HAL
open science

Statistical motion-based video indexing and retrieval

Ronan Fablet, Patrick Bouthemy, Patrick Pérez

To cite this version:

Ronan Fablet, Patrick Bouthemy, Patrick Pérez. Statistical motion-based video indexing and retrieval. RIAO'2000 : conference on content-based multimedia information acces, Apr 2000, Paris, France. pp.602 - 619. hal-02283233

HAL Id: hal-02283233

https://hal.science/hal-02283233v1

Submitted on 10 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Statistical Motion-Based Video Indexing and Retrieval

R. Fablet[1] P. Bouthemy[2] P. Pérez[2]

[1]IRISA/CNRS [2]IRISA/INRIA
Campus universitaire de Beaulieu,
35042 Rennes Cedex, France
e-mail: {rfablet,bouthemy,perez}@irisa.fr
www: http://www.irisa.fr.vista/Vista.english.html

6th Int. on Content-Based Multimedia Information Access, RIAO'2000, Paris, April 2000

Abstract

We propose an original approach for the characterization of video dynamic content with a view to supplying new functionalities for motion-based video indexing and retrieval with query by example. We have designed a statistical framework for motion content description without any prior motion segmentation, and for motion-based video classification and retrieval. Contrary to other proposed methods, we do not extract from a given video sequence a set of motion features but we identify a global probabilistic model, expressed as a temporal Gibbs random field. This leads to define a efficient statistical motion-based similarity measure, relying on the computation of conditional likelihoods, to discriminate various motion contents. We have carried out experiments on a set of 100 video sequences, representative of various motion situations (temporal textures as fire and crowd motions, sport videos, car sequences, low motion activity examples). We have obtained promising results both for the video classification step and for the retrieval process.

1 Introduction

Archiving video information is of growing importance in various application fields. Reliable and convenient access to visual information is then of major interest for an efficient use of these databases. This implies indexing and retrieval of visual documents by their content. A great deal of research amount is currently devoted to image and video database management, [AZP96,BMM99]. Nevertheless, it remains hard to easily identify relevant information with regards to a given query, due to the complexity of dynamic scene analysis.

Another important aspect of video database management lies in the definition of appropriate similarity measures associated to the description of video content. It should provide users with efficient tools for the classification of video sequences into various types (sports, news, movies, commercials, ...) [FLE95,NKFH98,VL98], for the retrieval of examples similar to a given video query [FB99,JVX99], or for efficient video browsing using high-level structuring such as macro-segmentation [RHM99,YYL96].

Once the video has been segmented into shots [BGG99,ZWZS97], the issue is to deliver an interpretation and a representation of the shot content.[1] Most approaches rely on the selection of a set of key-frames and on their description using texture or color features [FT98]. However, motion information should not be neglected and is part of the cues to be considered in the context of video indexing, in particular for activity classification or action recognition. In this paper, we focus on this aspect and we aim at providing a statistical scheme for shot description based on dynamic content analysis and the associated measure of motion-based shot similarity.

[1]In the sequel we will also use the term of sequence to designate an elementary shot.

To cope with motion information, two main categories of approaches can be distinguished. The first one exploits segmentation, tracking and characterization of moving elements in order to determine a spatio-temporal representation of the video shot [Courtney97,FTM98,GB98]. To this end, they use either parametric motion models or dense optical flow fields. Then, the description of motion content generally relies either on the extraction of qualitative pertinent features for the entities of interest (*e.g.*, related to the direction of the displacement [GB98], or to the the trajectory of the center of gravity of the tracked objects [DAKGK00]), or on the computation of global histograms of estimated dense optical flow fields [AC97,JVX99]. However, some kinds of video cannot be handled in such a way. For instance, when considering complex dynamic contents involving motion of rivers, flames, foliages in the wind, or crowds, it reveals impossible to extract and track relevant stable local primitives. Furthermore, as far as video indexing is concerned, the entities of interest may not be single objects but rather groups of objects, in particular when dealing with sport videos. No tool currently exists to automatically extract such entities.

These remarks led to consider another category of methods providing attractive analysis of motion information in the context of content-based video indexing. Our goal is to cope with the interpretation of dynamic contents without any explicit prior motion segmentation. Primary work in that direction [NP92] results in the definition of "temporal textures" which include for instance motions of rivers, foliages, flames, or crowds. Different techniques for "temporal texture" feature extraction have been proposed [BF98,FB99,NP92,OHSF98,SP96]. In [OHSF98], descriptors are extracted from the characterization of surfaces derived from spatio-temporal trajectories. In [NP92], features issued from spatial cooccurrences of normal flows are exploited to classify sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In previous work [BF98,FB99], we have considered global features extracted from temporal cooccurrence distributions of local motion-related information more reliable than the normal velocity. In this paper, we propose to extend the latter work in order to deliver in an unified way a probabilistic modeling of dynamic content and an associated statistical scheme for motion-based video indexing and retrieval.

The paper is organized as follows. The general ideas underlying our work are presented in Section 2. Section 3 describes the local non-parametric motion-related information that we use. In Section 4, we introduce the probabilistic modeling associated to the spatio-temporal distribution of local motion-related quantities. The proposed statistical framework for motion-based indexing and retrieval is presented in Section 5. Section 6 contains classification results and retrieval examples over a set of video sequences. Concluding remarks are reported in Section 7.

2 Problem statement

As pointed out above, proposed approaches for temporal texture analysis mainly rely on the extraction of a set of numerical descriptors. As a consequence, the comparison of shot content is performed in the feature space according to a given metric such as the Euclidean distance or more appropriate measures [SJ99]. Besides, to deal with video databases involving various dynamic contents, it is necessary to determine an optimal set of features and the associated similarity measure, using either principal component analysis [MSP96] or some feature selection techniques [JZ97]. Nevertheless, feature space is usually of high dimension, and the considered distance is likely not to capture properly uncertainty attached to feature measurements.

To cope with these issues, it seems more relevant to adopt a statistical point of view which may deliver a unified view for learning and classification. We introduce a motion classification approach for video indexing which relies on a statistical analysis of the spatio-temporal distribution of local non-parametric motion-related information. We aim at identifying probabilistic models corresponding to different dynamic content types to be discriminated. We exploit a

correspondence between cooccurrence measurements and Markov Random Field (MRF) models established in the context of spatial texture analysis in [Gimelfarb96], and we propose an extension to temporal textures (see Section 4). We consider only temporal models, which allows us to easily compute the involved likelihood functions. This property leads to define a general statistical framework for video indexing and retrieval (see Section 5). In particular, we have designed a technique for hierarchical video classification based on an approximated Kullback-Liebler divergence (see subsection 5.1). The retrieval process is stated as a Bayesian inference issue conducted through the extracted video hierarchy (see subsection 5.2).

3 Motion information

The first step of our approach is to define appropriate motion-related measurements whose spatio-temporal distributions will be interpreted. Since we aim at characterizing the actual dynamic content of the scene, we need to get rid of camera motion. Consequently, we first estimate the dominant image motion between two successive images, which is assumed to be due to camera motion. Then, we cancel it by wrapping the successive images to the first image of the shot through the combination of the successive estimated elementary dominant motions.

3.1 Dominant motion estimation

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

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

with $r = (x, y)$ and $\Theta = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6]$. The computation is achieved with the gradient-based multi-resolution incremental estimation method described in [OB95]. The following minimization problem is solved :

$$\widehat{\Theta} = \arg\min_\Theta \sum_r \rho\left(DFD(r, \Theta)\right) \tag{2}$$

where $DFD(r, \Theta) = I_{t+1}(s + w_\Theta(r)) - I_t(r)$ and $\rho()$ is Tukey's biweight function. The use of a robust estimator ensures the motion estimation not to be sensitive to secondary motions due to mobile objects in the scene. Criterion (2) is minimized by means of an iterative reweighted least-square technique embedded in a multiresolution framework and involving appropriate successive linearizations of the DFD expression.

3.2 Local motion-related measurements

In order to characterize the distribution of residual motion in the motion compensated image sequence, we aim at defining some local motion-related measurements. Dense optic flow field provides such local information. Nevertheless, as stressed previously, the relevance of the estimation cannot always be guaranteed in case of complex motion situations and the required computational load remains prohibitive in the context of video indexing involving large databases. As a consequence, we prefer considering a local motion-related measure directly computed from the spatio-temporal derivatives of the intensity function in the image.

By assuming intensity constancy along motion trajectories, the well-known image motion constraint relating the 2D apparent motion and the spatio-temporal derivatives of the intensity function can be expressed as follows :

$$\mathbf{v}(r) \cdot \nabla I^*(r) + \frac{\partial I^*(r)}{\partial t} = 0 \tag{3}$$

where $\mathbf{v}$ is the 2D motion vector in the image, and $I^*(p)$ the intensity function at point p in the wrapped image. Then, we can infer the normal velocity v_n^* in the compensated sequence :

$$v_n^*(r) = \frac{-I_t^*}{\|\nabla I^*(r)\|} \tag{4}$$

with $I_t^*(r) \triangleq \dfrac{\partial I^*(r)}{\partial t}$ is the temporal derivative of the intensity function I^*. $I_t^*(r)$ is approximated by a simple finite difference. Although this expression is explicitly related to apparent motion, it can be null whatever the motion magnitude, if the motion direction is perpendicular to the spatial intensity gradient. Moreover, the normal velocity estimate is also very sensitive to noise attached to the computation of the intensity derivatives.

As for example pointed out in [OB97], the module of the spatial gradient $\|\nabla I^*(r)\|$ can represent, to a certain extent, a pertinent measure of the reliability of the computed normal velocity. Furthermore, if the spatial intensity gradient is sufficiently distributed in terms of direction in the vicinity of point r, a weighted average of v_n in an appropriate neighborhood forms a relevant motion-related quantity. More precisely, we consider the following expression :

$$v_{obs}(r) = \frac{\sum_{s \in \mathcal{F}(r)} \|\nabla I^*(s)\|^2 \cdot |v_n^*(s)(s)|}{\max\left(\eta^2, \sum_{s \in \mathcal{F}(r)} \|\nabla I^*(s)\|^2\right)} \tag{5}$$

where $\mathcal{F}(r)$ is a small window centered on r, η^2 a predetermined constant related to the noise level in uniform areas. $I_t^*(p)$ is approximated by a simple finite difference. This motion-related measurement forms a more reliable quantity than the normal flow, yet simply computed from the intensity function and its derivatives. This motion information was successfully exploited for the detection of mobile objects in compensated sequences [OB97,FBG99].

Obviously, the information relative to motion direction has been lost, which prevents from discriminating for instance two opposed translations with the same magnitude. However, this is not a real shortcoming, since we are interested in interpreting the type of dynamic situations observed in the considered video shot and not a specific motion value.

Besides, the computation of cooccurrence measurements for the spatio-temporal distribution of motion-related information ($v_{obs}(r)$) requires to quantize these continuous values. To cope with erroneous values, we apply a quantization on a predefined interval. It indeed appears relevant to introduce a limit beyond which measures are no more regarded as usable. In practice, we will use N levels of quantizations in the interval $[0, 4]$. In the sequel, we will note Λ the quantized range of variations of ($v_{obs}(r)$) and x_k these discretized motion-related measurements for the kth image of the processed video sequence.

4 Temporal Gibbsian modeling

In previous work [FB99], we exploit the measurement of temporal cooccurrences of these motion-related quantities to extract the global motion-based features. This led to encouraging results. On the other hand, a relationship between Markov random fields and cooccurrence distributions [Gimelfarb96], and more generally, with filtering-based feature extraction techniques [ZWM98], have been established. Therefore, we aim at developing a probabilistic modeling framework to characterize the spatio-temporal distributions of the local motion-related quantities attached to a shot.

Compared to the extraction of a set of numerical features for motion-based content characterization, this probabilistic modeling supplies a more accurate description of the uncertainty attached to motion information. Besides, it enables to deliver tools for video classification and retrieval embedded in a statistical framework, which should ease user interaction (relevance feedback) and the combination of different kinds of content description (motion, color, texture or shape for instance).

4.1 Temporal Gibbs random fields

We consider a purely temporal modeling approach for two main reasons. First, analyzing the temporal evolution of the local motion information allows us to handle certain kinds of non-stationarity while being sufficient to characterize the motion classes of interest. Second, it makes feasible the computation of the conditional likelihood of a sequence of motion-related quantities *w.r.t.* a model. This property leads first to a simple and straightforward model estimation (see subsection 4.2), and, secondly, to define a well-funded statistical indexing and retrieval framework (see Section 5).

More precisely, a sequence of motion quantities $(x_k)_{k=0,..,L}$ is assumed to be the realization of a sequence of random fields $X = (X_0, \ldots, X_K)$, where X is a first-order Markov chain of length $K+1$:

$$P_V(X) = P_V(X_0) \prod_{k=1}^{K} P_V(X_k|X_{k-1}) \tag{6}$$

V refers to the underlying interaction potentials to be defined later. In addition, since we consider only purely temporal interaction, we assume that conditional probabilities are expressed by:

$$P_V(X_k|X_{k-1}) = \prod_{r\in\mathcal{R}} P_V\left(X_k(r)|X_{k-1}(r)\right) \tag{7}$$

where $\mathcal{R}$ is the image grid. The latter local conditional probabilities are equivalently expressed using a Gibbs formulation:

$$P_V\left(X_k(r)|X_{k-1}(r)\right) \propto \exp\left[V\left(X_k(r), X_{k-1}(r)\right)\right] \tag{8}$$

where $V(\cdot,\cdot)$ are the prior parameters attached to the model V. Besides, it leads to the definition of the local normalization constant $Z_k(r)$ at position r in image k of the sequence:

$$Z_k(r) = \sum_{\omega\in\Lambda} \exp\left[V\left(\omega, X_{k-1}(r)\right)\right] \tag{9}$$

The likelihood $P_V(X)$ is then given by:

$$P_V(X) = P_V(X_0) \prod_{k=1}^{K} \prod_{r\in\mathcal{R}} \frac{\exp\left[V\left(X_k(r), X_{k-1}(r)\right)\right]}{Z_k(r)} \tag{10}$$

This temporal modeling can be defined equivalently as a Markov random field with temporal neighborhood structure [GG84]. In addition, the following normalization constraint is imposed in order to guarantee the uniqueness of the potentials of the temporal Gibbs random field:

$$\forall\omega' \in \Lambda, \quad \sum_{\omega\in\Lambda} \exp V_a(\omega,\omega') = 1 \tag{11}$$

which implies : $\forall(r,k) \in \mathcal{R}\times[1,K], \quad Z_k(r) = 1$. As a consequence, contrary to Markov random field in the general case, the global likelihood function $P_V(X)$ can be here simply decomposed as a product of local temporal transitions:

$$P_V(X) = P_V(X_0) \prod_{k=1}^{K} \prod_{r\in\mathcal{R}} \exp\left[V\left(X_k(r)|X_{k-1}(r)\right)\right] \tag{12}$$

For given $P_V(X_0)$ and V, we have now a complete knowledge of $P_V(\cdot)$ law. This property will be exploited to derive a statistical framework for video indexing and retrieval in Section 5. Hereafter, the law $P_V(X_0)$ is supposed to be uniform.

At last, similarly to [Gimelfarb96,ZWM98], the global likelihood function can be expressed using an exponential formulation involving a dot product $\bullet$ between cooccurrences measurements $H(X)$ and the potentials of the model V:

$$P_V(X) = P_V(X_0) \cdot \exp\Big[V \bullet H(X)\Big] \tag{13}$$

where $H(X)$ is the set of temporal cooccurrence measurements defined by:

$$\forall(\omega,\omega') \in \Lambda^2,\ H(\omega,\omega'|X) = \sum_{k=1}^{k=K} \sum_{r\in\mathcal{R}} \delta(\omega - X_k(r)) \cdot \delta(\omega' - X_{k-1}(r)) \tag{14}$$

with $\delta()$ is the Kronecker symbol, and the dot product between cooccurrences measurements $H(X)$ and the potentials of the model V is given by:

$$V \bullet H(X) = \sum_{(\omega,\omega')\in\Lambda^2} H(\omega,\omega'|X) \cdot V(\omega,\omega') \tag{15}$$

4.2 Maximum Likelihood Estimation

Given a realization x of X, we aim at identifying the model V associated to X. To this end, we consider the Maximum Likelihood (ML) criterion. Thus, we aim at solving:

$$\widehat{V} = \arg\max_V LF(V) \tag{16}$$

where V stands for $\{V(\omega,\omega'), (\omega,\omega') \in \Lambda^2\}$ and $LF(V) = \ln(P_V(x))$.

In fact, this temporal modeling consists in a product of $|\mathcal{R}|$ independent Markov chains defined by their transition matrix $exp\,[V(\cdot,\cdot)]$:

$$P_V(X_k(r)|X_{k-1}(r)) \propto exp\,[V\,(X_k(r), X_{k-1}(r))] \tag{17}$$

As a consequence, the ML estimate is readily determined from the empirical estimation of the transition probability $P_V(X_k(r)|X_{k-1}(r))$. Thus, potentials $\widehat{V}(\cdot,\cdot)$ are given by:

$$\forall(\omega,\omega') \in \Lambda^2, \quad \widehat{V}(\omega,\omega') = \ln\left(\frac{\sharp\,\{(k,r) : x_k(r) = \omega, x_{k-1}(r) = \omega'\}}{\sharp\,\{(k,r) : x_{k-1}(r) = \omega'\}}\right) \tag{18}$$

Equivalently, using cooccurrence measurements, we obtain:

$$\forall(\omega,\omega') \in \Lambda^2, \quad \widehat{V}(\omega,\omega') = \ln\left(\frac{H(\omega,\omega'|x)}{\sum_{\beta\in\Lambda} H(\beta,\omega'|x)}\right) \tag{19}$$

where the cooccurrence matrix $H(x)$ is given by Eq.(14).

4.3 Model complexity reduction

As far as video indexing is concerned, a crucial point is to supply informative representations of content description while remaining parsimonious. To this end, we perform a selection of relevant potentials after the ML estimation of the model.

In fact, as stressed by Eq.(13), informative potentials are associated to high cooccurrence values. Therefore, we perform a ranking of the estimated potentials according to the associated cooccurrences, and, we consider an iterative procedure to select the proper model complexity. Exploiting the determined ranking, potentials of $\widehat{V}$ are introduced one by one in a model $\tilde{V}$ initially set to the constant model. We use the normalization constraint of Eq.(11) to determine

the values of potentials of $\tilde{V}$ which have not been introduced yet. In order to quantify the amount of information captured by $\tilde{V}$, we evaluate the likelihood ratio corresponding to the comparison of the reduced model $\tilde{V}$ with the estimate $\widehat{V}$ and given by:

$$LR_x(\tilde{V}, \widehat{V}) = P(x|\tilde{V})/P(x|\widehat{V}) \tag{20}$$

As soon as $LR_x(\tilde{V}, \widehat{V})$ exceeds a user-specified threshold λ_{LR} accounting for the amount of information to be relevant, the procedure is stopped and the obtained reduced model $\tilde{V}$ is stored as the model attached to the sequence x.

5 Statistical motion-based indexing and retrieval

In the context of video indexing, we focus on retrieval operation. Our goal is to retrieve in a video database examples similar to a video query *w.r.t.* terms of motion content. To this end, we exploit the statistical modeling approach described above. In the same manner than in [VL99], the retrieval process is embedded in a Bayesian framework. Furthermore, since in case of large databases direct search reveals untractable, we propose a hierarchical motion-based structuring of the considered video database using a statistical similarity measure.

5.1 Statistical hierarchical indexing

We aim at determining an efficient indexing structure to make the retrieval task easier. Visual-content search trees have proven to be well-suited to still image database management as shown in [CBD00,MSP96,Schweitzer99]. It involves the construction of a binary tree whose nodes are attached to subsets of elements of the base. To achieve this hierarchical structuring, two main categories of approaches can be considered. Top-down techniques [Schweitzer99] consist in successively splitting nodes of the tree, processing from the root to the leaves. As a consequence, an element misclassified in the first steps of the procedure appears in an undesirable branch of the final binary tree. Therefore, we prefer to use bottom-up techniques which seem to offer better performance in terms of classification accuracy, [CBD00,MSP96]. In practice, we consider an ascendant hierarchical classification procedure, [DGLS81].

Given a similarity measure, this classification scheme successively forms from the leaves to the root pairs of elements minimizing the similarity measure. In addition, a threshold D_{max} is introduced to prevent from merging nodes which are too far from each other. In our case, we want to take advantage of the statistical point of view introduced above to define an appropriate similarity measure. When considering two probabilistic distributions, the Kullback-Liebler (KL) divergence appears to provide relevant characterization of similarity of these two distributions [BV98]. In practice, we will use a easily tractable approximation of the KL distance.

We first assume that a temporal Gibbsian model V^n has been estimated for each sequence n in the video database. The Kullback-Liebler (KL) divergence $KL(p, q)$ is given by:

$$KL(p\|q) = \int \ln \frac{p}{q} dp \tag{21}$$

It can be viewed as the expectation of the log-likelihood ratio $ln\left(\frac{p}{q}\right)$ *w.r.t.* the distribution p. This expectation can be approximated using a Monte-Carlo procedure. In our case, if we consider an element n of the database, the sequence of motion-related quantities x^n represents a sample associated to the distribution modeled by V^n. More precisely, at each $(k, r) \in [0, K] \times \mathcal{R}$, the transition from $(x^n_{k-1}(r))$ to $(x^n_k(r))$ is a sample of the Markov chain attached to model V^n. As a consequence, when considering two elements of the database n_1 and n_2, their associated models V^{n_1} and V^{n_2}, and the sequences of computed motion-related quantities x^{n_1} and x^{n_2}, the KL divergence $KL(n_1\|n_2)$ is approximated as the empirical average of the logarithm of the ratio of the likelihood of the transitions from $(x^n_{k-1}(r))$ to $(x^n_k(r)$ computed respectively *w.r.t.* V^{n_1}

and V^{n_2}:

$$KL(n_1 \| n_2) \approx \sum_{k=1}^{K} \sum_{r \in \mathcal{R}} \ln \left(\frac{P_{V^{n_1}}\left(x_k^{n_1}(r) | x_{k-1}(r)\right)}{P_{V^{n_2}}\left(x_k^{n_1}(r) | x_{k-1}^{n_1}(r)\right)} \right) \tag{22}$$

This leads to approximate the KL divergence $KL(n_1 \| n_2)$ by the logarithm of the ratios of the likelihoods of the sequence of motion-related quantities n^1 computed respectively for the Gibbsian models V^{n_1} and V^{n_2}:

$$KL(n_1 \| n_2) \approx \ln \left(\frac{P(x^{n_1} | V^{n_1})}{P(x^{n_1} | V^{n_2})} \right) \tag{23}$$

It indeed quantifies the loss of information occurring when considering V^{n_2} instead of V^{n_1} to model the distribution attached to n^1. Finally, in order to deal with a symmetric similarity measure, we consider the similarity measure $D(n_1, n_2)$ between elements n_1 and n_2 given by:

$$D(n_1, n_2) = \frac{1}{2} \left[KL(n_1 \| n_2) + KL(n_2 \| n_1) \right] \tag{24}$$

We need to define this similarity measure not only between elements of the video database, but also between two clusters C^1 and C^2. D is then defined by:

$$D(C^1, C^2) = \max_{(n_1, n_2) \in C^1 \times C^2} D(n_1, n_2) \tag{25}$$

Exploiting this similarity measure, we achieve an ascendant hierarchical classification as follows. At each step, a pair is formed by merging the closest clusters according to D. If a cluster C is too far from all the others, i.e. $\min_{C'} D(C, C') > D_{max}$, it is kept alone to form a single cluster. D_{max} can be easily set since D is directly related to log-likelihood ratios. As initialization, each element of the video database forms a leave of the binary tree. Besides, a model has to be attached to each created cluster. Since our temporal modeling is directly determined from temporal cooccurrence measurements, the model associated to the merging of two clusters is simply estimated using Eq.(19). More precisely, when considering a set of sequence, the cooccurrence measurements defined in Eq.(14) are the sum of the cooccurrence measurements computed for each sequence of the considered group. Therefore, when merging two clusters C_1 and C^2, we first compute the cooccurrence matrix $H(C^1, C^2)$ as the sum of the cooccurrence matrices $H(C^1)$ and $H(C^2)$, and second, exploiting Eq.(19), we estimate the potentials of the associated model.

5.2 Statistical retrieval

In case of direct search in the video database, retrieval operations can be achieved according to a MAP criterion [VL99]. Given a video query q, we want to determine the best match n^*:

$$\begin{aligned} n^* &= \arg\max_{n \in \mathcal{N}} P(n|q) \\ &= \arg\max_{n \in \mathcal{N}} P(q|n) P(n) \end{aligned} \tag{26}$$

Let us note that this criterion also supplies a ranking of the elements of the database according to $P(n|q)$. In our case, we introduce no a priori: $P(n)$ is uniform over the database elements and equal to a constant. Besides, to each element n of the base is attached a temporal Gibbsian model V^n. We compute the sequence of motion-related measures x^q for the video query q and the conditional likelihood $P(q|n)$ is expressed as $P_{V^n}(x^q)$ (see Eq.(6)). Then, we infer:

$$n^* = \arg\max_{n} P_{V^n}(x^q) \tag{27}$$

In practice, a direct search reveals time-consuming for large databases. The resolution of criterion (27) is conducted through the hierarchy determined in subsection 5.1. We cannot

ensure to find the best match, but this provides a compromise between computational time and retrieval accuracy. More precisely, we proceed as follows. We explore the binary tree from the root to the leaves. At each step s, given a parent cluster C^s, we select the best child node C^{s+1} according to the MAP criterion:

$$C^{s+1} = \arg \max_{c \in \mathcal{N}(C^s)} P_{V^c}(x^q) \tag{28}$$

where $\mathcal{N}(C^s)$ is the set of children nodes of the cluster C^s in the hierarchy. This process is run until the selected cluster contains the desired number of elements (answers) than desired or until a given similarity threshold (precision) is reached.

6 Results

We have carried out experiments on a set of real image sequences. We have paid a particular attention to choose a video set representative of various motion situations : temporal textures (flames, crowds), sport videos (basket, rugby...), rigid motion situations (cars, train, ...), and low motion activity examples. Finally, we consider a database of 100 video sequences. Typically, each sequence is composed of about 10 images.

For each element of the database, we perform the estimation of the associated temporal Gibbsian model as described in Section 4. In practice, we consider 16 levels of quantization (*i.e.* $N = 16$).The proposed scheme for model complexity reduction allows us to keep only from 10% to 20% of the 256 ML estimates of the Gibbsian potentials.

At a second stage, we exploit this temporal modeling to determine an hierarchical representation of the database which expresses similarities in terms of motion content. In the subsequent, we present two types of experiments. First, in order to provide a comprehensive visualization of the extracted binary tree, we have also performed a classification on a smaller set of 20 sequences involving various types of dynamic content. Second, we display four results of retrieval operation with query by example. In both cases, we provide a comparison with the feature-based approach described in [FB99]. Mainly, for a given sequence x, this technique extracts a set of features from the temporal cooccurrence matrix $H(x)$ and exploits the Euclidean norm in this feature-space as the similarity measure. To provide a more quantitative evaluation of the comparison, we will note D_F this feature-based similarity measure. Hereafter, we will refer to the statistical technique presented in this paper as the *model-based statistical* technique, and to the approach described in [FB99] as the *feature-based geometrical* technique.

Hierarchical motion-based classification: The set of 20 sequences used to visualize an example of hierarchal classification is depicted in Fig.1. It is composed as follows. It includes two static shots of anchors, $anchor_1$ and $anchor_2$, from news program displaying a very weak motion activity. In addition, two other examples of low motion activity, *hall* and $belle - ile$, are also in the processed video set. Otherwise, four examples of rigid motion situations are introduced referring either to road traffic sequence, $road_1$ and $road_2$, or to landing or take-off in airport sequence, $airport_1$ and $airport_2$. At last, we add twelve sport video sequences involving football game, $football_1$ and $football_2$, hockey game, $hockey_1$, $hockey_2$, $hockey_3$ and $hockey_4$, basketball game, $basketball_1$, $basketball_2$ and $basketball_3$, and windsurfing, $windsurfing_1$ and $windsurfing_2$.

For this video set, we perform the automatic unsupervised construction of the hierarchy both using the statistical technique presented in this paper and the feature-based approach described in [FB99]. The obtained classification tree are displayed respectively in Fig.2 for the *model-based statistical* technique and in Fig.2 for the *feature-based geometrical* approach. In addition, we depict for the leaves of the tree the name of the associated element of the video set and for the other nodes the value of the maximum intra-cluster similarity measure.

The hierarchy obtained using the *model-based statistical* technique discriminates correctly the different kinds of dynamic contents. Traffic sequences, $road_1$ and $road_2$, airport videos, $airport_1$ and $airport_2$, and low motion activity situations, $anchor_1$ $anchor_2$, $hall$ and $belle-ile$, constitute a separate cluster in which relevant subclusters are formed associated to these four types of motion content. In addition, all sport video are properly grouped. In this group of sequences, subgroups have also been identified such as basketball sequences, $basketball_1$, $basketball_2$ and $basketball_3$, displaying very high motion activity, and hockey samples, $hockey_1$, $hockey_2$, $hockey_3$ and $hockey_4$. The classification tree resulting from the *feature-based geometrical* scheme appears slightly less relevant. In fact, the two airport sequences, $airport_1$ and $airport_2$, are not accurately classified.

Statistical motion-based retrieval: In Fig.4 and Fig.5, we display four examples of retrieval operation involving different types of video query. The first query is expressed as a news program consisting of a static shot on an anchor. A rigid motion situations is proposed as the second query. The third and fourth retrieval operations involve sport videos. A global view of the game field is considered in the third query, whereas the fourth one is a close-up shot on a basket player tracked by the camera during the shot.

For these four examples of retrieval with query by example, we display the three best replies obtained, on on hand, using the *model-based statistical* framework, and on the other hand, the *feature-based geometrical* scheme. Besides, in both cases, we also give the values of the measure of the relevance of each reply. More precisely, for the first framework, we give the values of the log-likelihood $\ln(P_{V^n}(x^q))$ corresponding to video query q and, in order to a posteriori evaluate the relevance of the replies, we have also estimated the model V^q associated to the query and we report the distances D between V^q and the different retrieved models V^n. For the second approach, we give the values of the feature-based geometrical similarity measure D_F.

For video query one and three, the results obtained with the two different techniques are equivalent end provide relevant answers with regards to motion content. In case of video query two and four, the video replies provided by the *model-based statistical* framework are visually more relevant. In both cases, the third reply proposed using the *feature-based geometrical* scheme would be not regarded as relevant by the user.

As a consequence, from these different results which remain partial, it seems that the use of a statistical framework improves both the classification operation and the retrieval process.

7 Conclusion

We have presented an original approach for dynamic content description in video sequences with a view to coping with motion-based indexing and retrieval with query by example. The proposed scheme exploits a temporal Gibbsian modeling of the cooccurrence distributions of local motion-related measurements computed along an image sequence from the spatio-temporal derivatives of the intensity function. To remain independent from camera work and to really handle scene motion, this analysis is performed in the reconstructed sequence compensated resulting from cancelling the dominant image motion assumed to be due to camera movement. This probabilistic modeling framework results first in a statistical hierarchical motion-based structuring of the video database using an approximated version of the Kullback-Liebler distance, and, second, in a Bayesian retrieval scheme based on the MAP criterion and performed through the extracted hierarchy. We have obtained promising results both for the classification stage and retrieval operations.

In future work, we plan to evaluate our approach on a still larger video database. We could also exploit our statistical modeling approach to characterize dynamic contents attached to specific entities of interest, such as moving objects extracted in an automatic [FBG99] or semi-automatic [GBD99] way.

References

[AC97] Ardizzone (E.) et Cascia (M. La). – Automatic video database indexing and retrieval. *Multimedia Tools and Applications*, vol. 4, 1997, pp. 29–56.

[AZP96] Aigrain (P.), Zhang (H-J.) et Petkovic (D.). – Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia Tools and Applications*, vol. 3, n3, September 1996, pp. 179–202.

[BF98] Bouthemy (P.) et Fablet (R.). – Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. *In : Proc. 14th Int. Conf. on Pattern Recognition, ICPR'98*, pp. 905–908. – Brisbane, August 1998.

[BGG99] Bouthemy (P.), Gelgon (M.) et Ganansia (F.). – A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, n7, 1999, pp. 1030–1044.

[BMM99] Brunelli (R.), Mich (O.) et Modena (C.M.). – A survey on the automatic indexing of video data. *Jal of Visual Communication and Image Representation*, vol. 10, 1999, pp. 78–112.

[BV98] Bonet (J.S. De) et Viola (P.). – Texture recognition using a non-parametric multi-scale statistical model. *In : Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'98*, pp. 641–647. – Santa-Barbara, June 1998.

[CBD00] Chen (J.-Y.), Bouman (C. A.) et Dalton (J. C.). – Hierarchical browsing and search of large image databases. *IEEE Trans. on Image Processing*, 2000. – To appear.

[Courtney97] Courtney (J.D.). – Automatic video indexing via object motion analysis. *Pattern Recognition*, vol. 30, n4, April 1997, pp. 607–625.

[DAKGK00] Dagtas (S.), Al-Khatib (W.), Ghafoor (A.) et Kashyap (R.L.). – Models for motion-based video indexing and retrieval. *IEEE Trans. on Image Processing*, vol. 9, n1, 2000, pp. 88–101.

[DGLS81] Diday (E.), Govaert (G.), Lechevallier (Y.) et Sidi (J.). – Clustering in pattern recognition. *In : Digital Image Processing*, pp. 19–58. – J.-C. Simon, R. Haralick, eds, Kluwer Academic Publisher, 1981.

[FB99] Fablet (R.) et Bouthemy (P.). – Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval. *In : Proc. 3rd Int. Conf. on Visual Information and Information Systems, VISUAL'99.* – Amsterdam, June 1999. LNCS n1614, pp. 221–228, Springer-Verlag.

[FBG99] Fablet (R.), Bouthemy (P.) et Gelgon (M.). – Moving object detection in color image sequences using region-level graph labeling. *In : Proc. 6th IEEE Int. Conf. on Image Processing, ICIP'99.* – Kobe, Japan, October 1999.

[FLE95] Fischer (S.), Lienhart (R.) et Effelsberg (W.). – Automatic recognition of film series. *In : Proc. ACM Multimedia retrieval.*

[FT98] Ferman (A. Muffit) et Tekalp (A. Murat). – Efficient filtering and clustering methods for temporal video segmentation and visual summarization. *Jal of Visual Communication and Image Representation*, vol. 9, n4, December 1998, pp. 336–351.

[FTM98] Ferman (A. Muffit), Tekalp (A. Murat) et Mehrotra (R.). – Effective content representation for video. *In : Proc. 5th IEEE Int. Conf. on Image Processing, ICIP'98*, pp. 521–525. – Chicago, October 1998.

[GB98] Gelgon (M.) et Bouthemy (P.). – Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. *In : Proc. 5th Eur. Conf. on Computer Vision, ECCV'98* – Freiburg, Germany, June 1998. LNCS n1406, pp. 595–609., Springer-Verlag.

[GBD99] Gelgon (M.), Bouthemy (P.) et Dubois (T.). – A region tracking technique with failure detection for an interactive video indexing environment. *In : Proc. 3rd Int. Conf. on Visual Information and Information Systems, VISUAL'99.* – Amsterdam, June 1999. LNCS n1614, pp. 261–268, Springer-Verlag.

[GG84] Geman (S.) et Geman (D.). – Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, n6, 1984, pp. 721–741.

[Gimelfarb96] Gimel'Farb (G.L.). – Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, n11, November 1996, pp. 1110–1114.

[JVX99] Jain (A.K.), Vailaya (A.) et Xiong (W.). – Query by video clip. *Multimedia Systems*, vol. 7, n5, 1999, pp. 369–384.

[JZ97] Jain (A.) et Zongker (D.). – Feature selection : evaluation, application and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, n2, February 1997, pp. 153–158.

[MSP96] Milanese (R.), Squire (D.) et Pun (T.). – Correspondence analysis and hierarchical indexing for content-based image retrieval. *In : Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, pp. 859–862. – Lausanne, September 1996.

[NKFH98] Naphade (M.R.), Kristjansson (T.T.), Frey (B.J.) et Huang (T.). – Probabilistic multimedia objects (Multijects) : a novel approach to video indexing and retrieval in multimedia systems. *In : Proc. 5th IEEE Int. Conf. on Image Processing, ICIP'98*, pp. 536–5450. – Chicago, October 1998.

[NP92] Nelson (R.) et Polana (R.). – Qualitative recognition of motion using temporal texture. *Computer Vision, Graphics, and Image Processing*, vol. 56, n 1, July 1992, pp. 78–99.

[OB95] Odobez (J.M.) et Bouthemy (P.). – Robust multiresolution estimation of parametric motion models. *Jal of Visual Communication and Image Representation*, vol. 6, n4, 1995, pp. 348–365.

[OB97] Odobez (J.M.) et Bouthemy (P.). – Separation of moving regions from background in an image sequence acquired with a mobile camera. *In : Video Data Compression for Multimedia Computing*, chap. 8, pp. 295–311. – H. H. Li, S. Sun, and H. Derin, eds, Kluwer Academic Publishern, 1997.

[OHSF98] Otsuka (K.), Horikoshi (T.), Suzuki (S.) et Fujii (M.). – Feature extraction of temporal texture based on spatio-temporal motion trajectory. *In : Proc. 14th Int. Conf. on Pattern Recognition, ICPR'98*, pp. 1047–1051. – Brisbane, August 1998.

[RHM99] Rui (Y.), Huang (T.) et Mehrota (S.). – Constructing table-of-content for videos. *Multimedia Systems*, vol. 5, n7, September 1999, pp. 359–368.

[Schweitzer99] Schweitzer (H.). – Organizing image databases as visual-content search trees. *Image and Vision Computing*, vol. 17, 1999, pp. 501–511.

[SJ99] Santini (S.) et Jain (R.). – Similarity measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, n9, 1999, pp. 871–883.

[SP96] Szummer (M.) et Picard (R.W.). – Temporal texture modeling. *In : Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, pp. 823–826. – Lausanne, September 1996.

[VL98] Vasconcelos (N.) et Lippman (A.). – A Bayesian framework for semantic content characterization. *In : Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'98*, pp. 566–571. – Santa-Barbara, June 1998.

[VL99] Vasconcelos (N.) et Lippman (A.). – Probabilistic retrieval: new insights and experimental results. *In : Workshop on Content-Based Access of Image and Video Libraries, CVPR'99*, pp. 62–66. – Denver, June 1999.

[YYL96] Yeung (M. M.), Yeo (B.-L.) et Liu (B.). – Extracting story units from long programs for video browsing nd navigation. *In : Proc. 3rd IEEE Int. Conf. on Multimedia Computing and Systems, ICMCS'96*, pp. 296–305. – Hiroshima, Japan, June 1996.

[ZWM98] Zhu (S.C.), Wu (T.) et Mumford (D.). – Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling. *Int. Journal of Computer Vision*, vol. 27, n2, 1998, pp. 107–126.

[ZWZS97] Zhang (H.J.), Wu (J.), Zhong (D.) et Smoliar (S.). – An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, vol. 30, n4, April 1997, pp. 643–658.

Figure 1: Set of the 20 processed video shots used to visualize an example of motion-based hierarchal classification. For each video, we display the median image of the sequence.

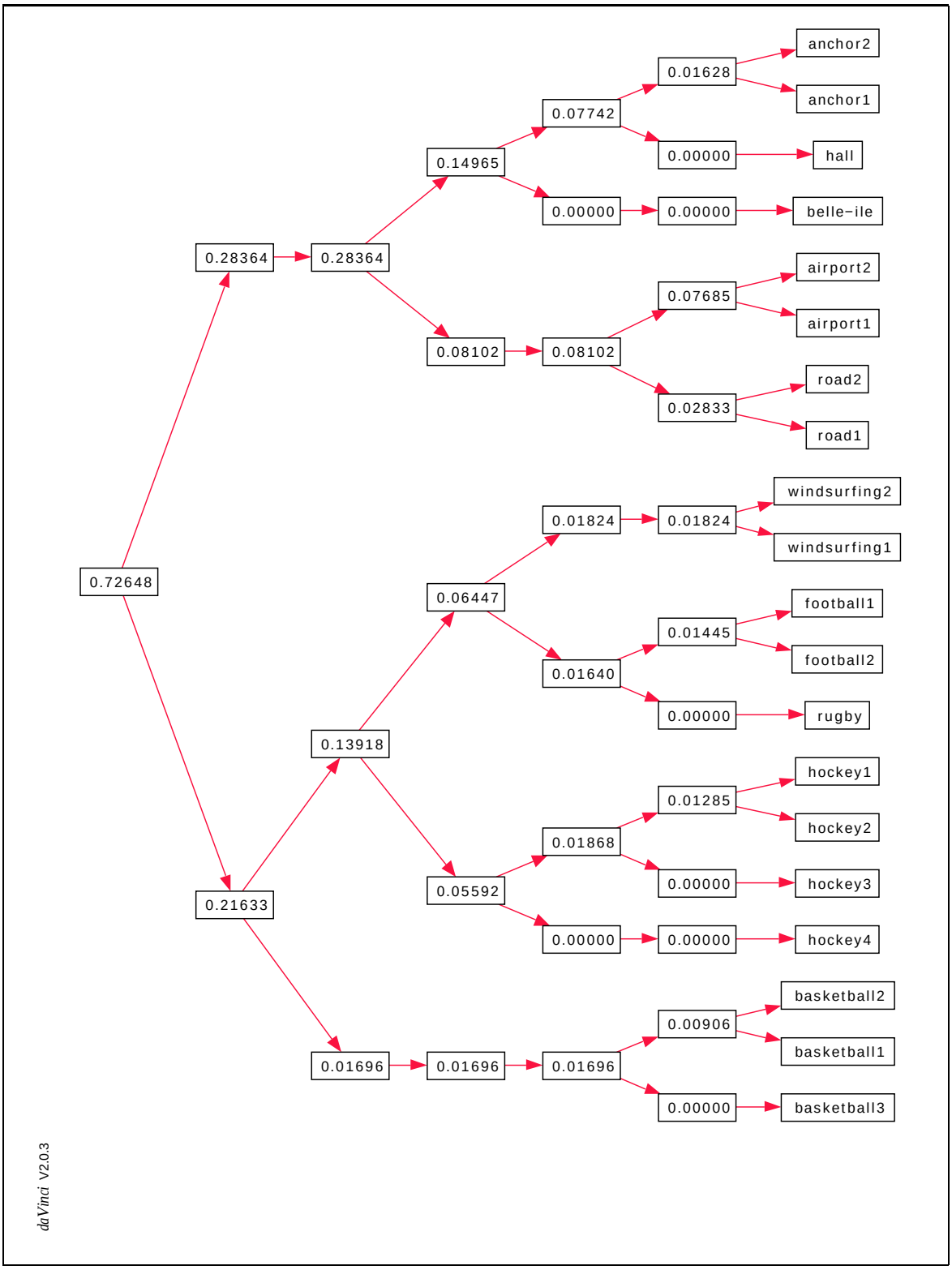

Figure 2: **Model-based statistical classification:** example of motion-based hierarchical classification for the set of 20 video sequences presented in Fig.1 with $D_{max} = 0.75$. At each leave of the tree, we report the name of the video sequence. For the other nodes of the tree, we display the maximum intra-cluster distance evaluated using expression D.

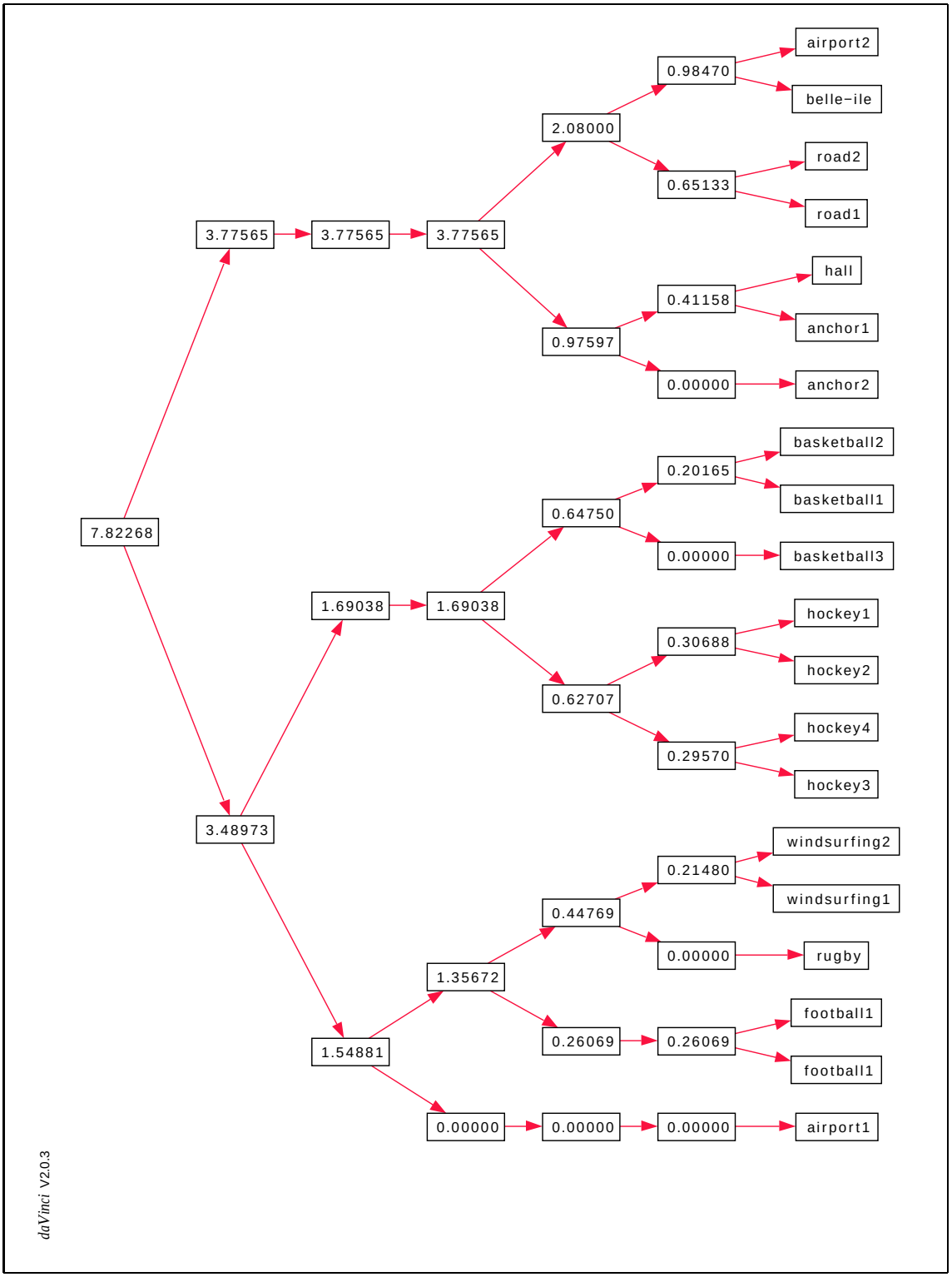

Figure 3: **Feature-based geometrical classification:** example of motion-based hierarchical classification for the set of 20 video sequences presented in Fig.1 with $D_{max} = 10$. At each leave of the tree, we report the name of the video sequence. For the other nodes of the tree, we display the maximum intra-cluster distance evaluated using the geometrical distance D_F in the feature space.

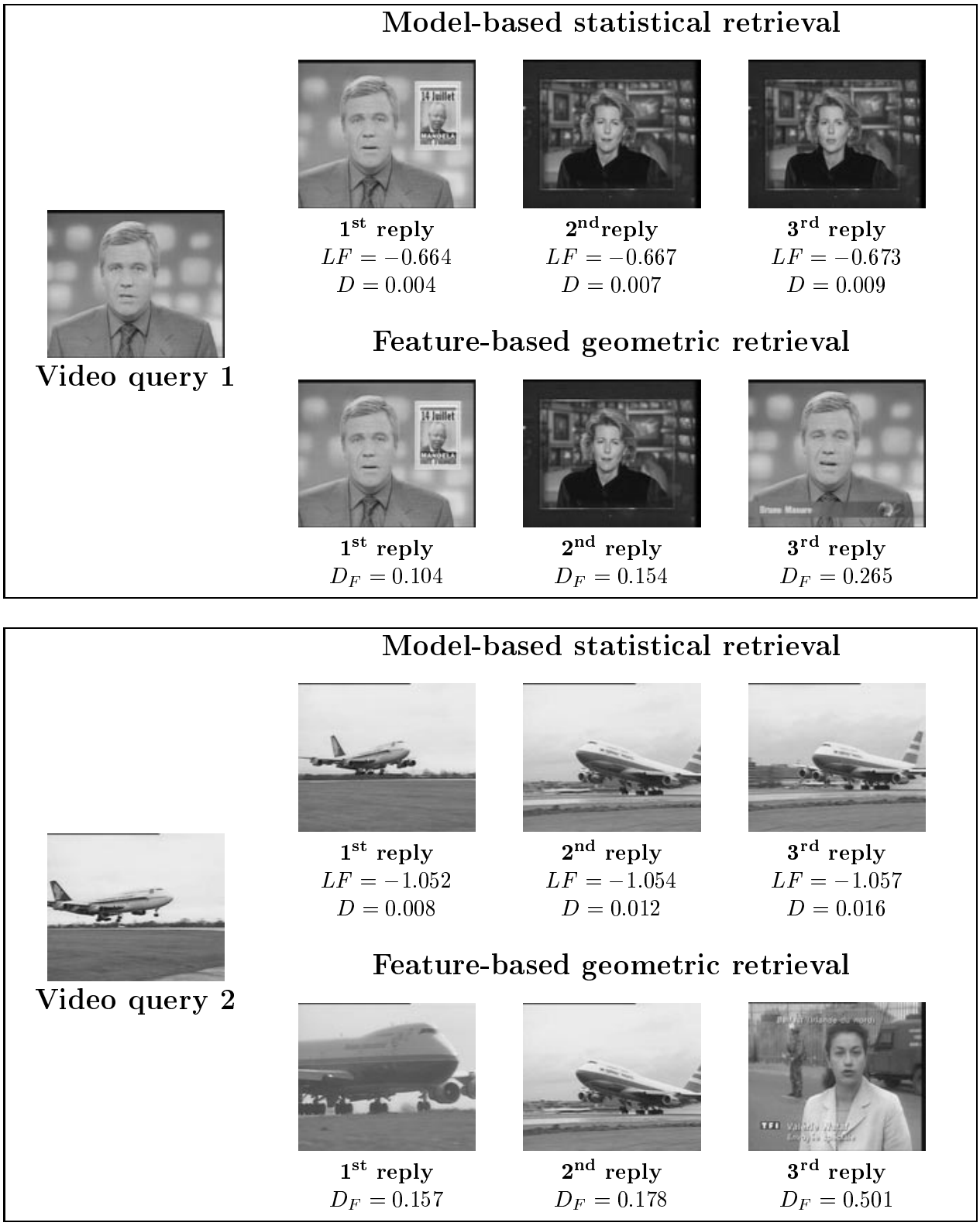

Figure 4: Example of retrieval operations involving three replies using the *model-based statistical* and *feature-based geometrical* techniques. Considering the first approach, we give for each reply n the value LF of the log-likelihood $\ln\left(P_{V^n}(x^q)\right)$ corresponding to video query q. In addition, to a posteriori evaluate the relevance of the replies, we have also estimated the model V^q associated to the query and we report the distances D between V^q and the different retrieved models V^n. Besides, for the second technique, we give the value D_F of the Euclidean distance in the feature space between the video query and the replies.

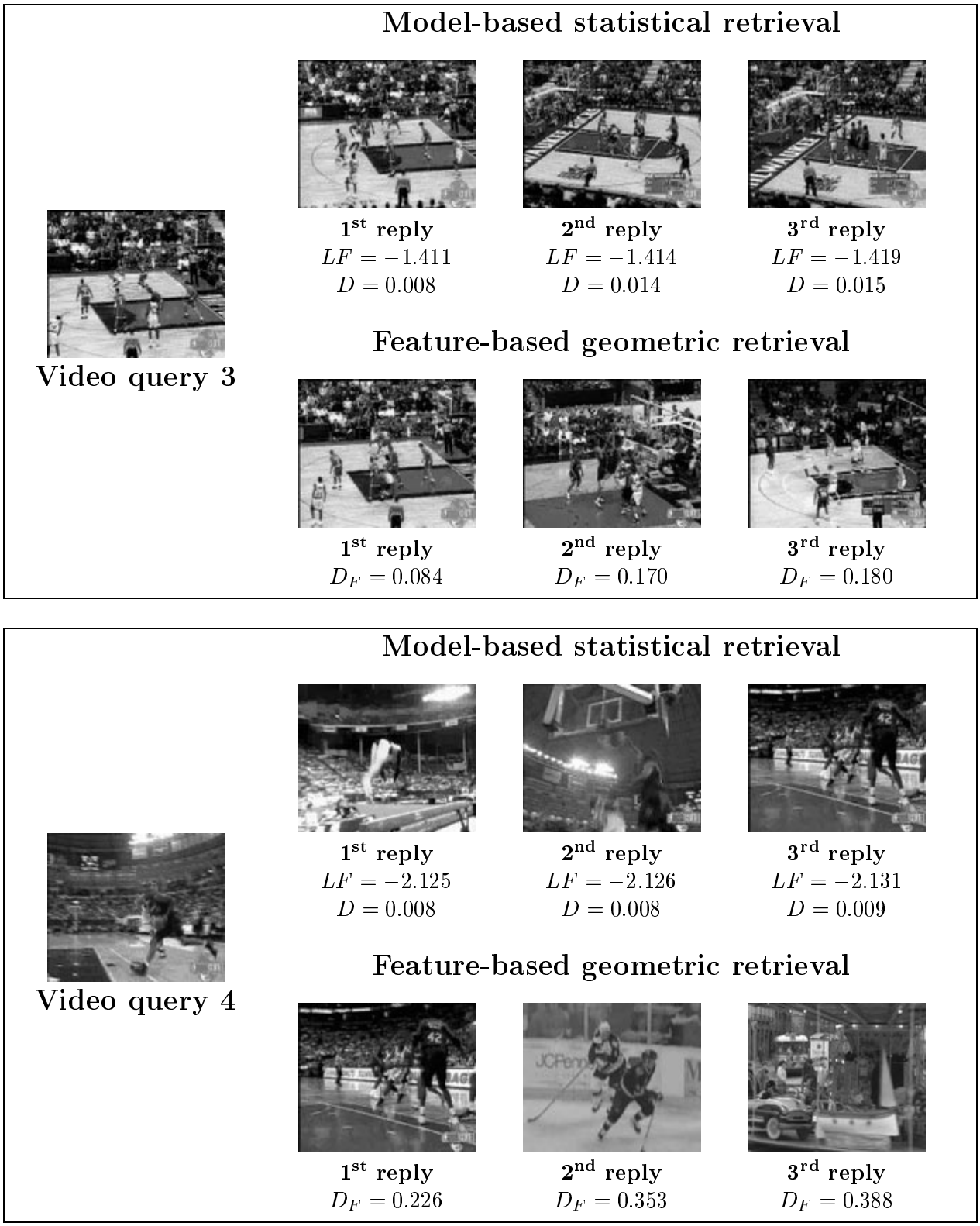

Figure 5: Example of retrieval operations involving three replies using the *model-based statistical* and *feature-based geometrical* techniques. Considering the first approach, we give for each reply n the value LF of the log-likelihood $\ln\left(P_{V^n}(x^q)\right)$ corresponding to video query q. In addition, to a posteriori evaluate the relevance of the replies, we have also estimated the model V^q associated to the query and we report the distances D between V^q and the different retrieved models V^n. Besides, for the second technique, we give the value D_F of the Euclidean distance in the feature space between the video query and the replies.