mathcal{S}}_d}$ can be updated. Since $\overline{\mathcal{S}_d}$ includes the areas likely to be of interest for video indexing, we then extract its connected components. Let us denote by $\{\mathcal{R}_i\}_{i\in\{1,\ldots,N_{reg}\}}$ the N_{reg} resulting regions. For each region $\mathcal{R}_i$, we perform the estimation of its activity model $\mathcal{M}^{\mathcal{R}_i}$. We then perform a second region-level labeling step applied to the original block-based partition as explained in Subsection 5.2 with $|\mathcal{L}| = N_{reg} + 1$. We consider $N_{reg} + 1$ different labels $\{l_0, .., l_{N_{reg}}\}$ corresponding to the updated model $\mathcal{M}^{\mathcal{S}_d}$ (label l_0) and to models $\{\mathcal{M}^{\mathcal{R}_i}\}_{i\in\{1,\ldots,N_{reg}\}}$. Once convergence is reached, regions $\mathcal{Q}_i$ formed by pixels with labels $l_i \neq l_0$ are regarded as the entities of interest for the processed video shot. Besides, for each region $\mathcal{Q}_i$, we store its associated model $\mathcal{M}^{\mathcal{Q}_i}$ as a descriptor of its dynamic content which will be used in the retrieval stage.

6. Retrieval with partial query

6.1. *Partial query*

We can now tackle retrieval with partial query by example. Considering a set of video documents, we first construct an index base composed of the set of entities and their associated activity descriptors extracted from all the stored video samples. This is achieved by segmenting each video into shots [5] and by applying the motion activity segmentation scheme described in Section 5. This provides us automatically and simultaneously with the extraction of meaningful entities and the associate characterization in terms of motion activity. The addition of an entity to the base is manually validated in order to reject areas which are not relevant for indexing purpose such as logos or score captions.

Otherwise, once a video query is submitted by the user, we extract automatically from the submitted video sample local relevant entities and the user specifies which one is of interest to perform the retrieval of similar examples from the indexed video base.

6.2. *Bayesian retrieval*

Considering a database $\mathcal{D}$ comprising a set of extracted entities with their non-parametric statistical motion characterization, the retrieval process is formulated as a Bayesian inference issue based on the MAP criterion [15, 30]. Given a video query q and a region $\mathcal{R}_q$ as partial query, we aim at delivering to the user examples similar to q w.r.t. dynamic content. The best match d^* is given by:

$$\begin{aligned} d^* &= \arg\max_{d\in\mathcal{D}} P(d|q) \\ &= \arg\max_{d\in\mathcal{D}} P(q|d)P(d) \end{aligned} \tag{28}$$

The prior distribution $P(d)$ allows us to express *a priori* knowledge on the video content relevance over the database. It could be inferred from semantical description attached to each type of video sequences or from relevance feedback by interacting with the user in the retrieval process [25]. In the current implementation of our retrieval scheme, we will set no a priori ($P(d)$ distribution is uniform). A statistical model of motion activity $\mathcal{M}^d$ is attached to each entity d of the database. Furthermore, a cooccurrence distribution $\Gamma^{\mathcal{R}_q}$ relative to motion-related measurements $x^{\mathcal{R}_q}$ are also attached to $\mathcal{R}_q$. Hence, the conditional likelihood $P(q|d)$ is formally expressed as $P_{\mathcal{M}^d}(x^{\mathcal{R}_q})$ and we get:

$$d^* = \arg\max_{d\in\mathcal{D}} P_{\mathcal{M}^d}(x^{\mathcal{R}_q}) \tag{29}$$

From the exponential expression of the law $P_{\mathcal{M}^d}$ (relation (11)), we further deduce:

$$d^* = \arg\max_{d\in\mathcal{D}} \left[\Psi_{\mathcal{M}^d} \bullet \Gamma^{\mathcal{R}_q}\right] \tag{30}$$

In addition, the computation of the conditional likelihoods $P_{\mathcal{M}^d}(x^{\mathcal{R}_q})$ supplies us with a ranking of the elements d of the base $\mathcal{D}$ since we can evaluate how the different statistical models $\mathcal{M}^d$ fits to the motion-related measurements computed in the query area $\mathcal{R}_q$. In practice, we provide the user with a given number of ranked replies.

7. Results

7.1. Extraction of entities of interest

We have first carried out experiments for the extraction of entities of interest based on motion activity. All experiments have been performed with the values of parameters involved in the motion activity segmentation scheme given in Table 1.

In Figure 2, we show partial corresponding to the different steps of the motion activity segmentation scheme. Figure 2.a contains to the key-frame of the processed shot, Figure 2.b the binary support of the dominant image motion, Figure 2.c the map of quantized motion-related measurements and Figure 2.d the extracted regions of interest w.r.t. motion activity. This scheme enabling simultaneous extraction and characterization of entities of interest within key-frames of video shots also appears efficient in terms of computational time. It indeed requires about 0.2 second of CPU time to process images of size 120×160 comprising 20 blocks of size 32×32 blocks on a Sun Creator workstation 360MHZ.

Concerning the parameter setting (given in Table 1), it appears that the selected values of the constant η and of the regularization coefficient β have a rather weak influence on the obtained results. Similarly, threshold μ can be set in the range $[0.1, 0.5]$ without major changes in the segmentation results. The remaining parameters, i.e., block size, number of levels and bounds of quantization of motion-related measurements, are closely related. Let us denote by N the number of quantization levels and B the size of blocks. Indeed, B has to be set in accordance to N. More precisely, we need to estimate N^2 potentials for motion activity models within blocks using $2B^2$ samples. Therefore, if we consider 32×32 blocks, the number of quantization levels should not be greater than 32. Setting N to 16 is a reasonable trade-off between accuracy of the representation of motion activity and computation complexity . Besides, for image sizes comprised between 120×160 and 352×288 pixels as in our video base, the use of 32×32 blocks proves sufficient enough to locate entities of interest in the scene. The use of the PR estimator should indeed help dealing with rather small blocks. Therefore, the choice of σ value obviously depends both on N and B. For large values of B and N, the PR regularization is not necessary, which leads to low value for σ^2 (lower than to 0.2). For 32×32 blocks and 16 quantization levels, it turns out that σ^2 should take a value in the range $[0.2, 1.0]$.

The considered video set involves different kinds of sport videos. Two main classes of shots can be distinguished: the first one consists of close-up of a particular area of the play field, and the second one displays global views of the scene. In the first case, the entities of interest are obviously the tracked players, whereas in the second case this is no more a single player but

Table 1. Parameter setting or selected options for our motion activity segmentation scheme

Motion-related measurements	quantization within $[0, 4]$ on 16 levels, $\eta = 5$
Support of the dominant image motion	$\mu = 0.2$
Motion activity model estimation	PR estimator using a Gaussian kernel with variance $\sigma^2 = 0.25$
Markovian region-level labeling	32×32 blocks (i.e 20 blocks for images of size 120×160) $\beta = 0.5$

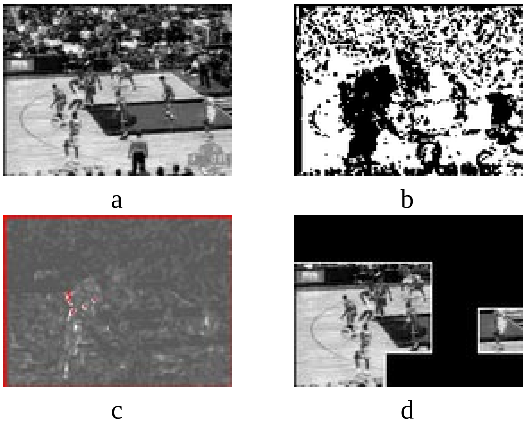


Fig. 2. Results of motion activity segmentation: (a) selected key-frame for the processed shot, (b) support of the dominant image motion in white ($\mu = 0.2$), (c) map of motion-related measurements quantized within $[0, 4]$ on 16 levels (visualized with grey-levels within $[0, 256]$), (d) result of the motion activity segmentation using the PR estimator with $\sigma = 0.5$, $\beta = 0.5$ and 32×32 blocks. The black area holds for the region regarded as the static background and the two areas delimited by a white line are the extracted entities of interest.

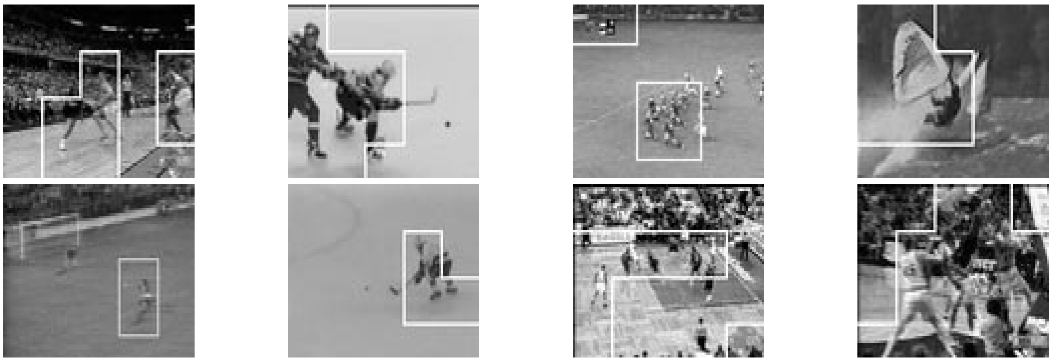


Fig. 3. Examples of block-based segmentation of entities of interest within shot key-frames according to motion activity. The extracted areas are delimited in white. For these different examples, the parameters used are those given in Table 1.

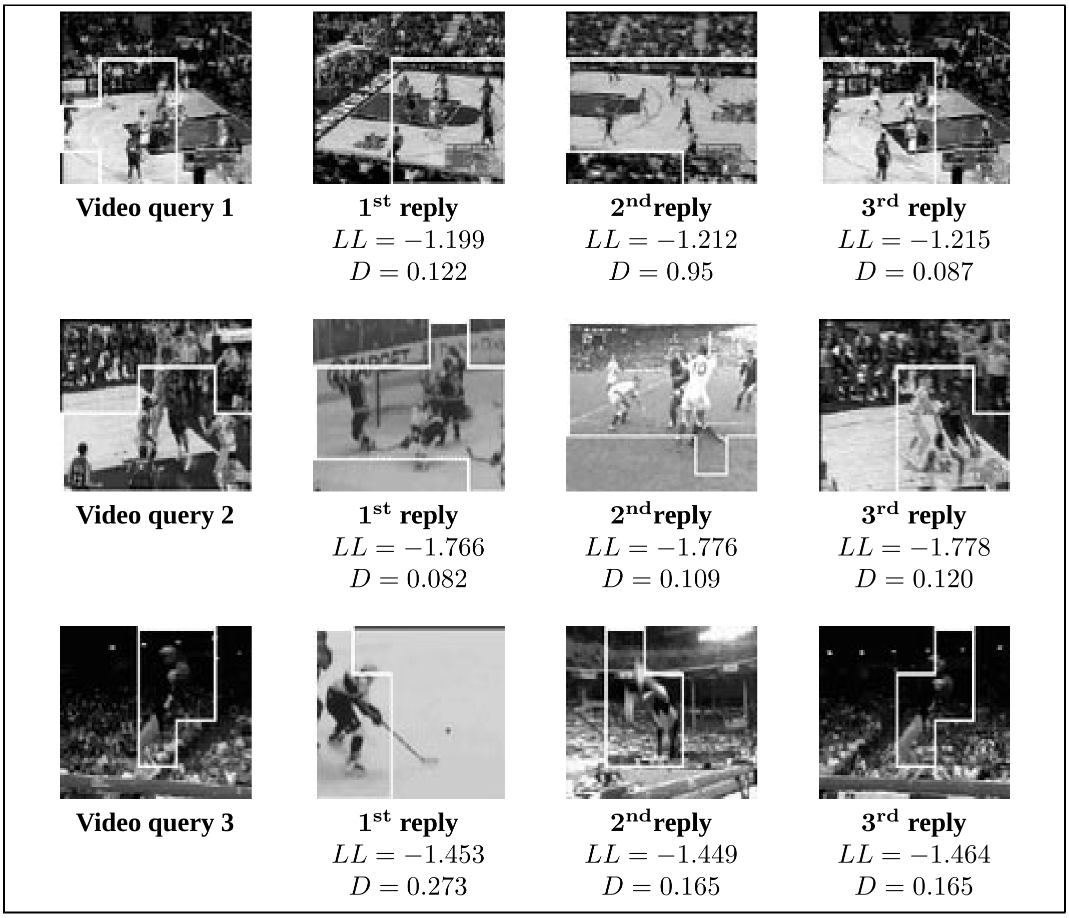


Fig. 4. Examples of retrieval with partial query. We give for each reply d the value LL of the log-likelihood $\ln\left(P_{\mathcal{M}^d}(x^{\mathcal{R}_q})\right)$ corresponding to partial video query $\mathcal{R}_q$. To a posteriori evaluate the relevance of the replies, we have also estimated model $\mathcal{M}^{\mathcal{R}_q}$ for the query and we report the distances $D(\mathcal{M}^{\mathcal{R}_q}, \mathcal{M}^d)$ between $\mathcal{M}^{\mathcal{R}_q}$ and the different retrieved models $\mathcal{M}^d$.

rather a group of players or a particular area of the play field. We display in Figure 3 some examples of entities extracted w.r.t motion activity in shots of the processed video set.

7.2. Retrieval operations with partial query

We have conducted several experiments of retrieval operations with partial query. We have considered a set of one hundred video shots involving different dynamic contents. They are color image sequences downloaded from different web sites in a compressed format (MPEG, AVI, MOV). In practice, we process uncompressed grey-level sequences of about ten to twenty images. We have focused on sport shots such as rugby, football, basketball and hockey. In Figure 4, we report three examples of retrieval operations. The three best replies are given for each query. For all the processed examples, the system delivers relevant replies in terms of motion properties. To appreciate the relevance of the replies, we give for each reply d the value of the conditional likelihood $P_{\mathcal{M}_d}(x_q^{\mathcal{R}})$. To further quantify the similarity between the retrieved entities of interest and the query, we have also determined the value of the similarity measure D computed between the statistical motion activity models estimated within the query area and those attached to the retrieved ones Let us recall that we do not need to estimate the activity model corresponding to the query in the retrieval phase; this is just performed here to enable a complementary quantitative evaluation of the results.

To evaluate the statistical diversity of the motion activity contents available in the processed base of entities of interest, we cannot supply any direct 2D absolute mapping of the elements of the base since our approach only allows to compare pairs of motion models through the similarity measure D. However, we select one "reference" model to map the base onto a 1D axis w.r.t. this model and supply such a 1D mapping for several "reference" models. We have then considered the three partial queries reported in Fig.4 as reference models, and we have computed the distances $\{D(\mathcal{M}^{\mathcal{R}_q}, \mathcal{M}^d)\}_{d \in \mathcal{D}}$ between the motion model of the partial query $\mathcal{R}_q$ and those of the elements d of the base $\mathcal{D}$. To compute distance histograms, we need to quantize the distances $\{D(\mathcal{M}^{\mathcal{R}_q}, \mathcal{M}^d)\}_{d \in \mathcal{D}}$. In fact, we have used the set of values $\{1.0 - \exp D(\mathcal{M}^{\mathcal{R}_q}, \mathcal{M}^d)\}_{d \in \mathcal{D}}$ since these quantities are comprised within $[0, 1]$. Fig.5 depicts histograms of these distances resulting from a linear quantization within $[0, 1]$ using 25 levels. Plots 5.(a), 5.(b) , 5.(c) resp. refer to the distance histograms computed w.r.t. resp. the video query 1, 2 and 3 of Fig.4. These histograms, in particular plots 5.(a) and 5.(b), show that the distribution of motion activity contents within the processed base is widespread. Besides, from these 1D mappings of the processed base, it can be envisaged that our motion modeling framework may also be exploited for motion classification since different modes seem to be discriminated in the first two plotted histograms.

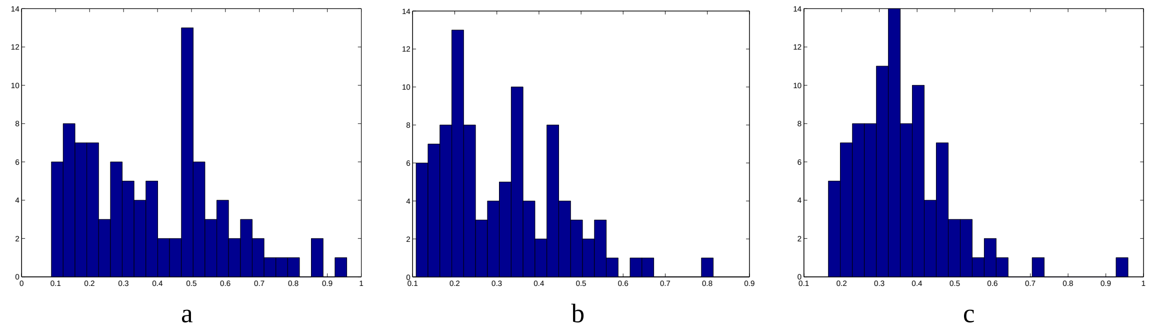
8. Conclusion

We have presented in this paper an original approach for motion-based video retrieval able to handle partial query. It relies on an automatic and efficient extraction of entities of interest within key-frames of each video shot, which results from motion activity characterization. Motion information is expressed as non parametric statistical models which account for a large range of dynamic scene content. This statistical framework can then be straightforwardly exploited to perform the retrieval with partial query, as validated by representative real experiments.

In future work, we plan to evaluate our approach on a larger video base. An important issue will be to define a procedure to evaluate in a quantitative way our retrieval scheme (for instance, in terms of correctly classified or misclassified scenes). In particular, it will require to define a preliminary categorization of the video base, which is not so easy. Besides, we could also address the tracking of the extracted entities of interest in video shots. Then, the trajectory of the tracked entities could be also exploited for retrieval purpose. To still benefit from a Bayesian framework, this would require to propose an appropriate statistical framework to combine trajectory and motion activity characterizations.

References

1. D. A. Adjeroh, M. C. Lee, and I. King. A distance measure for video sequences. *Computer Vision and Image Understanding,* 75(1-2):25–45, August 1999.
2. P. Aigrain, H-J. Zhang, and D. Petkovic. Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia Tools and Applications,* 3(3):179–202, September 1996.
3. M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing,* 18(4):349–369, 1989.
4. A. Del Bimbo, E. Vicario, and D. Zingoni. Symbolic description and visual querying of image sequences using spatio-temporal logic. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 7(4):609–621, 1997.



Qualitative evaluation of the statistical diversity of the base of entities of interest]Qualitative evaluation of the statistical diversity of the base of entities of interest w.r.t. motion activity content : histograms of the similarity measures of motion activity computed between a given partial query $\mathcal{R}_q$ and each element d of the database. In order to quantize the similarity measures $D(\mathcal{M}^{\mathcal{R}_q}, \mathcal{M}^d)$, we consider the values $1.0 - expD(\mathcal{M}^{\mathcal{R}_q}, \mathcal{M}^d)$ which are within $[0, 1]$ and a linear quantization over $[0, 1]$ with 25 bins. We report three examples relative to every partial query introduced in Fig. 4 : 1D mapping w.r.t. (a) video query 1, (b) video query 2, (c) video query 3.

Fig. 5. [

5. P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, 9(7):1030–1044, 1999.
6. R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, 1999.
7. S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong. VideoQ- an Automatic content-based video search system using visual cues. In *Proc. ACM Multimedia Conf.*, pages 313–324, Seattle, November 1997.
8. P.B. Chou and C.M. Brown. The theory and practice of Bayesian image modeling. *International Journal of Computer Vision*, 4(3):185–210, 1990.
9. J.D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–625, April 1997.
10. S. Dagtas, W. Al-Khatib, A. Ghafoor, and R.L. Kashyap. Models for motion-based video indexing and retrieval. *IEEE Trans. on Image Processing*, 9(1):88–101, 2000.
11. Y. Deng and B.S. Manjunath. Content-based search of video using color, texture and motion. In *Proc. of 4th IEEE Int. Conf. on Image Processing, ICIP'97*, pages 543–547, Santa-Barbara, October 1997.
12. R. Fablet and P. Bouthemy. Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval. In *Proc. of 3rd Int. Conf. on Visual Information Systems, VISUAL'99*, LNCS Vol 1614, pages 221–228, Amsterdam, June 1999. Springer.
13. R. Fablet and P. Bouthemy. Statistical motion-based object indexing using optic flow field. In *Proc. of 15th Int. Conf. on Pattern Recognition, ICPR'2000*, volume 4, pages 287–290, Barcelona, September 2000.
14. R. Fablet, P. Bouthemy, and M. Gelgon. Moving object detection in color image sequences using region-level graph labeling. In *Proc. of 6th IEEE Int. Conf. on Image Processing, ICIP'99*, pages 939–943, Kobe, October 1999.
15. R. Fablet, P. Bouthemy, and P. Pérez. Non parametric statistical analysis of scene activity for motion-based video indexing and retrieval. Technical Report 4005, INRIA, September 2000.
16. M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. of 5th Eur. Conf. on Computer Vision, ECCV'98*, LNCS Vol 1406, pages 595–609, Freiburg, June 1998. Springer.
17. M. Gelgon and P. Bouthemy. A region-level motion-based graph representation and labeling for tracking a spatial image region. *Pattern Recognition*, 33(4):725–745, 2000.
18. S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.
19. G.L. Gimel'Farb. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(11):1110–1114, 1996.
20. M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. of 2nd Eur. Conf. on Computer Vision, ECCV'92*, pages 282–287, Santa Margherita, May 1992.
21. A.K. Jain, A. Vailaya, and W. Xiong. Query by video clip. *Multimedia Systems*, 7(5):369–384, 1999.
22. A. Mitiche and P. Bouthemy. Computation and analysis of image motion: a synopsis of current problems and methods. *International Journal of Computer Vision*, 19(1):29–55, 1996.
23. R. Mohan. Video sequence matching. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'98*, pages 3697–3700, Seattle, May 1998.
24. M.R. Naphade, T.T. Kristjansson, B.J. Frey, and T. Huang. Probabilistic multimedia objects (Multijects) : a novel approach to video indexing and retrieval in multimedia systems. In *Proc. of 5th IEEE Int. Conf. on Image Processing, ICIP'98*, pages 536–5450, Chicago, October 1998.

25. C. Nastar, M. Mitschke, and C. Meilhac. Efficient query refinement for image retrieval. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'98*, pages 547–552, Santa Barbara, June 1998.
26. R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *Computer Vision, Graphics, and Image Processing*, 56(1):78–99, 1992.
27. J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, 1995.
28. J.M. Odobez and P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In *Video Data Compression for Multimedia Computing*, chapter 8, pages 295–311. H. H. Li, S. Sun, and H. Derin, eds, Kluwer, 1997.
29. E. Parzen. On estimation of probability density function and mode. *Annals Math. Statist.*, 33:1065–1076, 1962.
30. N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'2000*, pages 216–221, Hilton Head, June 2000.
31. N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, 2000.
32. V. Vinod. Activity based video shot retrieval and ranking. In *Proc. of 14th Int. Conf. on Pattern Recognition, ICPR'98*, pages 682–684, Brisbane, August 1998.
33. H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent access to digital video: The informedia project. *IEEE Computer*, 29(5):46–52, 1996.
34. S.C. Zhu, T. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

Ronan Fablet graduated from the Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (SUPAERO), France, in 1997. He has been preparing a Phd degree in Signal Processing and Telecommunications from the University of Rennes, France, since October 1998. His main research interests are statistical modeling, motion analysis, video indexing and retrieval.

Patrick Bouthemy graduated from Ecole Nationale Supérieure des Télécommunications, Paris, in 1980, and received the Ph.D degree in Computer Science from the University of Rennes, France, in 1982. From December 1982 until February 1984, he was employed by INRS-Telecommunications, Montreal, P.Q., Canada, in the Department of Visual Communications. Since April 1984, he has been with INRIA, at IRISA in Rennes. He is currently "Directeur de Recherche" Inria and head of Vista project. His major research interests are concerned with image sequence processing and 2D motion analysis based on statistical approaches (MRF models, robust estimation, Bayesian estimation), fluid motion analysis, tracking, dynamic scene interpretation, video indexing. He is Associate Editor of the IEEE Transactions on Image Processing.