# Non Parametric Statistical Analysis of Scene Activity for Motion-Based Video Indexing and Retrieval

Ronan Fablet, Patrick Bouthemy, Patrick Pérez

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Non Parametric Statistical Analysis of Scene Activity for Motion-Based Video Indexing and Retrieval*

Ronan Fablet, Patrick Bouthemy and Partick Pérez

**N˚4005**

Septembre

THÈME 3


ISSN 0249-6399 ISRN INRIA/RR--4005--FR+ENG

# Non Parametric Statistical Analysis of Scene Activity for Motion-Based Video Indexing and Retrieval

Ronan Fablet, Patrick Bouthemy and Partick Pérez

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Vista

Rapport de recherche n ° 4005 — Septembre — 34 pages

**Abstract:** This report describes an original approach for content-based video indexing and retrieval. We provide a global interpretation of the dynamic content of video shots without any prior motion segmentation and without any use of dense optic flow fields. To this end, we exploit the spatio-temporal distribution within a shot of appropriate local motion-related measurements issued from the spatio-temporal derivatives of the intensity function. These distributions are then represented by causal Gibbs models. The considered statistical modeling framework makes possible the exact computation of the conditional likelihood function of a video shot to belong to a given motion or more generally activity class. This property allows us to develop a general statistical framework for video indexing and retrieval with query by example. We build a hierarchical structure of the processed video base according to motion content similarity. We consider a similarity measure inspired from Kullback-Leibler divergence. Then, retrieval with query by example is performed through this binary tree using the MAP criterion. We have obtained promising results on a set of various real image sequences.

**Key-words:** Non-parametric motion analysis - Video databases - Motion-based indexing - Query by example - Causal Gibbs models - Maximum likelihood estimation - Temporal cooccurrence

*(Résumé : tsvp)*

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

# Caractérisation statistique non paramétrique du mouvement pour l'indexation et la recherche de vidéos par le contenu

**Résumé :** Ce rapport décrit une approche originale pour l'indexation de vidéos par le contenu. Nous cherchons à fournir une analyse globale du contenu dynamique des plans vidéo sans segmentation préalable au sens du mouvement, ni utilisation de techniques d'estimation de champs denses de vitesses. Pour ce faire, nous exploitons des distributions spatio-temporelles de mesures locales de mouvement calculées à partir des gradients spatio-temporelles de la fonction intensité dans l'image. Ces distributions sont alors représentées par des modèles de Gibbs causaux. De plus, afin d'effectuer une caractérisation du mouvement liée à la scène observée, les mesures de mouvement sont calculées dans la séquence d'images obtenue après compensation du mouvement dominant entre images supposé dû au mouvement de la caméra. La modélisation statistique proposée permet d'effectuer le calcul exact de la vraisemblance conditionnelle des mesures de mouvement conditionnellement à une classe de mouvement ou plus généralement d'activité. Cette propriété nous permet de définir un cadre statistique général pour l'indexation de vidéos par le contenu et la recherche de vidéos par l'exemple. Ainsi, nous pouvons construire une structuration hiérarchique d'une base de vidéos relativement aux contenus de mouvement. Elle consiste en la détermination d'un arbre binaire pour lequel chaque noeud est associé à un des modèles causaux de Gibbs préalablement appris sur chaque vidéo de la base. Cette classification hiérarchique exploite une mesure de similarité s'appuyant sur des distances de Kullback-Leibler. La recherche d'exemples vidéo similaires à une vidéo proposée comme requête peut alors être effectuée à travers la représentation hiérarchique de la base de vidéos et un critère bayésien du type MAP. Nous avons obtenu des résultats prometteurs sur un ensemble significatif de séquences vidéo réelles.

**Mots-clé :** Analyse non-paramétrique du mouvement - Indexation vidéo - Recherche par l'exemple - Modèles de Gibbs causaux - Maximum de vraisemblance - Cooccurrence temporelle

# 1 Introduction and related work

Image sequence archives are at the core of various application fields such as meteorology (satellite image sequences), road traffic surveillance, medical imaging, or TV broadcasting (audio-visual archives including movies, documentaries, news, ...). An entirely manual annotation of visual documents is no more able to cope with the rapidly increasing amount of video data. Besides, the efficient use of these databases requires to offer reliable and relevant access to visual information. As a consequence, this implies to index and retrieve visual documents by their content. A great deal of research is currently devoted to image and video database management [1, 6]. Nevertheless, it remains hard to easily identify the relevant information for a given query, due to the complexity of image and scene interpretation.

Furthermore, new needs appear for tools and functionalities concerned with efficient video navigation and browsing, with the classification of video sequences into different genres (sports, news, movies, commercials, documentaries, ... ) [36], with the retrieval of examples similar to a given video query [11, 14, 24], or with high-level video structuring such as macro-segmentation [37, 31]. Such applications require to combine content-based video description with the definition of an appropriate measure of video similarity.

As far as content-based video indexing is concerned, the primary task generally consists in segmenting the video into elementary shots [5, 39][1]. This stage is usually associated to the recognition of typical forms of video shooting such as static shot, panning , traveling or zooming [5]. At a second stage, it appears necessary to provide an interpretation and a representation of the shot content. In that context, dynamic content analysis is of particular interest. Mainly, two kinds of approaches are considered to characterize dynamic content in video sequences. A first class of approaches, based on parametric or dense motion field estimation, includes image mosaicing [16, 21], segmentation, tracking and characterization of moving elements in order to determine a spatio-temporal representation of the video shot [16, 8, 15]. The description of the motion content may then rely on the extraction of pertinent qualitative features for the extracted entities of interest, such as the direction of the displacement [16], or on the analysis of the trajectories of the center of gravity of the tracked objects [9]. However, these techniques turn out to be unadapted to certain classes of sequences with complex dynamic contents such as motion of rivers, flames, foliages in the wind, crowds, etc. Furthermore, as far as video indexing is concerned,

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

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 this kind of entities. Therefore, in the context of video indexing, it seems relevant to adopt a global point of view that avoids any explicit motion segmentation step.

This leads to consider a second category of methods for motion-based video indexing and retrieval. Our goal is to interpret dynamic contents without any prior motion segmentation and without any complete motion estimation in terms of parametric motion models or optical flow fields. Preliminary work in that direction has proposed the extraction of "temporal texture" features, [11, 4, 27, 30, 34]. Motions of rivers, foliages, flames, or crowds, for instance, can indeed be regarded as temporal textures. In [30], temporal texture features are extracted from the description of surfaces related to spatio-temporal trajectories. In [27], features issued from spatial cooccurrences of the normal flow field are exploited to classify sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In our previous work concerned with motion-based video classification and retrieval [11, 4], we have considered global features extracted from temporal cooccurrence distributions of local motion-related measurements which were proved more reliable than normal velocities. In this paper, we introduce a non parametric probabilistic modeling of the dynamic content of video shots evaluated by these temporal cooccurrences. It allows us to design an original, coherent and efficient framework for both motion-based video indexing and retrieval.

The remainder of the paper is organized as follows. Section 2 outlines the general ideas underlying our work. Section 3 describes the local non parametric motion-related measurements that we use. In Section 4, we introduce the statistical modeling of the spatio-temporal distribution of the motion-related quantities computed from a video sequence and the associated estimation scheme. Section 5 deals with the application to content-based video indexing. This involves the design of a hierarchical video classification scheme and of an appropriate video similarity measure based on the Kullback-Leibler divergence. Both are then exploited to satisfy queries by example with a statistical framework. In Section 6, we report experimental results of video classification and retrieval examples over a set of video sequences. Section 7 contains concluding remarks.

## 2 Problem statement

As previously pointed out, the description of shot content must be combined with the definition of an appropriate measure of shot similarity to handle video navigation, browsing or retrieval. Usually, shot content characterization relies on the extraction of a set of numerical features or descriptors, and the comparison of shot contents is performed in the feature space according to a given distance such as the Euclidean distance or more elaborated measures [32]. As a consequence, to cope with video databases involving various dynamic contents, it is necessary to determine an optimal set of features and the associated similarity measure. These issues can be tackled using Principal Component Analysis [25] or some other feature selection techniques [23]. Nevertheless, feature space is usually of high dimension, and the considered distance is likely not to capture properly the uncertainty attached to feature measurements.

Statistical methods appear more suited in that context. In addition, they also provide a unified view for learning and classification. Furthermore, a Bayesian scheme can then be adopted to properly formalize the retrieval process. In [35], modeling of DCT coefficients by Gaussian distribution mixture is exploited for image texture indexing and the retrieval operation is formulated in a Bayesian framework w.r.t. MAP criterion. This statistical approach is shown to outperform classical techniques using distances in the feature space.

We follow such a statistical approach in the context of motion-based video indexing. Our goal is to define a direct and general characterization of motion information allowing us to provide within the same framework efficient statistical tools for video database classification and video retrieval with query by example. To this end, we have designed a motion classification (or, more generally, scene activity classification) method relying on a statistical analysis of the spatio-temporal distribution of local non-parametric motion-related measurements. We aim at identifying probabilistic models corresponding to different dynamic content types. Indeed, in recent work [18, 40], a correspondence has been established between cooccurrence distributions and Markov random field models in the context of spatial texture analysis. We propose an extension to temporal textures while introducing only causal statistical models. More precisely, we consider causal Gibbs models. Since the exact conditional likelihood function can be straightforwardly computed in that context, this allows us to develop a general and efficient statistical framework for video indexing and retrieval with query by example.

# 3 Local motion-related measurements

We have to define appropriate local motion-related measurements to be used for classification. Since our goal is to characterize the actual dynamic content of the scene, we have first to cancel camera motion. As a consequence, we estimate the dominant image motion between two successive images which is assumed to be due to camera motion. Then, to cancel it, we warp the successive images to the first image of the video shot by combining the elementary dominant motions successively estimated over consecutive image pairs.

## 3.1 Dominant motion estimation

To model the transformation between two successive images, we consider a 2D affine motion model. A 2D quadratic model involving eight parameters, i.e. corresponding to the 3D rigid motion of a planar surface, could be alternatively considered. However, it is computationally more demanding while not significantly offering more adequacy in most situations. The displacement $\mathbf{w}_\Theta(p)$, at pixel $p$, related to the affine motion model parameterized by $\Theta$ 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 computation is achieved with the gradient-based multi-resolution incremental estimation method described in [28]. The following minimization problem is solved:

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

where $DFD(p, \Theta) = I_{t+1}(p + \mathbf{w}_\Theta(p)) - I_t(p)$ with $I$ the intensity function in the image, $\mathcal{R}$ denotes the image grid, and $\rho$ is robust M-estimator, here Tukey biweight function. 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. 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 [28].

## 3.2 Local motion-related measurements

To characterize the nature of residual motion in the motion compensated image sequence, we need to specify appropriate local motion-related measurements. A dense

optic flow field provides such local information [24, 2]. Nevertheless, as stressed previously, the accuracy and 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. Hence, we prefer to consider local motion-related measurements directly computed from the spatio-temporal derivatives of the intensity function in the image.

By assuming intensity constancy along 2D motion trajectories, the image motion constraint relating the 2D residual motion and the spatio-temporal derivatives of the intensity function can be expressed as follows [20]:

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

where $\mathbf{w}(p)$ is the 2D residual motion vector at pixel $p$, and $I^*$ the intensity function in the warped image. We can infer the residual normal velocity $v_n^*(p)$ in the motion compensated sequence at pixel $p$:

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

where $I_t^*(p) \triangleq \frac{\partial I^*(p)}{\partial t}$ is the temporal derivative of the intensity function $I^*$. $I_t^*(p)$ 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 residual 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 pointed out in [29], the norm of the spatial image gradient $\|\nabla I^*(s)\|$ 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 pixel $p$, an appropriately weighted average of $v_n^*(p)$ in a local neighborhood appears as a relevant motion-related quantity. More precisely, we consider the following expression :

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

where $\mathcal{F}(p)$ is a small window centered on $p$ and $G^2$ a predetermined constant related to the noise level in uniform areas. This motion-related measurement forms a more

reliable quantity than the normal flow, yet simply computed from the intensity function and its derivatives. This local motion information was successfully exploited for the detection of mobile objects in motion compensated sequences [29, 22, 13].

Besides, we have to cope with the limitations of the gradient-based image motion constraint (3). As a matter of fact, this relation is no longer valid in occluded regions, over motion discontinuities, and even on sharp intensity discontinuities. In addition, it cannot handle large displacement magnitude. Therefore, we adopt a multiscale strategy to compute $v_{obs}(p)$ at a reliable scale and we exploit an appropriate test to validate its use. More precisely, we build a Gaussian pyramid of the considered image and the next one. At each pixel $p$, we determine the lowest scale for which the image motion constraint (3) is valid using the statistical test designed in [19]. Then, $v_{obs}(p)$ is computed at the selected scale. If for a given pixel $p$ the image motion constraint remains invalid at all scales, no motion quantity is computed at $p$.

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

The computation of the temporal cooccurrences of the motion-related measurements $\{v_{obs}(p)\}_{p\in\mathcal{R}}$ requires to quantize these continuous variables. By definition, these quantities are positive and, for a given pixel $p$, $v_{obs}(p)$ is theoretically inferior to the greatest actual displacement magnitude in the window $\mathcal{F}(p)$. We could merely apply a linear quantization within $[0, v_{obs}^{max}]$ with $v_{obs}^{max} = \max_{p\in\mathcal{R}} v_{obs}(p)$. In that case, we would face two main problems. First, since we aim at evaluating content similarity between video shots, a common range of quantized motion-related quantities to all image sequences has to be selected. As illustrated in Fig.1, it does not make sense to directly compare the histograms of basketball and anchor shots if a linear quantization over $[0, v_{obs}^{max}]$ is retained, whereas maximum values $v_{obs}^{max}$ greatly differ between these two shots. Secondly, although we consider a multiscale strategy combined with a validity test of the image motion constraint, we may still get spurious motion quantities of usually irrelevant high magnitude of displacement in the scene since the validity test may fail in some specific situations. The range of quantities $\{v_{obs}(p)\}_{p\in\mathcal{R}}$ is indeed within $[0.0, 15.4]$ in the first example of Fig.1. Thus, a linear quantization within $[0, v_{obs}^{max}]$ would result in the loss of the main part of information contained in the empirical distribution $\{v_{obs}(p)\}_{p\in\mathcal{R}}$ as illustrated in Fig.1 for the basketball shot. Therefore, we prefer to consider a linear quantization within a predefined interval $[0, V_{max}]$. Applying this quantization scheme, the direct com-

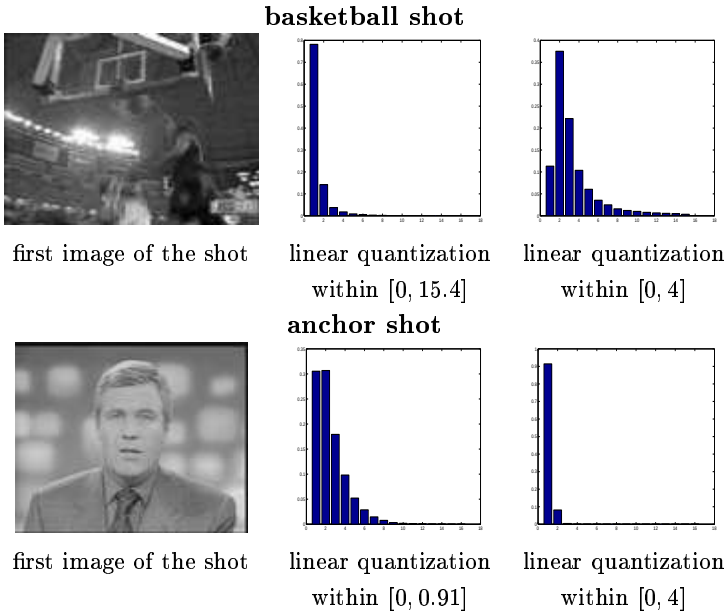



Figure 1: *Quantization of motion-related measurements* $\{v_{obs}(p)\}_{p\in\mathcal{R}}$. *We display two examples of quantization of the motion-related quantities for a basketball shot and an anchor shot. The first column depicts the first image of the processed shot; the middle one the histogram resulting from a linear quantization on 16 levels within the interval* $[0, v_{obs}^{max}]$, $v_{obs}^{max} = 15.4$ *in the first example and* $v_{obs}^{max} = 0.91$ *in the second one; the last one contains the histogram resulting from a linear quantization within* $[0, 4]$ *over 16 levels.*

parison of the quantized versions of motion-related measurements becomes relevant. For instance in Fig.1, the motion activity is greater in the basketball shot compared to the anchor shot as confirmed by the histograms of quantized motion-related values obtained with $V_{max} = 4$.

Let denote $\Lambda$ the discretized range of variations for $\{v_{obs}(p)\}_{p\in\mathcal{R}}$. In the sequel, we note $x_k$ these quantized motion-related measurements for the $k$th image of the video sequence.

# 4 Temporal Gibbs models

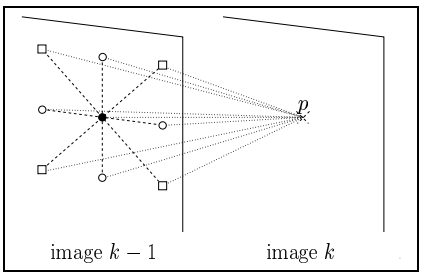
## 4.1 Causal Gibbs random fields

We now present our statistical modeling framework for the characterization of motion information within a video shot. Our goal is to associate a probabilistic model to a sequence of quantized motion-related quantities. As mentioned in Section 2, we consider Gibbs models since they offer a direct correspondence with cooccurrence measurements which we previously exploited for video indexing in [11, 4]. Furthermore, we have investigated a purely causal modeling for two main reasons. First, we are concerned with the temporal evolution of the distribution of motion-related quantities. Besides, such a causal approach allows us to handle temporal non-stationarities while being sufficient to discriminate motion classes of interest. Secondly, from a theoretical point of view, it is quite beneficial to be able to compute the exact likelihood function attached to a model to properly establish a content-based video similarity measure. Whereas this is generally not possible with classical spatial Markov random fields [17] due to the unknown partition function, this becomes easily feasible with our causal Gibbs models. Thus, this attractive property allows us to design a general statistical framework for video classification and retrieval based on likelihood computation.

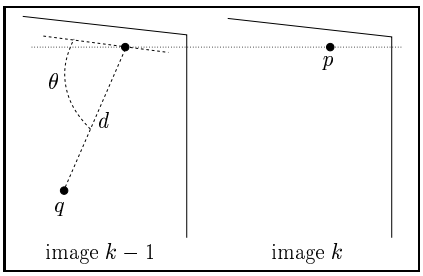
We assume that the sequence of the motion-related quantities along a given video shot $x = (x_k)_{k=0,..,K}$ is the realization of a random field $X = (X_0, \ldots, X_K)$, and that $X$ is a first-order Markov chain:

$$P_\Psi(X = x) = P_\Psi(X_0 = x_0) \prod_{k=1}^{K} P_\Psi(X_k = x_k | X_{k-1} = x_{k-1}) \tag{6}$$

where $\Psi$ refers to the underlying interaction potentials to be defined later. In addition, we assume that the random variables $(X_k(p))_{p\in\mathcal{R}}$ at time $k$ are independent conditionally to $X_{k-1}$. Thus, we assume that conditional probabilities $P_\Psi(x_k|x_{k-1})$




*Causal temporal neighborhood of pixel p*




*Parameterization of the temporal pair $(p,q)$ using polar coordinates $a=(d,\theta)$*

Figure 2: *Causal temporal neighborhood comprising up to 9 pairs. We denote $\eta_p^1$ the temporal neighborhood formed by one single site p with symbol $\bullet$, $\eta_p^5$ the set of 5 neighbors of p represented by symbols $\bullet$ and $\circ$, and $\eta_p^9$ the complete set of 9 neighbors of p (symbols $\bullet$, $\circ$ and $\square$).*

factorize as:

$$\begin{aligned} P_\Psi(x_k|x_{k-1}) &= \prod_{p\in\mathcal{R}} P_\Psi\left(x_k(p)|x_{k-1}\right) \\ &= \prod_{p\in\mathcal{R}} P_\Psi\left(x_k(p)|x_{k-1}(\eta_p)\right) \end{aligned} \tag{7}$$

where $\mathcal{R}$ is the image grid, and $\eta_p$ designates the set of sites in image $k-1$ which interact with site $p$ in image $k$. $\eta_p$ will be called the temporal neighborhood of site $p$ and is specified in Fig.2. We consider a small set of temporal interactions. Each pair $(p,q)$, with $q\in\eta_p$, can be characterized by the polar coordinates $a=(d,\theta)$ (see Fig.2). Let us denote $\mathcal{A}$ the set of polar coordinates $a$ corresponding to the temporal pairs defined in Fig.2. In the sequel, we will use the term clique to designate a temporal pair. In practice, we consider three different neighborhoods $\eta^1$, $\eta^5$ and $\eta^9$ (Fig.2). The simplest case $\eta^1$ includes the temporal clique defined by $a=(0,0)$ whereas $\eta^5$ and $\eta^9$ respectively refer to the cases with 5 cliques and 9 cliques.

Let us introduce a Gibbs model to express the conditional probability $P_\Psi\left(x_k|x_{k-1}\right)$. This comes to write:

$$P_\Psi\left(x_k(p)|x_{k-1}(\eta_p)\right) = \frac{\exp\left[\sum_{a\in\mathcal{A}} \Psi_a\left(x_k(p), x_{k-1}(p_a)\right)\right]}{Z_k(p, x_{k-1}(\eta_p))} \tag{8}$$

where $\{\Psi_a(\nu,\nu')\}_{(\nu,\nu')\in\Lambda^2,a\in\mathcal{A}}$ denote the potentials attached to the Gibbs model $\Psi$ for each label pair $(\nu,\nu')$ and temporal clique $a$. Thus, model $\Psi$ is defined by $|\mathcal{A}|\cdot|\Lambda|^2$ potentials. $p_a$ is the temporal neighbor of $p$ for clique $a$ for the considered neighborhood $\eta_p$, and $Z_k(p)$ the local normalization constant. $Z_k(p)$ is given by:

$$Z_k(p, x_{k-1}(\eta_p)) = \sum_{\nu\in\Lambda} \exp\left[\sum_{a\in\mathcal{A}} \Psi_a\left(\nu, x_{k-1}(p_a)\right)\right] \tag{9}$$

Let us point out that $P_\Psi$ is not uniquely defined (see Appendix A for further details). More precisely, for a given pair $(a,\nu')\in\mathcal{A}\times\Lambda$, the potentials $\{\Psi_a(\nu,\nu')\}_{\nu\in\Lambda}$ are defined up to an additive constant. To guarantee the uniqueness of the potentials of the causal Gibbs model, we add the following normalization constraint:

$$\forall(a,\nu')\in\mathcal{A}\times\Lambda,\quad \sum_{\nu\in\Lambda}\exp\Psi_a(\nu,\nu') = 1 \tag{10}$$

For convenience, we fix the constant potential defined by $\forall(\nu,\nu')\in\Lambda^2,\ \ \Psi_a(\nu,\nu') = -ln|\Lambda|$ as $\Psi_a\equiv 0$, and $\Psi\equiv 0$ as the constant model for all clique types.

Contrary to Markov random fields [17], this causal modeling leads to an expression of the global likelihood function $P_\Psi(x)$ as a simple product of local transitions:

$$P_\Psi(x) = P_\Psi(x_0)\prod_{k=1}^{K}\prod_{p\in\mathcal{R}}\frac{\exp\left[\sum_{a\in\mathcal{A}}\Psi_a\left(x_k(p), x_{k-1}(p_a)\right)\right]}{Z_k(p, x_{k-1}(\eta_p))} \tag{11}$$

Thus, for given $P_\Psi(x_0)$ and potentials $\Psi$, $P_\Psi$ is entirely known, which provides us with a general statistical framework for motion-based video classification and retrieval as described in Section 5. Following [18, 40], we can now rewrite the causal expression of relation (6) using the temporal cooccurrence measurements attached to the clique $a$ as follows:

$$P_\Psi(x) = P_\Psi(x_0)\frac{\exp\left[\sum_{a\in A}\Psi_a\bullet\Gamma_a(x)\right]}{Z_\Psi(x)}, \tag{12}$$

where $Z_\Psi(x)$ is the global normalization factor given by:

$$Z_\Psi(x) = \prod_{k=1}^{K}\prod_{p\in\mathcal{R}} Z_k(p, x_{k-1}(\eta_p)), \tag{13}$$

$\Gamma_a(x) = \{\Gamma_a(\nu, \nu'|x)\}_{(\nu,\nu')\in\Lambda^2}$ is the cooccurrence matrix for the clique type $a$ defined as:

$$\Gamma_a(\nu, \nu'|x) = \sum_{k=1}^{k=K} \sum_{(p,p_a)} \delta(\nu - x_k(p))\delta(\nu' - x_{k-1}(p_a)) \tag{14}$$

$\delta()$ denotes the Kronecker symbol, and $\bullet$ is the dot product between cooccurrence matrix $\Gamma_a$ and prior interaction potential $\Psi_a$ defined as follows:

$$\Psi_a \bullet \Gamma_a(x) = \sum_{(\nu,\nu')\in\Lambda^2} \Psi_a(\nu, \nu') \cdot \Gamma_a(\nu, \nu'|x) \tag{15}$$

Our statistical framework for motion information modeling in image sequences can be claimed as non parametric in two ways. First, the probabilistic model $\Psi$ is quite general, in particular it does not refer to a 2D parametric motion model. We can even assess that our description of motion information can be viewed as a general characterization of scene activity. Secondly, from a statistical point of view, our approach is also non parametric in the sense that the conditional likelihood $P_\Psi(x_k(p)|x_{k-1}(\eta_p))$ is not assumed to follow a known parametric law (Gaussian,...).

## 4.2 Maximum likelihood estimation

Given a realization $x$ of $X$, the causal temporal Gibbs model defined by its potentials $\{\Psi_a(\nu, \nu'),\ a \in \mathcal{A},\ (\nu, \nu') \in \Lambda^2\}$ can be estimated using the Maximum Likelihood (ML) criterion:

$$\widehat{\Psi}_{ML} = \arg\max_{\Psi} LF_\Psi(x) \quad \text{with } LF_\Psi(x) = \ln(P_\Psi(x)) \tag{16}$$

We hereafter assume that $P_\Psi(x_0)$ follows a uniform law. Using the formulation given in relation (12), we get:

$$\widehat{\Psi}_{ML} = \arg\max_{\Psi} \sum_{a\in A} \Psi_a \bullet \Gamma_a(x) - \log Z_\Psi(x) \tag{17}$$

Contrary to the difficult issue of ML estimation of Markov model potentials where the partition function is unknown, we do not need to use time-consuming stochastic techniques [38]. Since we can here directly compute the partial derivatives of the log-likelihood function, solving the issue expressed in (16) can then be achieved using

usual optimization techniques. The derivatives of $LF\Psi(x)$ w.r.t. $\Psi_a$ potentials are given by:

$$\forall(a,\nu,\nu'),\ \frac{\partial LF_\Psi(x)}{\partial \Psi_a(\nu,\nu')} = \Gamma_a(\nu,\nu'|x) - \sum_{(k,p)\in S_{a\nu'}} P_\Psi\left(x_k(p)=\nu|x_{k-1}(\eta_p)\right) \tag{18}$$

with $S_{a\nu'} = \{(k,p) \in \{1,.,K\} \times \mathcal{R} \ / \ x_{k-1}(p_a) = \nu'\}$. Let us point out that the computation of $\frac{\partial LF_\Psi(x)}{\partial \Psi_a(\nu,\nu')}$ is independent of the normalization supplying the uniqueness of the representation of the model. Setting to 0 the partial derivatives expressed in (18), we get the equations to be simultaneously verified by $\widehat{\Psi}_{ML}$:

$$\sum_{(k,p)\in S_{a\nu'}} P_{\widehat{\Psi}}\left(x_k(p)=\nu|x_{k-1}(\eta_p)\right) = \Gamma_a(\nu,\nu'|x) \tag{19}$$

This naturally confirms that informative potentials of model $\widehat{\Psi}$ correspond to high cooccurrence values. We will exploit this property to reduce the model complexity as explained in subsection 4.5. In practice, the maximization in (16) is carried out using a classical conjugate gradient algorithm as outlined in Fig.3. Let us mention that, since the computation of the partial derivatives of $LF\Psi(x)$ is independent of the normalization constraint, this ensures the descent direction of the optimization algorithm to be independent of the normalization constraint (10)too. In addition, the log-likelihood function $LF_\Psi(x)$ may have several local minima w.r.t. $\Psi$, whereas the existence of a unique global minimum is guaranteed in case of Markov models [18]. Hence, it is important to define an appropriate optimization scheme. As described later in subsection 4.4, we have adopted an incremental strategy which has proven robust and accurate enough.

### 4.3 Estimation of the simple temporal model

For the simplest model including only one clique, i.e. $\mathcal{A} = \{a_0\} = \{(0,0)\}$, parameter estimation is readily performed. The considered model is indeed equivalent to a product of $|\mathcal{R}|$ independent Markov chains with:

$$P_\Psi(x_k(p)|x_{k-1}(p)) \propto \exp\left[\Psi\left(x_k(p),x_{k-1}(p)\right)\right] \tag{20}$$

Normalizing $\Psi$ according to (10), we get the following ML estimate for a given sequence $x$ of local motion-related quantities:

$$\widehat{\Psi}_{ML}(\nu,\nu') = \log\left(\Gamma_{a_0}(\nu,\nu'|x)/\sum_{\nu\in\Lambda}\Gamma_{a_0}(\nu,\nu'|x)\right) \tag{21}$$

- *initialization of the Gibbs model* $\Psi^0$
- *initialization of the descent direction* $\mathbf{d_0} \equiv \mathbf{0}$
- *iteration* $l$*:*
  - *computation of the gradient* $\nabla LF_{\Psi^l}(x)$ *(according to relation (18))*
  - *update of the descent direction* $\mathbf{d_l}$ *:*

$$\mathbf{d_l} = \nabla \mathbf{LF}_{\Psi^l}(\mathbf{x}) + \frac{\|\nabla \mathbf{LF}_{\Psi^l}(\mathbf{x})\|^2}{\|\nabla \mathbf{LF}_{\Psi^{l-1}(\mathbf{x})}\|^2} \mathbf{d_{l-1}}$$

  - *search for the coefficient* $\lambda_l$ *which verifies:*

$$\lambda_l = \arg\min_{\lambda} LF_{\Psi^l + \lambda \mathbf{d_l}}(x)$$

  - *update of the model estimate:*

$$\Psi^{l+1} = \Psi^l + \lambda_l \mathbf{d_l}$$

- *stopping criterion :* $\|\nabla LF_{\Psi^l}(x)\|_\infty < \gamma$ *where* $\gamma$ *is a predefined constant.*

Figure 3: *Maximum likelihood estimation of model* $\Psi = (\Psi_a(\nu, \nu'))_{a \in \mathcal{A}, (\nu,\nu') \in \Lambda^2}$ *by applying a conjugate gradient technique to criterion (16)*

.

Since the potentials of the model verify relation (10), the likelihood function is simply given by:

$$P_\Psi(x) = \exp\Big[\Psi \bullet \Gamma_{a_0}(x)\Big] \tag{22}$$

### 4.4 Estimation of the extended temporal models

Let us now consider the case of an extended temporal neighborhood $\eta^5$ or $\eta^9$ (see Fig.2). To perform the ML estimation, we adopt an incremental strategy. First, we determine a ranking of the different cliques according to their relevance in the model. For each $a \in \mathcal{A}$, we evaluate the ML estimate of the specific model $\Psi^a$ with potentials set as constant for all cliques $b$ other than $a$:

$$\widehat{\Psi}^a = \arg \max_{\Psi^a : \forall b \neq a, \Psi_b \equiv 0} LF_\Psi(x) \tag{23}$$

As developed in subsection 4.3, the ML estimated potentials $\widehat{\Psi}^a_a$ defined on the clique $a$ are given by:

$$\forall (\nu, \nu') \in \Lambda^2, \quad \widehat{\Psi}^a_a(\nu, \nu') = \log \left( \Gamma_a(\nu, \nu'|x) / \sum_{\nu \in \Lambda} \Gamma_a(\nu, \nu'|x) \right) \tag{24}$$

We can rank cliques $a \in \mathcal{A}$ according to the values of the conditional likelihoods of the processed sequence of motion-related quantities $x$ w.r.t. $\widehat{\Psi}^a$ and form the set $\tilde{\mathcal{A}}$ of ranked cliques:

$$\tilde{\mathcal{A}} = \{a_1, .., a_{|\mathcal{A}|}\} \text{ with } LF_{\widehat{\Psi}^{a_1}}(x) > LF_{\widehat{\Psi}^{a_2}}(x) > .. > LF_{\widehat{\Psi}^{a_{|\mathcal{A}|}}}(x) \tag{25}$$

The incremental ML estimation of the Gibbs model $\Psi$ is then carried out in as follows. At iteration $l \in \{1, |\mathcal{A}|\}$, it consists in estimating the model $\widehat{\Psi}^l$ which satisfies the maximization criterion (16) under the constraint:

$$\forall b \in \{a_{l+1}, ..., a_{|\mathcal{A}|}\}, \quad \Psi_b \equiv 0 \tag{26}$$

This minimization is achieved using an iterative conjugate gradient procedure which exploits the computation of the derivatives of the log-likelihood function (see Fig.3). For the initialization at each minimization step, we take $\Psi^l = \widehat{\Psi}^{l-1}$. Finally, at iteration $|\mathcal{A}|$, we obtain the ML estimate $\widehat{\Psi}_{ML}$ defined on the whole considered temporal neighborhood structure.

## 4.5 Model complexity reduction and model structure selection

When considering $n$ cliques (i.e., $|\mathcal{A}| = n$) with $N$ levels of quantization (i.e., $|\Lambda| = N$) for the local motion-related measurements, $n \times N^2$ potentials $\{\Psi_a(\nu, \nu')\}_{a \in \mathcal{A},\ (\nu,\nu') \in \Lambda^2}$ have to be estimated. Typically, $N = 16$ and $n = 1, 5$, or 9. The number of potentials rapidly increases with the number of considered cliques. As far as video indexing is concerned, it is crucial to supply parsimonious content representations while keeping the characterization of the video content accurate enough. To this end, we aim at reducing the global model complexity while keeping the most pertinent information in the selected model. Two aspects are considered.

### 4.5.1 *modification of the range of* $\Lambda$

Some quantization levels may seldom appear in the sequence of local motion-related quantities $x$. In that case, the potentials associated to these quantization levels are of weak importance as stressed by relation (19). To select the relevant quantization levels, we compute the number of occurrences of each level $\nu \in \Lambda$ in the sequence $x$. For each level $\nu^0$ with an occurrence number lower than a given threshold, potentials $\{\Psi_a(\nu_0, \nu), \Psi_a(\nu, \nu_0)\}_{(a,\nu) \in \mathcal{A} \times \Lambda}$ are set to $-\infty$ (a very low value in practice), which corresponds to a null probability. These potentials are let unchanged in the estimation process.

### 4.5.2 *Selection of informative ML potentials*

The second phase of complexity reduction intervenes after ML parameter estimates are computed and is two-fold. First, for each clique, we store only pertinent potentials of the global estimated model $\widehat{\Psi}_{ML}$ while setting the other ones to a constant value (determined using the normalization constraint (10)). Second, we eliminate cliques which bring negligible information. This model complexity reduction can be regarded as a pruning procedure applied to the set of potentials of the ML estimate of the causal Gibbs model $\widehat{\Psi}_{ML}$. To achieve this, we resort to likelihood ratio tests which enable to specify the amount of information to be kept. For both aspects of complexity reduction, we compute the ratio of the conditional likelihood of sequence $x$ w.r.t a proposed reduced model $\Psi^*$ over the conditional likelihood of x w.r.t. $\widehat{\Psi}_{ML}$:

$$LR_x(\Psi^*, \widehat{\Psi}_{ML}) = P_{\Psi^*}(x) / P_{\widehat{\Psi}_{ML}}(x) \tag{27}$$

This ratio is compared to a user-specified threshold $\lambda_{LR}$. It indeed allows us to specify the tolerated error between the ML estimate of the Gibbs model and the

reduced model actually stored. $LR_x(\Psi^*, \widehat{\Psi}_{ML})$ can be viewed as an evaluation of the loss of information occurring if we substitute $\Psi^*$ for $\widehat{\Psi}_{ML}$.

Let us describe in more details how the incremental complexity reduction strategy is performed. As shown in equation (19), for each clique type $a$ the informative potentials of ML estimate $\widehat{\Psi}_a$ correspond to high cooccurrence values. For a given clique $a$, potentials $\widehat{\Psi}_a(\nu, \nu')$ are one by one introduced in a model $\Psi^*$ (initially, $\Psi^*_a \equiv 0$ and $\forall b \neq a, \quad \Psi^*_b = \widehat{\Psi}_b$), according to their corresponding value $\Gamma_a(\nu, \nu'|x)$ in the cooccurrence matrix, the highest values being introduced first. At each step, we compute the likelihood ratio (27). As soon as this ratio exceeds $\lambda_{LR}$, we consider the selected potentials as the representative potentials of ML estimate $\widehat{\Psi}_a$ associated to the sequence $x$. Let us note $\tilde{\Psi}$ the reduced model consisting of the potentials selected as above mentioned for each clique $a$.

Then, to select the representative cliques, we exploit the ranking over cliques $(a_1, .., a_{|\mathcal{A}|})$ defined in subsection 4.4. We consider the different reduced models $(\tilde{\Psi}^k)_{k \in \{1,..,|\mathcal{A}|\}}$ such that:

$$\begin{cases} \tilde{\Psi}^k_a = \tilde{\Psi}_a, & \forall a \in \{a_1, .., a_k\} \\ \tilde{\Psi}^k_a \equiv 0, & \forall a \in \{a_{k+1}, .., a_{|\mathcal{A}|}\} \end{cases} \tag{28}$$

We compute the likelihood ratios $LR_x(\tilde{\Psi}^k, \widehat{\Psi}_{ML})$ and stop at step $k^*$ where the ratio $LR_x(\tilde{\Psi}^{k^*}, \widehat{\Psi}_{ML})$ exceeds $\lambda_{LR}$. The corresponding reduced model $\tilde{\Psi}^{k^*}$ is finally selected as the model attached to the sequence $x$.

## 5 Motion-based video classification and retrieval

We now discuss the application of our modeling framework to motion-based video classification and retrieval issues. Considering a set of video sequences, we are interested in retrieving in this database examples similar to a video query in terms of motion content or more generally of scene activity. The general idea is to define an appropriate similarity measure between image sequences and to determine the closest matches according to this similarity measure. As far as feature-based techniques are concerned, the retrieval process generally makes use of classical distances in the feature space such as the Euclidean or Mahanolabis distances, [24, 25].

In our case, we first benefit from our statistical modeling of scene activity to define an appropriate similarity measure w.r.t. motion content. We then exploit this similarity measure to achieve a hierarchical classification over a video set. In a

third step, we tackle video retrieval with query by example formulated as a Bayesian inference issue.

### 5.1 Statistical similarity measure related to scene activity

Given video shots characterized by a statistical model of scene activity, we have to evaluate the degree of similarity of their content. We define a similarity measure inspired from Kullback-Leibler divergence [3]. Considering two distributions $\mu$ and $\mu'$, the Kullback-Leibler (KL) divergence $KL(\mu\|\mu')$ is defined by:

$$KL(\mu\|\mu') = \int \ln \frac{\mu}{\mu'} d\mu \tag{29}$$

It can be viewed as the expectation of the log-likelihood ratio $ln\,(\mu/\mu')$ w.r.t. distribution $\mu$. 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 $\Psi^n$. More precisely, for each $(k,p) \in [1,K] \times \mathcal{R}$, the transition from $x^n_{k-1}(\eta_p)$ to $x^n_k(p)$ is a sample of the causal Gibbs model $\Psi^n$. If we consider two elements of the video base $n_1$ and $n_2$, their associated models $\Psi^{n_1}$ and $\Psi^{n_2}$, and the sequences of computed motion-related quantities $x^{n_1}$ and $x^{n_2}$, the KL divergence $KL(\Psi^{n_1}\|\Psi^{n_2})$ is approximated as the empirical average of the log-ratio of the conditional likelihoods of the transitions from $x^n_{k-1}(\eta_p)$ to $x^n_k(p)$ computed respectively w.r.t. $\Psi^{n_1}$ and $\Psi^{n_2}$:

$$KL(\Psi^{n_1}\|\Psi^{n_2}) \approx \frac{1}{K|\mathcal{R}|} \sum_{k=1}^{K} \sum_{p\in\mathcal{R}} \ln \left( \frac{P_{\Psi^{n_1}}\left(x^{n_1}_k(p)|x^{n_1}_{k-1}(\eta_p)\right)}{P_{\Psi^{n_2}}\left(x^{n_1}_k(p)|x^{n_1}_{k-1}(\eta_p)\right)} \right) \tag{30}$$

Due to the causal nature of our modeling framework, this comes to approximate the KL divergence $KL(\Psi^{n_1}\|\Psi^{n_2})$ by the log-ratios of the likelihoods of the sequence of motion-related quantities $x^{n^1}$ computed respectively for the Gibbs models $\Psi^{n_1}$ and $\Psi^{n_2}$:

$$KL(\Psi^{n_1}\|\Psi^{n_2}) \quad \approx \quad \frac{1}{K|\mathcal{R}|} \ln \left( \frac{P_{\Psi^{n_1}}(x^{n_1})}{P_{\Psi^{n_2}}(x^{n_1})} \right) \tag{31}$$

It indeed quantifies the loss of information occurring when considering $\Psi^{n_2}$ instead of $\Psi^{n_1}$ to model the motion 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:

$$D(n_1,n_2) = \frac{1}{2}\left[KL(\Psi^{n_1}\|\Psi^{n_2}) + KL(\Psi^{n_2}\|\Psi^{n_1})\right] \tag{32}$$

It should be noticed that this similarity measure is not a metric since it does not verify the triangular inequality. However, it can be easily computed and interpreted, since it is expressed as a logarithm of a likelihood ratio.

### 5.2 Hierarchical motion-based indexing and retrieval

In case of large databases, it is obviously relevant to appropriately structure the considered video set. We focus here on hierarchical representations which have been successfully exploited for still image bases with a view to tackling browsing or retrieval issues [37, 25, 33, 7]. Such indexing structures rely on binary trees. The tree nodes will correspond to subsets of shots of the processed video base. To achieve this hierarchical structuring, either top-down [33] or bottom-up [25] strategies can be adopted. As pointed out in [7], bottom-up techniques seem to offer better performance in terms of classification accuracy. In fact, since top-down methods consist in successively splitting the nodes of the tree from the root to the leaves, an element misclassified at the top of the hierarchy will appear in an undesirable branch of the final binary tree. Therefore, we retain bottom-up clustering and more particularly, we consider an ascendant hierarchical classification procedure, [10].

We also need to define the similarity measure $D$ between clusters of videos. For two clusters $C^1$ and $C^2$, $D$ is defined by:

$$D(C^1, C^2) = \max_{(n_1,n_2)\in C^1\times C^2} D(n_1, n_2) \tag{33}$$

We can now construct an ascendant hierarchical classification based on $D$. It proceeds incrementally as follows. At a given iteration, 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. $\min_{C'\neq C} D(C, C') > D_{max}$, it is kept alone to form a single cluster. $D_{max}$ is a given threshold. For two clusters $C_1$ and $C_2$, $\exp[-D(C_1, C_2)]$ can be expressed as the average of two likelihood ratios comprised in $[0, 1]$ (relation 33). Therefore, we set as $D_{max} = -\ln\tau$ where $\tau$ is a threshold in $[0, 1]$. Threshold $\tau$ quantifies the information loss we tolerate in terms of accuracy of description of motion distributions when substituting models attached to $C_2$ for those attached to $C_1$, and conversely (Typically, $\tau = 0.1$). The merging procedure is performed from the leaves and iterated until no new cluster can be built. A leave of the tree is created for each element of the considered video base.

For retrieval purpose, a scene activity model has to be attached to each created cluster. In the case of the simple temporal model, since the model is directly determined from temporal cooccurrence measurements, the model associated to the

merging of two clusters can be straightforwardly estimated using relation (21). Indeed, when considering the set of sequences comprised in th newly created cluster, the corresponding cooccurrence measurements are directly determined as the sum of the cooccurrence measurements computed for each individual sequence of the considered cluster. When merging two clusters $C_1$ and $C^2$, we first compute the cooccurrence matrix $\Gamma(C^1, C^2)$ as the sum of the cooccurrence matrices $\Gamma(C^1)$ and $\Gamma(C^2)$, and second, exploiting relation (21), we estimate the potentials of the Gibbs model associated to the new cluster formed by the union of $C_1$ and $C_2$. Otherwise, when coping with the extended temporal Gibbs models, such a straightforward updating is no more available. We do not estimate the model associated with the union of the $x$ sequences to save computation, and we prefer to select either $\Psi^{C_1}$ or $\Psi^{C_2}$ as the model representative of the new cluster resulting from the merging of nodes $C_1$ and $C_2$. We keep the model which maximizes the conditional likelihood computed for the motion-related quantity sequence issued from the union of all the elements of the two clusters.

### 5.3 Bayesian retrieval

As in [35], the retrieval process is formulated as a Bayesian inference issue. Given a video query $q$, we aim at determining the best match $d^*$ in the stored set $\mathcal{D}$ of video sequences according to the MAP criterion:

$$d^* = \arg\max_{d \in \mathcal{D}} P(d|q) = \arg\max_{d \in \mathcal{D}} P(q|d)P(d) \tag{34}$$

The distribution $P(d)$ allows us to formulate *a priori* knowledge on the video content relevance over the database. It can be inferred from semantical descriptions attached to each type of video sequences. It could also exploit relevance feedback during the retrieval process [26]. Indeed, the likelihood of the different possible replies could be weighted according to some evaluation of former retrieval operations performed by the user. In the remainder, we will in fact incorporate no *a priori* ($P(d)$ distribution is taken uniform).

Furthermore, criterion (34) also supplies a ranking of the elements $\{d\}_{d \in \mathcal{D}}$ according to $P(q|d)P(d)$, which quantifies how relevant is the selection of $d$ w.r.t. the motion content of query $q$. In our case, to each element $d$ of the database is attached a causal Gibbsian model $\Psi^d$. We compute the sequence of motion-related measurements $x^q$ for video query $q$ and the conditional likelihood $P(q|d)$ is expressed using $P_{\Psi^d}$. Then, we can infer:

$$n^* = \arg\max_{n} P_{\Psi^d}(x^q) \tag{35}$$

Let us stress that we do not need to estimate a model for the query.

In addition, we can take advantage from an hierarchical representation of the video base described in the previous section to satisfy a video query. When dealing with large databases, solving criterion (35) in an exhaustive way reveals quite time consuming. Therefore, we exploit the constructed binary tree to obtain a suboptimal but efficient solution of criterion (35). If the convergence to the best match is not guaranteed, this can be viewed as a trade-off between reply accuracy and search complexity. The retrieval process is carried out through the binary tree from the root to the leaves as follows. As initialization, we select the best node $C^0$ at the root $\mathcal{T}_{root}$ of the search tree according to:

$$C^0 = \arg \max_{C \in \mathcal{T}_{root}} P_{\Psi^C}(q) \tag{36}$$

At each step $k$, given a parent cluster $C^k$, we select the best child node $C^{k+1}$ according to the MAP criterion:

$$C^{k+1} = \arg \max_{C \in C^k} P_{\Psi^C}(x^q) \tag{37}$$

This procedure is iterated until a given maximal number of answers or a given precision is reached.

# 6 Results

We have evaluated the whole proposed framework for scene activity modeling, content-based video indexing and retrieval, on a database containing samples of real videos. We have paid a particular attention to choose examples representative of various motion situations. The database includes temporal textures (samples of fire and sequences of river), video shots exhibiting an important scene activity such as sport video (basket, horse riding,...), rigid motion situations (cars, train, ...), and sequences with a low motion activity. We have built a database of 150 sequences of 10 images issued from 70 video shots. In addition, elements issued from the same video shot are not temporally adjacent.

The experiments reported in this Section have been performed using values of parameters set as follows. In the stage concerned with motion-related measurements, we fix $V_{max} = 4.0$ and $|\Lambda| = 16$. This appears relevant and accurate enough in previous work [11, 14, 4]. For model complexity reduction, we set $\lambda_{LR} = 0.99$. At last, $D_{max} = 2.3$ ($\tau = 0.1$), in the hierarchical classification stage.

### 6.1 Model complexity reduction

In a first step, we have estimated the causal Gibbs model attached to each element of the database for the neighborhood $\eta^5$ (see Section 4). For the considered database, we finally kept from only 5% to 20% of the 1280 potentials (here, $|\Lambda| = 16$, $|\mathcal{A}| = 5$ and $16^2 \times 5 = 1280$) of each ML model attached to each video shot after the model complexity reduction phase. We report two examples of model complexity reduction respectively for shot $anchor_1$ and shot $basket_1$ with the temporal neighborhood $\eta^5$. The median images of these two sequences are displayed in Fig.4. Video $anchor_1$ is a static of an anchor in a news program. The scene activity is very weak and only Gibbs potentials related to low values of motion magnitude are kept. This leads to select 5% of the estimated ML Gibbs potentials and one clique over the five initial ones. The second example $basketball_3$ involves important scene activity. The stored Gibbs model is more complex with two selected cliques and 10% of the estimated potentials kept.

### 6.2 Statistical hierarchical motion-based classification

To provide a comprehensive visualization of the statistical hierarchical motion-based classification described in Section 5, we have performed a classification on a subset of 20 sequences displayed in Fig.4. It contains two static shots of anchors, $anchor_1$ and $anchor_2$, from news program involving a very weak scene activity. Two other examples of low motion activity, *hall* and *Concorde*, are included. Four examples of rigid motion situations are introduced corresponding to road traffic sequences, $highway_1$ and $highway_2$, and to airport sequences, *landing* and $take-off$. Ten sport video sequences are added involving shots of rugby game, $rugby_1$ and $rugby2_2$, hockey game, $hockey_1$, $hockey_2$, and $hockey_3$, basketball game, $basketball_1$, $basketball_2$ and $basketball_3$, and windsurfing, $windsurfing_1$ and $windsurfing_2$. At last, two samples of temporal textures with high scene activity, $fire$ and $river$, are also considered.

For this experiment, we exploit extended temporal models corresponding to $\eta^5$. The obtained unsupervised hierarchical classification, shown in Fig.5, correctly separates the different kinds of dynamic contents. Traffic sequences, $road_1$ and $road_2$, airport videos, *landing* and $take-off$, and low motion activity situations, $anchor_1$, $anchor_2$, *hall* and *Concorde*, constitute a separate cluster in which relevant subclusters have been created associated to these two types of motion content. In addition, all sport video shots are properly grouped. In this last group, pertinent subgroups have also been identified such as the one comprising the three basketball sequences displaying very high motion activity, and the one with the three hockey shots.

| | number of samples | I | II | III | IV |
|---|---|---|---|---|---|
| I | 19 | 100 % | 0 % | 0 % | 0 % |
| II | 18 | 0 % | 89 % | 0 % | 11 % |
| III | 71 | 0 % | 1 % | 94 % | 5 % |
| IV | 6 | 0 % | 0 % | 0 % | 100 % |

Table 1: Tab. 1. *Evaluation of the performance of the retrieval system w.r.t. an* a priori *classification of the video base of* 150 *elements. Class (I) refers to weak scene activity, class (II) to rigid motion situations, class (III) to wide-angle shots and close shots of sport videos, class (IV) to temporal texture samples. We supply the classification rates for the second retrieved answer obtained when considering in turn each element of the base as a query.*

### 6.3 Statistical motion-based retrieval with query by example

For the retrieval experiments performed over the base of 150 videos, we have considered simple temporal models with neighborhood $\eta^1$. Fig.6 reports four experiments of retrieval operations with query by video example. The first query is a news program which consists in a static shot on an anchor. A rigid motion situation (plane take-off) is proposed as the second query. The third and fourth retrieval operations concern sport videos. The third query is a close shot on a basketball player tracked by the camera during the shot, whereas the camera delivers a global view of the game field in the last example. We deliver the three best replies according to the computed log-likelihood values $P_{\Psi^C}(x^q)$ (as given in relation (37)). For all the considered queries, the retrieval process supplies quite relevant replies. In particular, when considering the two examples involving sport videos with an important motion activity, the close-up situation is well discriminated from the other ones. To *a posteriori* evaluate the relevance of the replies, we have also estimated the model $\Psi^q$ associated to the query $q$ and we report the values of the distance $D(q, n)$ given by relation (32) between $\Psi^q$ and the different retrieved models $\Psi^n$. The ranking supplied by log-likelihood values is confirmed by the values of distance $D$ for each reply.

To carry out a more quantitative evaluation of our motion-based retrieval system, we have analyzed the relevance of the replies retrieved when considering in turn each element of the video base as a query. To this end, we need to define *a priori* classes w.r.t. motion content. We consider four classes which seemed to be relevant

as illustrated by the classification experiment reported in Fig.5. More precisely, class (I) refers to weak scene activity contents, class (II) to rigid motion situations, class (III) to wide-angle shots and close shots of sport games, class (IV) to temporal texture samples. It should be stressed that quantitatively evaluating a system devoted to query by example according to semantic classes necessarily remains partial and somewhat subjective. For evaluation purpose, we consider two measures. First we count how many times the query shot appears as the best answer. Let us note that this is not guaranteed *a priori* since the retrieval process is conducted through the hierarchical representation of the base and not according to an exhaustive search. For the processed video base, the first retrieved answer is indeed the query shot with a rate of 76% (within the remaining 24%, i.e. 36 video samples, the best reply belongs to the same *a priori* class for 30 queries). Secondly, we have evaluated the relevance of the second retrieved answer in terms of correct classification w.r.t. the *a priori* scene activity classes described above. The obtained results with simple temporal Gibbs models are given in Table 1. Mainly, for classes (I), (II), (III) and (IV), the rate of correct classification is within $[89\%, 100\%]$, which appears as quite promising. These results also reveal the limitations of the evaluation of our retrieval system involving query by example w.r.t. semantic *a priori* classes. For instance, we obtain a misclassification rate of 16% for the class (II) related to rigid motion situations. The corresponding video shots indeed involve rigid objects close to the camera and undergoing large displacements. Thus, they could appear as more similar to the close shots of sport games than to rigid motion situations such as the traffic sequences involved in the classification experiments reported in Fig.5. However, this evaluation should be considered as a first validation of our approach> We plan to evaluate it on a larger database.

# 7 Conclusion

We have described an original method for the global characterization of motion content in video sequences able to handle a very large range of dynamic scene contents. We have proposed a general statistical framework for video classification, indexing and retrieval with query by example. It relies on a statistical modeling of the distribution of local motion-related measurements using causal Gibbs models estimated using the ML criterion. Besides, we have designed an efficient model complexity reduction scheme based on likelihood ratios. This statistical modeling leads to a general statistical framework for motion-based hierarchical classification of a video database and motion-based retrieval with query by example according to the MAP

criterion. We have obtained promising results both for the classification stage and the retrieval process on a video database involving various types of motion content and scene activity.

In future work, we plan to validate our approach on a still larger video base. In that context, as pointed out in [7], the hierarchical indexing structure can be regarded as a relevant alternative to retrieval with query by example, since it allows users to navigate the database according to their interest. Multiscale causal Gibbs model will be also investigated. Ongoing work aims at exploiting this novel approach for motion modeling and characterization to automatically segment entities of interest in the shot and to satisfy partial queries [12]. It could be also useful to extract shots of interest in video sequences with a view to creating video summaries.

## Acknowledgments

The authors are thankful to INA, Département Innovation, Direction de la Recherche, for providing the MPEG-1 news sequences, which are excerpts of the INA/GDR-ISIS video corpus, and to MIT for supplying the sequences of temporal textures $fire$ and $river$.

## A Uniqueness of the causal Gibbs random field modeling

We will demonstrate that the constraint defined by relation (10) guarantees the uniqueness of the potentials associated to a model. We indeed show that the considered causal Gibbs model is defined up to additive constants and that criterion (10) leads to fix these constants. In addition to the notation defined in Section 4, let us denote $\mathcal{X}$ the space of the sequences of motion-related quantities. Let us consider two causal Gibbs models $\Psi^1$ and $\Psi^2$ verifying: $\forall x \in \mathcal{X},\ \ P_{\Psi^1}(x) = P_{\Psi^2}(x)\ \ (A)$.

We want to show that: $\exists \lambda \in \mathbb{R}^{|\mathcal{A}|\cdot|\Lambda|},\ \ \forall (a,\nu,\nu') \in \mathcal{A}\times\Lambda^2,\ \ \Psi^1_a(\nu,\nu') = \Psi^2_a(\nu,\nu') + \lambda_{(a,\nu')}$. From statement (A), given $(\alpha,\beta) \in \Lambda \times \Lambda^{|\mathcal{A}|}$, it is obvious that $\forall (k,r) \in [1,K]\times\mathcal{R}$:

$$P_{\Psi^1}\left(X_k(p) = \alpha | X_{k-1}(\eta_p) = \beta\right) = P_{\Psi^2}\left(X_k(p) = \alpha | X_{k-1}(\eta_p) = \beta\right) \qquad (38)$$

Using expression (8), we obtain that $\forall(\alpha,\beta) \in \Lambda \times \Lambda^{|\mathcal{A}|}$, $\exists \xi(\beta) \in \mathbb{R}$ such that:

$$\begin{aligned} \exp\left[\sum_{a\in\mathcal{A}} \Psi_a^1(\alpha,\beta_a) - \Psi_a^2(\alpha,\beta_a)\right] &= \left[\sum_{\nu\in\Lambda} \exp \sum_{a\in\mathcal{A}} \Psi_a^1(\nu,\beta_a)\right] / \left[\sum_{\nu\in\Lambda} \exp \sum_{a\in\mathcal{A}} \Psi_a^2(\nu,\beta_a)\right] \\ &= \xi(\beta) \end{aligned} \tag{39}$$

where $\xi(\beta) = \{\xi_a(\beta)\}_{a\in\mathcal{A}}$. Let us define $\Delta\Psi = \Psi^1 - \Psi^2$ , $\alpha^*$ a particular value in $\Lambda$, and, for $(\alpha,\beta) \in \Lambda \times \Lambda^{\mathcal{A}}$, $f_{(\alpha,\beta)} = \{\delta(\nu-\alpha)\cdot\delta(\nu'-\beta_a)\}_{(a,\nu,\nu')\in\mathcal{A}\times\Lambda^2}$, and $\Delta f_{(\alpha,\beta)} = f_{(\alpha,\beta)} - f_{(\alpha^*,\beta)}$. From relation (39), we infer that $\forall(\alpha,\beta) \in \Lambda \times \Lambda^{|\mathcal{A}|}$:

$$\begin{aligned} \left\langle \Delta\Psi, \Delta f_{(\alpha,\beta)} \right\rangle &= \sum_{(a,\nu,\nu')\in\mathcal{A}\times\Lambda^2} \Delta\Psi_a(\nu,\nu') \cdot \left(f_{(\alpha,\beta)}(a,\nu,\nu') - f_{(\alpha^*,\beta)}(a,\nu,\nu')\right) \\ &= \ln\xi(\beta) - \ln\xi(\beta) = 0 \end{aligned} \tag{40}$$

where $\langle\cdot,\cdot\rangle$ is a dot product. As a consequence, $\Delta\Psi$ is in the subspace of dimension $|\mathcal{A}|\cdot|\Lambda|$ orthogonal to the space spanned by $\left\{\Delta f_{(\alpha,\beta)}\right\}_{(\alpha,\beta)\in\Lambda\times\Lambda^{\mathcal{A}}}$. The set $\left(g_{(a_0,\nu_0')}\right)_{(a_0,\nu_0')\in\mathcal{A}\times\Lambda}$ defined by:

$$\forall(a,\nu,\nu') \in \mathcal{A}\times\Lambda^2, \quad g_{(a_0,\nu_0')}(a,\nu,\nu') = \delta(a-a_0)\cdot\delta(\nu'-\nu_0') \tag{41}$$

is a base of this subspace. Then, $\Delta\Psi$ can be expressed as a linear combination of $\left(g_{(a_0,\nu_0')}\right)_{(a_0,\nu_0')\in\mathcal{A}\times\Lambda}$ i.e. $\exists\lambda \in \mathbb{R}^{|\mathcal{A}|\cdot|\Lambda|}$ such that:

$$\forall(a,\nu,\nu') \in \mathcal{A}\times\Lambda^2, \quad \Delta\Psi_a(\nu,\nu') = \sum_{(a_0,\nu_0')\in\mathcal{A}\times\Lambda} \lambda_{(a_0,\nu_0')} \cdot g_{(a_0,\nu_0')}(a,\nu,\nu') = \lambda_{(a,\nu')} \tag{42}$$

Therefore, considering the normalization constraint (10) comes to fix quantities $\left(\lambda_{(a,\nu')}\right)_{(a,\nu')\in\mathcal{A}\times\Lambda}$ and guarantees to uniquely define the potentials $\{\Psi_a(\nu,\nu')\}_{(a,\nu,\nu')\in\mathcal{A}\times\Lambda^2}$ associated to model $\Psi$■

# 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.


Figure 4: *Set of the 20 video shots used to supply an example of motion-based hierarchical classification given in Fig.5. For each video, we display the median image of the shot.*

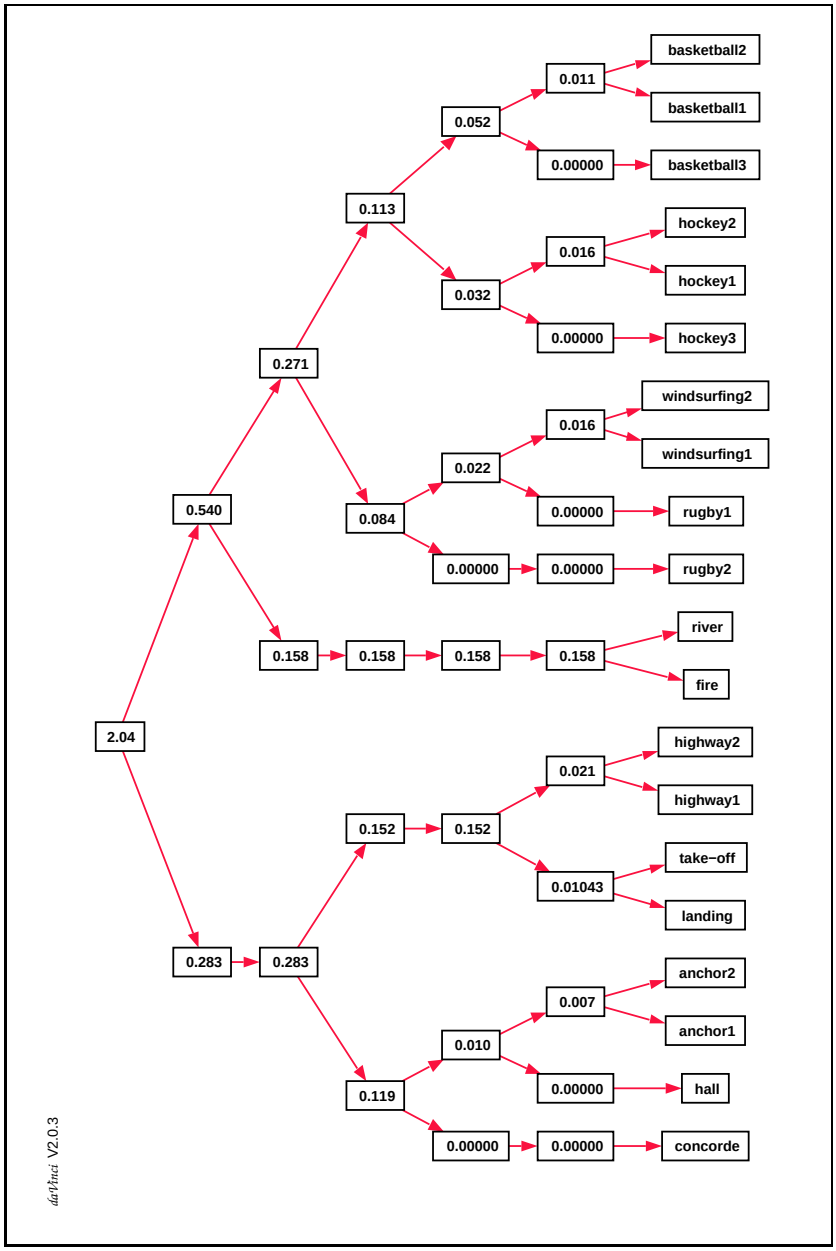
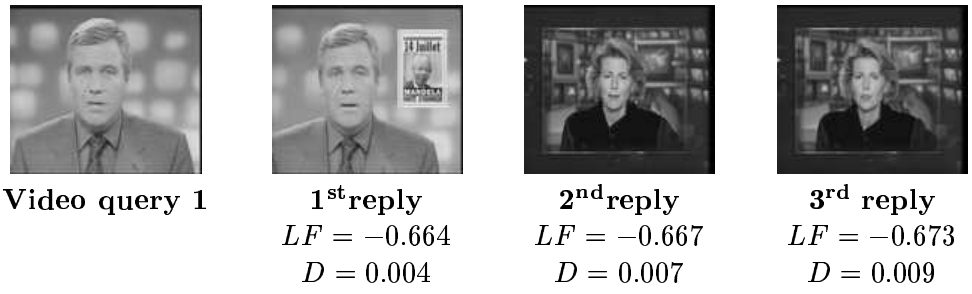



Figure 5: **Motion-based statistical classification:** *obtained motion-based hierarchical classification for the set of 20 video sequences presented in Fig.4 with* $D_{max} = 2.3$ *(*$\tau = 0.1$*). At each leave of the tree, we report the name of the video sequence. For the other nodes of the tree, we display the maximum intra-cluster distance evaluated using expression* $D$ *of relation (33).*



| | | | |
|---|---|---|---|
| **Video query 1** | **1st reply** $LF = -0.664$ $D = 0.004$ | **2nd reply** $LF = -0.667$ $D = 0.007$ | **3rd reply** $LF = -0.673$ $D = 0.009$ |
| **Video query 2** | **1st reply** $LF = -1.052$ $D = 0.008$ | **2nd reply** $LF = -1.054$ $D = 0.012$ | **3rd reply** $LF = -1.057$ $D = 0.016$ |
| **Video query 3** | **1st reply** $LF = -1.411$ $D = 0.008$ | **2nd reply** $LF = -1.414$ $D = 0.014$ | **3rd reply** $LF = -1.419$ $D = 0.015$ |
| **Video query 4** | **1st reply** $LF = -2.125$ $D = 0.008$ | **2nd reply** $LF = -2.126$ $D = 0.008$ | **3rd reply** $LF = -2.131$ $D = 0.009$ |

Figure 6: **Results of retrieval operations involving three replies.** *For each reply $n$, we give the value $LF$ of the log-likelihood* $\ln\left(P_{\Psi^n}(x^q)\right)$ *corresponding to video query $q$. To* a posteriori *evaluate the relevance of the replies, we have also estimated the model $\Psi^q$ associated to the query $q$ and we report the values of the distance $D(n,q)$, given by relation (32) between $\Psi^q$ and the different retrieved models $\Psi^n$.*

[2] E. Ardizzone and M. La Cascia. Automatic video database indexing and retrieval. *Multimedia Tools and Applications*, 4(1):29–56, 1997.

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

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

[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. *Jal of Visual Communication and Image Representation*, 10(2):78–112, 1999.

[7] J.-Y. Chen, C. A. Bouman, and J. C. Dalton. Hierarchical browsing and search of large image databases. *IEEE Trans. on Image Processing*, 9(3):442–455, 2000.

[8] J.D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–625, April 1997.

[9] 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.

[10] E. Diday, G. Govaert, Y. Lechevallier, and J. Sidi. Clustering in pattern recognition. In *Digital Image Processing*, pages 19–58. J.-C. Simon, R. Haralick, eds, Kluwer edition, 1981.

[11] 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.

[12] R. Fablet and P. Bouthemy. Statistical motion-based retrieval with partial query. In *Proc. of 4th Int. Conf. on Visual Information Systems, VISUAL'2000*, Lyon, November 2000.

[13] 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.

[14] 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.

[15] A. Muffit Ferman, A. Murat Tekalp, and R. Mehrotra. Effective content representation for video. In *Proc. of 5th IEEE Int. Conf. on Image Processing, ICIP'98*, pages 521–525, Chicago, October 1998.

[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] 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.

[18] G.L. Gimel'Farb. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(11):1110–1114, November 1996.

[19] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using markov random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(2):1217–1232, 1993.

[20] B. Horn and B. Schunk. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[21] M. Irani and P. Anandan. Video indexing based on mosaic representation. *Proc. of the IEEE*, 86(5):905–921, May 1998.

[22] 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.

[23] A. Jain and D. Zongker. Feature selection: evaluation, application and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.

[24] A.K. Jain, A. Vailaya, and W. Xiong. Query by video clip. *Multimedia Systems*, 7(5):369–384, 1999.

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

[26] 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.

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

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

[29] 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.

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

[31] Y. Rui, T. Huang, and S. Mehrota. Constructing table-of-content for videos. *Multimedia Systems*, 5(7):359–368, September 1999.

[32] S. Santini and R. Jain. Similarity measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[33] H. Schweitzer. Organizing image databases as visual-content search trees. *Image and Vision Computing*, 17:501–511, 1999.

[34] M. Szummer and R.W. Picard. Temporal texture modeling. In *Proc. of 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, pages 823–826, Lausanne, September 1996.

[35] 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.

[36] 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.

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

[38] L. Younes. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Poincaré*, 24(2):269–294, 1988.

[39] H.J. Zhang, J. Wu, D. Zhong, and S. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, April 1997.

[40] S.C. Zhu, T. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling. *Int. Journal of Computer Vision*, 27(2):107–126, 1998.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399