

# Statistical motion-based retrieval with partial query

Ronan Fablet<sup>1</sup> and Patrick Bouthemy<sup>2</sup>

<sup>1</sup>IRISA / CNRS

<sup>2</sup>IRISA / INRIA

Campus universitaire de Beaulieu, 35042 Rennes Cedex, France

e-mail : rfablet@irisa.fr, bouthemy@irisa.fr

<http://www.irisa.fr/vista>

*4th Int. Conf. on Visual Information Systems, VISUAL'2000, Lyon, November 2000*

**Abstract** We present an original approach for motion-based retrieval involving partial query. More precisely, we propose an unified statistical framework both to extract entities of interest in video shots and to achieve the associated content-based characterization to be exploited for retrieval issues. These two stages rely on the characterization of scene activity in video sequences based on a non-parametric statistical modeling of motion information. Areas comprising relevant scene activity are extracted from an ascendant hierarchical classification applied to the adjacency graph of an initial block-based partition of the image. Therefore, given a video base, we are able to construct a base of samples of entities of interest characterized by their associated scene activity model. 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 video base composed of a hundred samples.

## 1 Introduction

Efficient use of video archives is of growing importance in various application fields and requires to index and retrieve visual documents based on their content. In particular, one challenging task is to retrieve samples in a video base from a query formulated by the user.

Three main schemes can be distinguished for video retrieval issues: retrieval using textual query, retrieval with query by sketch, retrieval with query by example. These three procedures for query formulation are also associated to different description schemes of visual documents from semantic characterization to low-level feature extraction. Textual query simply consists in using natural language to express a query [9, 15, 17]. It indeed requires to assign a list of key-words to each document of the video base. However, manual annotating cannot cope with the tremendous amount of video data to be analyzed, and automatic semantic characterization of videos from visual content is still beyond reach for video bases involving non specific types of content [9, 14]. For example, it remains impossible to deal with queries such as “retrieving red cars going to the right”. To introduce

more flexibility in the retrieval process, a sketch drawn by the user to express his query can be considered [4]. For instance, an arrow can be used to indicate the direction of the displacement of the object corresponding to the drawn sketch. Nevertheless, this scheme does not allow to express a wide range of queries. For instance, how to sketch a query such as “retrieve rugby game samples”? The third class of retrieval schemes consists in handling queries formulated as video samples. Query by example then relies on the computation of a set of descriptors from the video query. These descriptors are compared to the features extracted and stored for each element of the video base. The latter retrieval approach currently appears more suited to deal with the variety of dynamic contents involved in non-dedicated video bases. In addition, it can benefit from the great deal of research devoted to the definition of tools for content-based video indexing based on the extraction of numerical features [1, 3, 5–7].

As far as retrieval with query by example is concerned, the proposed approaches mainly consider global queries [5–7, 16]. They exploit a global characterization of video content, considering static (color, texture) or dynamic content (motion). However, from user point of view partial query appears more flexible since it supplies the user with the opportunity of focusing on specific entities in the scene. As a consequence, the video analysis and indexing process should automatically perform both the extraction of some meaningful entities and the associated content-based characterization.

In this paper, we propose an original contribution to this issue. A non-parametric probabilistic modeling of motion information is exploited to design an appropriate scheme for the extraction of relevant entities in video sequences in terms of scene activity, and to simultaneously supply their motion-based characterization. Then, we can design an efficient statistical framework for motion-based retrieval involving partial query. The remainder of this paper is organized as follows. Section 2 outlines the general ideas underlying our work. In Section 3, the non-parametric statistical modeling of motion information is described. Section 4 is concerned with the scheme for automatic extraction of entities of interest related to scene activity. Section 5 deals with motion-based retrieval using partial query. Experiments are reported in Section 6, and Section 7 contains concluding remarks.

## 2 Problem statement

In order to handle video retrieval with partial query, the key issue is, on the one hand, to extract entities of interest (i.e., areas), and, on the other hand, to supply their associated characterization. As far as the former point is concerned, motion information represents an important cue to extract entities of interest in video sequences. In particular, motion segmentation techniques [8] can supply the complete partition of the image into regions of homogeneous motion, mainly in terms of 2D parametric motion models. However, they remain highly computationally expensive and not reliable enough to deal with large video sets. Besides, in case of complex dynamic scenes such as articulated motions, a sin-

gle object may be divided into several different regions. Grouping them into a meaningful entity remains a difficult issue. On the other hand, motion detection techniques [8, 13] separate moving objects from the background but no further characterization of the associated areas is available.

Using the latter approaches based on 2D parametric motion models, the description of motion content attached to the extracted entities of interest mainly consists in the computation of their 2D trajectories. Even if it reveals relevant for query by sketch for instance, it may not appear general enough to handle various types of motion content such as areas including several moving objects (sport videos) or temporal texture samples (falls, crowd). Therefore, we aim at supplying a characterization of motion information expressing scene activity in a more flexible way. Scene activity will indeed be described using the non-parametric probabilistic modeling approach that we have recently introduced in [6]. This framework also provides a statistical similarity measure between video samples in terms of motion properties. The latter point makes feasible the extraction of areas comprising relevant scene activity from an ascendant hierarchical classification applied to the adjacency graph of an initial partition of the image into blocks.

To cope with retrieval using partial query, we exploit this segmentation scheme to build a base of entities of interest extracted within the key-frames of the elementary shots of the processed video set. We indeed store each extracted entity and its associated statistical scene activity model. Given a query (video sample), we similarly extract entities of interest in the proposed sample. 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 [6, 14].

### 3 Statistical modeling of scene activity

#### 3.1 Local motion-related information

To characterize motion content within a given region, our approach relies on an analysis of the distribution of some local measurements. Two kinds of local motion-related quantities can be exploited. On the one hand, one can resort to dense optic flow fields [7, 16]. In our context, their use reveals time consuming and we may cope difficult situations which reveal complex for these approaches. As a consequence, we prefer considering local motion-related quantities directly derived from the spatio-temporal derivatives of the intensity function [5, 6, 11].

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 preceding and following images in the shot onto the selected key-frame.

*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 six affine motion parameters are computed with the gradient-based incremental estimation method described in [12]. The use of a robust estimator ensures the 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.

*Local motion-related quantity:* To describe the residual motion in the compensated image sequence, the following local motion-related quantity is considered:

$$v_{obs}(p) = \left( \sum_{s \in \mathcal{F}(p)} \|\nabla I^*(s)\| \cdot |I_t^*(s)| \right) / \max \left( \eta^2, \sum_{s \in \mathcal{F}(p)} \|\nabla I^*(s)\|^2 \right) \quad (1)$$

where  $I^*(p)$  is the intensity function at point  $p$  in the warped image,  $\mathcal{F}(p)$  is a  $3 \times 3$  window centered on  $p$ ,  $\eta^2$  a predetermined constant related to the noise level in uniform areas (typically,  $\eta = 5$ ), and  $I_t^*$  is 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 local motion-related quantity was already successively used for motion detection issues [13], and for motion-based video indexing and retrieval [5, 6].

Our scene activity modeling approach relies on the evaluation of cooccurrence measurements to characterize the distribution of  $v_{obs}$  quantities. This requires to quantize the continuous variables  $v_{obs}$ . We apply a quantization on a pre-defined interval. It indeed appears relevant to introduce a limit beyond which these local measures are no more regarded as usable since gradient-based motion measurements are known to correctly handle velocity of rather small magnitude. In practice, sampling within  $[0, 4]$  on 16 levels proves accurate enough. Let  $\mathcal{A}$  be the discretized range of variations for  $v_{obs}(p)$ .

### 3.2 Temporal Gibbs modeling of scene activity

In order to characterize the scene activity in an area of interest of a key-frame, we exploit the probabilistic framework presented in [6] which relies on non-parametric scene activity models. We briefly outline this technique developed for the global characterization of motion content and specify it to the case of a given spatial area (further details can be found in [6]).

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,  $\mathcal{R}$  the region of interest in the image  $k_0$  and  $x^{\mathcal{R}} = \{x_{k_0}^{\mathcal{R}}, x_{k_0+1}^{\mathcal{R}}\}$  the restriction of  $\{x_{k_0}, x_{k_0+1}\}$  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:

$$P_{\Psi}(x^{\mathcal{R}}) = P_{\Psi}(x_{k_0}^{\mathcal{R}}) \prod_{r \in \mathcal{R}} P_{\Psi}(x_{k_0+1}^{\mathcal{R}}(r) | x_{k_0}^{\mathcal{R}}(r)) \quad (2)$$

where  $\Psi$  refers to the scene activity model associated to area  $\mathcal{R}$  and  $P_{\Psi}(x_{k_0}^{\mathcal{R}})$  to the a priori distribution of  $x_{k_0}^{\mathcal{R}}$ . We will consider in practice a uniform law. In addition,  $P_{\Psi}(x_{k_0+1}^{\mathcal{R}}(r)|x_{k_0}^{\mathcal{R}}(r))$  is expressed using an equivalent Gibbsian formulation as:

$$P_{\Psi}(x_{k_0+1}^{\mathcal{R}}(r)|x_{k_0}^{\mathcal{R}}(r)) = \exp[\Psi(x_{k_0+1}^{\mathcal{R}}(r), x_{k_0}^{\mathcal{R}}(r))] \quad (3)$$

with the normalization constraint:

$$\forall \nu' \in A, \sum_{\nu \in A} \exp[\Psi(\nu, \nu')] = 1 \quad (4)$$

This modeling framework is causal since we only evaluate temporal interactions i.e. cooccurrence of two given values at the same grid point at two successive instants. The advantages are two-fold. On one hand, it allows us to handle certain kinds of temporal non-stationarity. On the other hand, it ensures an exact computation of the conditional likelihood of a sequence of motion-related quantities w.r.t. a model. This is of key importance to achieve model estimation in an easy way and to define an appropriate measure of motion-based similarity from the Kullback-Leibler divergence. More precisely, the conditional likelihood  $P_{\Psi}(x^{\mathcal{R}})$  can be expressed according to an exponential formulation (Gibbs model):

$$P_{\Psi}(x^{\mathcal{R}}) = \exp[\Psi \bullet \Gamma^{\mathcal{R}}] \quad (5)$$

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

$$\Gamma^{\mathcal{R}}(\nu, \nu') = \sum_{r \in \mathcal{R}} \delta(\nu - x_{k_0+1}^{\mathcal{R}}(r)) \cdot \delta(\nu' - x_{k_0}^{\mathcal{R}}(r)) \quad (6)$$

where  $\delta$  is the Kronecker symbol.  $\Psi \bullet \Gamma^{\mathcal{R}}$  is the dot product between the cooccurrence distribution  $\Gamma^{\mathcal{R}}$  and the potentials specifying  $\Psi$ :

$$\Psi \bullet \Gamma^{\mathcal{R}} = \sum_{(\nu, \nu') \in A^2} \Psi(\nu, \nu') \cdot \Gamma^{\mathcal{R}}(\nu, \nu') \quad (7)$$

In fact, this modeling approach is non-parametric in two ways. First, it does not correspond to 2D parametric (affine or quadratic) motion models [12]. Second, from a statistical point of view, it does also not on parametric distributions (Gaussian) to model the law  $P_{\Psi}(\nu|\nu')$ .

Furthermore, the ML (Maximum Likelihood) estimation of the model  $\hat{\Psi}^{\mathcal{R}}$  fitting to the motion distribution attached to the region  $\mathcal{R}$  reveals straightforward. It simply comes to perform an empirical estimation of the distribution  $\{P_{\hat{\Psi}^{\mathcal{R}}}(\nu, \nu')\}_{(\nu, \nu') \in A^2}$ . Therefore, the potentials of the model  $\hat{\Psi}^{\mathcal{R}}$ , which verifies  $\hat{\Psi}^{\mathcal{R}} = \arg \max_{\Psi} P_{\Psi}(x^{\mathcal{R}})$ , are given by [6]:

$$\hat{\Psi}^{\mathcal{R}}(\nu, \nu') = \ln \left( \Gamma^{\mathcal{R}}(\nu, \nu') / \sum_{\vartheta \in A} \Gamma^{\mathcal{R}}(\vartheta, \nu') \right) \quad (8)$$

## 4 Spatial segmentation based on scene activity

Given a video shot, we aim at extracting areas of interest in its key-frame in an automatic way. Here, meaningful entities are assumed to correspond to areas comprising pertinent scene activity. Hence, we exploit the statistical scene 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 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 of the image. The goal is to build meaningful clusters from this initial set of blocks based on motion content. To this end, we have first to define an appropriate measure of content similarity (subsection 4.1). Second, we exploit this similarity measure to create a hierarchical classification from the initial set of spatial regions (subsection 4.2).

### 4.1 Statistical measure of motion-based activity similarity

We exploit again the statistical scene activity modeling framework described in Section 3. We start from the initial elementary blocks, and we progressively merge neighboring blocks. Let us note  $\mathcal{B}_i$  a current block. The point is to decide to merge it or not to another block  $\mathcal{B}_j$ . We compute the ML estimate of the scene activity model  $\Psi^{\mathcal{B}_i}$  associated to block  $\mathcal{B}_i$  using relation (8). Then, as explained in [6], we consider an approximation of the Kullback-Liebler divergence to evaluate the similarity between two probabilistic models. More precisely, for two blocks  $\mathcal{B}_i$  and  $\mathcal{B}_j$ , the similarity measure  $D(\mathcal{B}_i, \mathcal{B}_j)$  is defined by:

$$D(\mathcal{B}_i, \mathcal{B}_j) = \frac{1}{2} [KL(\mathcal{B}_i \parallel \mathcal{B}_j) + KL(\mathcal{B}_j \parallel \mathcal{B}_i)] \quad (9)$$

where  $KL$  is the Kullback-Liebler divergence approximated as (see [6]):

$$KL(\mathcal{B}_i \parallel \mathcal{B}_j) \approx \frac{1}{|\mathcal{B}_i|} \ln (P_{\Psi^{\mathcal{B}_i}}(x^{\mathcal{B}_i}) / P_{\Psi^{\mathcal{B}_j}}(x^{\mathcal{B}_i})) \quad (10)$$

where  $x^{\mathcal{B}_i}$  is the sequence of quantized motion-related measurements for block  $\mathcal{B}_i$ , and  $|\mathcal{B}_i|$  the size of the block  $\mathcal{B}_i$ . Since  $\Psi^{\mathcal{B}_i}$  is the ML estimate of the scene activity model associated to the block  $\mathcal{B}_i$ ,  $KL(\mathcal{B}_i \parallel \mathcal{B}_j)$  is positive and equals 0 if the two statistical distributions are identical. In fact, this ratio quantifies the loss of information occurring when considering  $\Psi^{\mathcal{B}_j}$  instead of  $\Psi^{\mathcal{B}_i}$  when characterizing motion information within  $\mathcal{B}_i$ . Using the exponential expression of the law  $P_{\Psi^{\mathcal{B}_i}}$  (Eq.5),  $KL(\mathcal{B}_i \parallel \mathcal{B}_j)$  is rewritten as:

$$KL(\mathcal{B}_i \parallel \mathcal{B}_j) \approx \frac{1}{|\mathcal{B}_i|} (\Psi^{\mathcal{B}_i} \bullet \Gamma^{\mathcal{B}_i} - \Psi^{\mathcal{B}_j} \bullet \Gamma^{\mathcal{B}_i}) \quad (11)$$

with  $\Gamma^{\mathcal{B}_i}$  the matrix of temporal cooccurrence values of motion-related quantities for area  $\mathcal{B}_i$  given by relation (6).

## 4.2 Hierarchical graph labeling

We need now to design a merging procedure relying on this similarity measure to extract relevant areas from the initial set of elementary blocks. Ascendant hierarchical classification is an attractive and flexible tool to supply a hierarchical representation of video sets and to facilitate the discrimination of different types of global dynamic content [6]. Our idea is to adopt this kind of approach, usually applied to classification problems, to solve the considered segmentation issue. This will allow to design a simple but efficient segmentation method able to handle quite various types of situations. Nevertheless, we must also take into account the spatial relations between the different regions of the image partition since we cope with a segmentation problem. Hence, we have developed a method exploiting the adjacency graph of these spatial regions.

It is first required to define the similarity measure  $D$  not only between elementary blocks but also between two regions  $\mathcal{R}_1$  and  $\mathcal{R}_2$ . The similarity measure is then extended as follows:

$$D(\mathcal{R}_1, \mathcal{R}_2) = \max_{(B_1, B_2) \in \mathcal{B}_{\mathcal{R}_1} \times \mathcal{B}_{\mathcal{R}_2}} D(B_1, B_2) \quad (12)$$

where  $\mathcal{B}_{\mathcal{R}_i}$  is the set of elementary blocks belonging to region  $\mathcal{R}_i$ . Using  $D$ , an ascendant hierarchical classification can be conducted as follows. As initialization, each elementary blocks of the primary image partition forms a leave in the hierarchy. At each step, we merge the two closest regions  $\mathcal{R}_1$  and  $\mathcal{R}_2$  according to  $D$  to form a new region at level  $l + 1$  of the hierarchy. Besides, a similarity constraint is introduced to prevent from merging  $\mathcal{R}_1$  and  $\mathcal{R}_2$  if they are not enough similar:

$$\text{if } D(\mathcal{R}_1, \mathcal{R}_2) > D_{max}, \mathcal{R}_1 \text{ and } \mathcal{R}_2 \text{ are not merged.} \quad (13)$$

where  $D_{max}$  is a given threshold. For two regions  $\mathcal{R}_1$  and  $\mathcal{R}_2$ ,  $\exp[-D(\mathcal{R}_1, \mathcal{R}_2)]$  is expressed as an average of two likelihood ratios comprised in  $[0, 1]$  (relation 9). As a consequence, the parameter  $D_{max}$  is set as  $D_{max} = -\ln \mu$  where  $\mu$  is a threshold in  $[0, 1]$ . The threshold  $\mu$  indeed quantifies the information loss we tolerate in terms of accuracy of description of motion distributions when substituting models attached to  $\mathcal{R}_1$  by those attached to  $\mathcal{R}_2$ , and reciprocally.

Besides, to cope with the segmentation of the image into connected regions, this hierarchical classification is performed under a connectivity constraint. It consists in merging only regions which are connected. It simply comes to substitute for the similarity measure  $D$  in the merging constraint (13) a new similarity measure  $D^*$  defined by:

$$\begin{cases} D^*(\mathcal{R}_1, \mathcal{R}_2) = D(\mathcal{R}_1, \mathcal{R}_2), & \text{if } \mathcal{R}_1 \text{ and } \mathcal{R}_2 \text{ are connected} \\ D^*(\mathcal{R}_1, \mathcal{R}_2) = D_{max}, & \text{otherwise} \end{cases} \quad (14)$$

In order to extract entities of interest, we select a level in the extracted hierarchy. In practice, we take level  $L - 1$  where level  $L$  is the root of the binary tree

since the goal is usually to segment one relevant area from the background but other choices could be easily considered if required. This results in the extraction of at least two spatial areas, i.e. the entity of interest and background.

### 4.3 Separation of entities of interest from static background

The last step of the extraction of entities of interest within the considered key-frame consists in selecting the relevant areas in the set of regions  $\{\mathcal{R}_i\}$  determined by our scene activity segmentation approach. We aim at identifying the areas that do not correspond to the static background.

To this end, we also resort to the statistical modeling framework introduced in Section 3. More precisely, dynamic contents attached both to the static background and to the set of potential regions of interest  $\{\mathcal{R}_i\}$  are described by means of a temporal Gibbsian model. Given a region  $\mathcal{R}_i$ , the associated model  $\Psi^{\mathcal{R}_i}$  is estimated using relation (8). Besides, we can infer the prior model  $\Psi_{static}$  relative to the static background denoted as  $\mathcal{R}_{static}$  since it theoretically corresponds to a distribution involving only null motion-related measurements in the warped sequence. As a consequence, the associated temporal cooccurrence histogram  $\Gamma_{static}$  is given by:

$$\forall(\nu, \nu') \in A^2, \begin{cases} \Gamma_{static}(\nu, \nu') = 0, & \text{if } (\nu, \nu') \neq (0, 0) \\ \Gamma_{static}(0, 0) = 1 \end{cases} \quad (15)$$

and the ML model  $\Psi_{static}$  is then defined as:

$$\forall(\nu, \nu') \in A^2, \begin{cases} \Psi_{static}(\nu, \nu') = -\log |A|, & \text{if } \nu' \neq 0 \\ \Psi_{static}(\nu, 0) = \log \epsilon, & \text{if } \nu \neq 0 \\ \Psi_{static}(0, 0) = \log(1 - (|A| - 1)\epsilon), & \text{otherwise} \end{cases} \quad (16)$$

where  $\epsilon$  refers to a set precision (typically,  $\epsilon = 10^{-6}$ ). It prevents from computing  $\log(0)$  and correspond to the weight given to cooccurrence configurations that never appear in the temporal cooccurrence distribution.

Then, we compute the set of similarity measures ( $D(\mathcal{R}_{static}, \mathcal{R}_i)$ ) computed between the static background  $\mathcal{R}_{static}$  and the potential regions of interest ( $\mathcal{R}_i$ ) using relation (9). The area relative to the lowest similarity measure is assumed to be associated to the static background. Therefore, for indexing purpose, we finally store all the regions ( $\mathcal{R}_i$ ) except the latter region in a base of entities of interest. Besides, for each selected region  $\mathcal{R}_i$ , we also store its associated model  $\Psi^{\mathcal{R}_i}$  as a representation of its dynamic content used in the retrieval stage. Besides, for storage issues, we can achieve a model complexity reduction by evaluating likelihood ratios as described in [6].

## 5 Retrieval with partial query

### 5.1 Partial query

We now tackle retrieval with partial query by example. The first step consists in creating an indexed base comprising a set of meaningful entities extracted from the processed video set. Thus, each video is segmented into shots [2] and we then apply the scene activity segmentation scheme described in Section 4. This framework supplies us automatically and simultaneously with the extraction of meaningful entities and the associate characterization with regard to scene activity. The introduction of an entity is at last manually validated in order to sometimes reject areas which are not relevant for indexing purpose such as logos or score captions.

On the other hand, once a video query is provided by the user, we extract automatically from the submitted video sample local relevant entities and the user specifies if one of them represents an element of interest in order to perform the retrieval of similar examples from the indexed video base.

### 5.2 Bayesian retrieval

Similarly to [6, 15], the retrieval process is formulated as a Bayesian inference issue according to the MAP criterion. Given a video query  $q$  and a region  $\mathcal{R}_q$  as partial query, the retrieval over a base of entities  $\mathcal{D}$  of samples similar to partial query  $\mathcal{R}_q$  comes to solve for:

$$d^* = \arg \max_{d \in \mathcal{D}} P(d|\mathcal{R}_q) = \arg \max_{d \in \mathcal{D}} P(\mathcal{R}_q|d)P(d) \quad (17)$$

The 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 [10]. In the current implementation of our retrieval scheme, we will set no a priori ( $P(d)$  distribution is uniform).

In our case, a statistical model of scene activity  $\Psi_d$  is attached to each entity  $d$  of the database. Furthermore, we have also determined the sequence of motion-related measurements  $x^{\mathcal{R}_q}$  for the partial query  $\mathcal{R}_q$ . Therefore, the conditional likelihood  $P(\mathcal{R}_q|d)$  is formally expressed as  $P_{\Psi_d}(x^{\mathcal{R}_q})$  and criterion (17) is given by:

$$d^* = \arg \max_{d \in \mathcal{D}} P_{\Psi_d}(x^{\mathcal{R}_q}) \quad (18)$$

From the exponential expression of the law  $P_{\Psi_d}$  (relation (5)), we further deduce:

$$d^* = \arg \max_{d \in \mathcal{D}} [\Psi^d \bullet \Gamma^{\mathcal{R}_q}] \quad (19)$$

with  $\Gamma^{\mathcal{R}_q}$  the cooccurrence distribution within  $\mathcal{R}_q$  computed using relation (6). Otherwise, the computation of the conditional likelihoods  $\{P_{\Psi_d}(x^{\mathcal{R}_q})\}_{d \in \mathcal{D}}$  also supplies us with a ranking of the element  $d$  of the base  $\mathcal{D}$  since it evaluates how the different statistical models  $\Psi^d$  fits to the motion-related measurements computed in the query area  $\mathcal{R}_q$ .

## 6 Results

### 6.1 Extraction of entities of interest

We have first carried out experiments for the extraction of entities of interest based on scene activity. We have processed different kinds of sport videos. Two main classes of shots can be distinguished: the first one involves close-up of a particular area of the play field, and the second one displays a global view of the scene. In the first case, the entities of interest are obviously the tracked players, whereas in the second case these are not a single player but rather a group of players or a particular area of the play field.

Figure 1 contains eight examples corresponding to these two different cases. In all these situations, our method extracts areas of interest which are relevant and accurate enough in the context of video indexing and retrieval.

This scheme for scene activity segmentation appears effective for motion-based indexing since it requires about 0.15 second of CPU time to process three successive  $160 \times 120$  images (for a Sun Creator workstation 360MHZ).

### 6.2 Retrieval operations with partial query

At a second stage, we have conducted retrieval operations with partial query. We have considered a set of one hundred video shots involving different dynamic contents. We have focused on sport sequences such as rugby, football, basketball and hockey. In Figure 2, we report three examples of retrieval operations. The three best replies are given for each query. For all the processed examples, the system provides 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  $P_{\psi_d}(x^{\mathcal{R}_q})$ . 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 probabilistic scene activity models estimated within the query area and those attached to the retrieved ones.

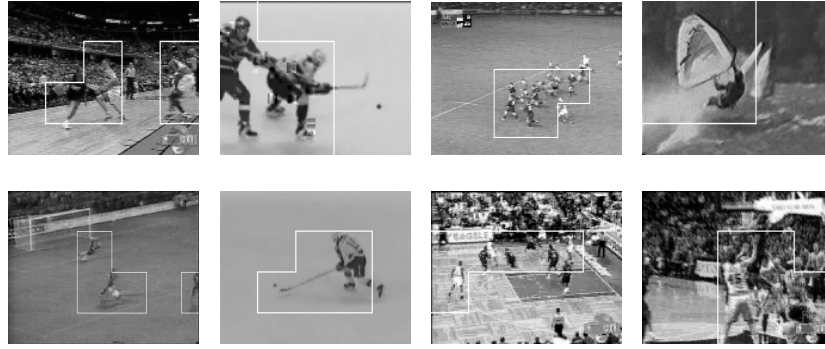
## 7 Conclusion

We have presented in this paper an original approach for motion-based video retrieval handling partial query. It relies on an automatic and efficient extraction of entities of interest from a video sequence based on scene activity characterization. Motion information is expressed as non parametric statistical scene activity models able to deal with a large range of dynamic scene content. This statistical framework can then be straightforwardly exploited to solve the retrieval process using a Bayesian framework.

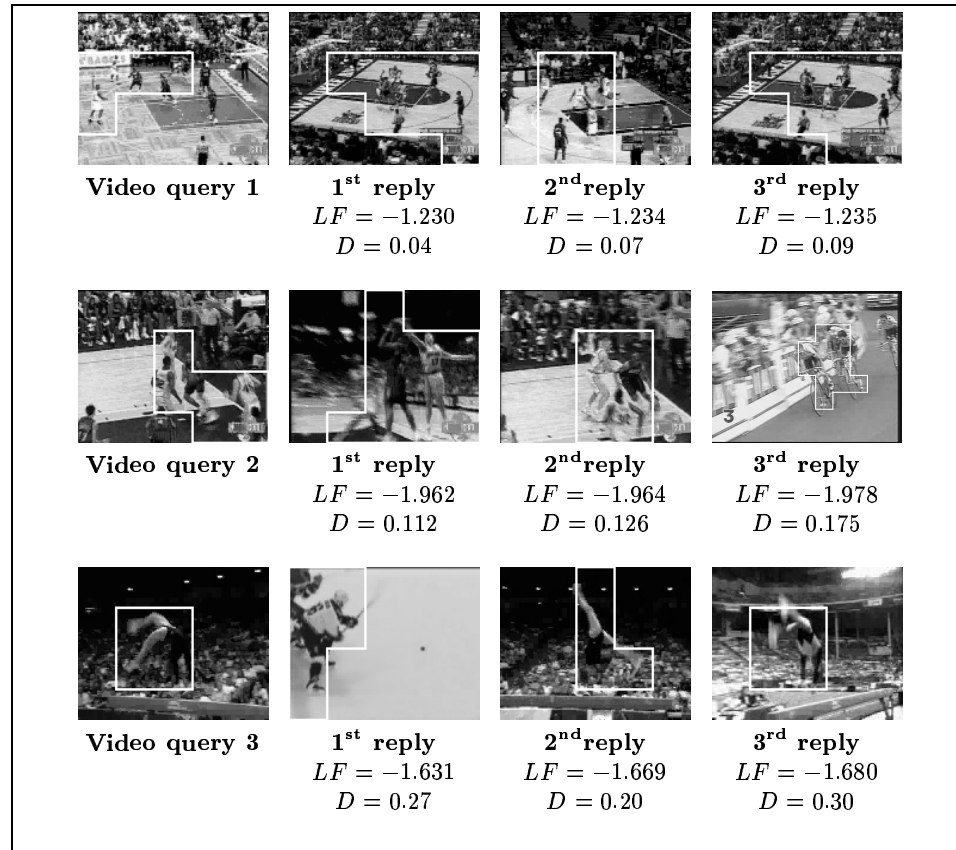
In future work, we plan to evaluate our approach on a larger video base and to address the tracking of the extracted entities of interest in video shots.

## References

1. 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.
2. 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.
3. R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *Jal of Vis. Comm. and Im. Repr.*, 10(2):78–112, 1999.
4. 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.*, Seattle, November 1997.
5. 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.
6. R. Fablet, P. Bouthemy, and P. Pérez. Statistical motion-based video indexing and retrieval. In *Proc. of 6th Int. Conf. on Content-Based Multimedia Information Access, RIAO'2000*, pages 602–619, Paris, April 2000.
7. A.K. Jain, A. Vailaya, and W. Xiong. Query by video clip. *Multimedia Systems*, 7(5):369–384, 1999.
8. A. Mitiche and P. Bouthemy. Computation and analysis of image motion: a synopsis of current problems and methods. *Int. Journal of Computer Vision*, 19(1):29–55, 1996.
9. 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.
10. 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*, Santa Barbara, June 1998.
11. R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *Computer Vision, Graphics, and Image Processing*, 56(1):78–99, July 1992.
12. J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Vis. Comm. and Im. Repr.*, 6(4):348–365, 1995.
13. 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.
14. N. Vasconcelos and A. Lippman. A Bayesian framework for semantic content characterization. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'98*, pages 566–571, Santa-Barbara, June 1998.
15. 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*, Hilton Head, June 2000.
16. 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.
17. H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent access to digital video: The informedia project. *IEEE Computer*, 29(5):46–52, 1996.



**Figure1.** Examples of segmentation based on motion-based activity. Entities of interest are delimited in white.



**Figure2.** Examples of retrieval with partial query. we give for each reply  $d$  the value  $LF$  of the log-likelihood  $\ln(P_{\Psi^d}(x^{\mathcal{R}_q}))$  corresponding to video query  $q$ . To a posteriori evaluate the relevance of the replies, we have also estimated model  $\Psi^q$  for the query and we report the distances  $D$  between  $\Psi^{\mathcal{R}_q}$  and the different retrieved models  $\Psi^d$ .