



Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval

Ronan Fablet, Patrick Bouthemy, Patrick Pérez

► To cite this version:

Ronan Fablet, Patrick Bouthemy, Patrick Pérez. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. IEEE Transactions on Image Processing, 2002, 11 (4), pp.393 - 407. <10.1109/TIP.2002.999674>. <hal-02283295>

HAL Id: hal-02283295

<https://hal.science/hal-02283295v1>

Submitted on 16 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.



HAL Authorization

Non-Parametric Motion Characterization using Causal Probabilistic Models for Video Indexing and Retrieval

R. Fablet¹ P. Bouthemy^{2*}

¹IRISA/CNRS ²IRISA/INRIA

Campus universitaire de Beaulieu

35042 Rennes Cedex, France

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

P. Pérez^{2,3}

³Microsoft Research

St. George House, 1 Guildhall St.

Cambridge, CB2 3NH, UK

e-mail: pperez@microsoft.com

*Corresponding author

Abstract— This paper describes an original approach for content-based video indexing and retrieval. We aim at providing 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 derived from the spatio-temporal derivatives of the intensity function. These distributions are then represented by causal Gibbs models. To be independent of camera movement, the motion-related measurements are computed in the image sequence generated by compensating the estimated dominant image motion in the original sequence. The statistical modeling framework considered makes the exact computation of the conditional likelihood of a video shot belonging to a given motion or more generally to an activity class feasible. 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 database according to motion content similarity. This results in a binary tree where each node is associated to an estimated causal Gibbs model. 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.

Keywords— Non-parametric motion analysis, video databases, motion-based indexing, query-by-example, statistical modeling, maximum likelihood estimation, spatio-temporal cooccurrences

I. 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, etc.). An entirely manual annotation of visual documents is no longer able to cope with the rapidly increasing amount of these video data. In addition, the efficient use of these databases requires a reliable and relevant means to access visual information. This implies indexing and retrieving visual documents by their content. A great deal of research is currently devoted to image and video database management [1], [5], [10]. Nevertheless, it remains hard to 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 navigation and browsing within videos [8], [12], with the classification of video sequences into different genres (sports, news, movies, commercials, documentaries, etc.) [44], with the retrieval of examples similar to a given video query [17], [20], [30], or with high-level video structuring such as macro-segmentation [38], [45]. Such applications require the combination of content-based video descriptions 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 ele-

mentary shots [5], [9], [47]¹. This stage is usually associated with the recognition of typical forms of video shooting such as static shot, panning, traveling or zooming [9]. 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. Two types of approaches are usually considered for characterizing dynamic content in video sequences. A first class of approaches, based on parametric or dense motion field estimation, includes image mosaicing [22], [27], segmentation, tracking and characterization of moving elements in order to determine a spatio-temporal representation of the video shot [13], [21], [22]. The description of the motion content may then rely on the extraction of pertinent qualitative features of the entities of interest, such as the direction of the displacement [22], or on the analysis of the trajectories of the center of gravity of the tracked objects [14]. However, these techniques turn out to be unsuitable for certain classes of sequences with complex dynamic contents such as the motion of rivers, flames, foliage in the wind, crowds, etc. Furthermore, as far as video indexing is concerned, the entities of interest may not be single objects but rather groups of objects, particularly when dealing with sport videos. No tool currently exists to automatically extract these kinds of entities. Therefore, in the context of video indexing, it seems appropriate to adopt a global point of view that avoids any explicit motion segmentation step.

The unsuitability of parametric or dense motion field estimation leads us to consider a second category of methods for motion-based video indexing and retrieval. The 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 works in this direction have led to the extraction of “temporal texture” features, [7], [17], [34], [37], [42]. Motion of rivers, foliage, flames, or crowds, for instance, can indeed be regarded as temporal textures. In [37], temporal texture features are extracted from the description of surfaces generated by spatio-temporal trajectories. In [34], 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 [7], [17], we considered global features extracted from temporal cooccurrence distributions of local motion-related measurements, which 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. This modeling allows us to design an original, coherent and efficient

¹Henceforth, for convenience, the term “sequence” will be used to designate an elementary shot.

framework for both motion-based video indexing and motion-based video retrieval.

The remainder of the paper is organized as follows. Section II outlines the general ideas underlying our work. Section III describes the local non-parametric motion-related measurements that we use. In Section IV, we introduce our method based on the statistical modeling of the spatio-temporal distribution of the motion-related quantities computed from a video sequence and the associated estimation scheme. Section VI 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 tools are then exploited to satisfy queries by example within a statistical framework. In Section VII, we report experimental results of video classification and retrieval examples over a set of video sequences. Section VIII contains concluding remarks.

II. 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 [5]. Usually, shot content characterization relies on the extraction of a set of numerical features or descriptors, and the comparison of shot content is performed in the feature space according to a given distance such as the Euclidean distance or more elaborate measures [39]. 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 [31] or some other feature selection techniques [29]. Unfortunately, the feature space is usually of high dimension, and the distance metric used is likely not to properly capture the uncertainty attached to feature measurements. Consequently, statistical methods may be a more suitable approach, as 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 [43], modeling of DCT coefficients by Gaussian distribution mixtures is exploited for image texture indexing and the retrieval operation is formulated in a Bayesian framework w.r.t. the MAP (Maximum A Posteriori) 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 for video retrieval with query-by-example. To this end, we have designed a motion classification (or, more generally, motion 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. In recent works [24], [48], 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 readily computed in this context, this allows us to develop a general and efficient statistical framework for video indexing and retrieval with query-by-example.

III. 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. To this end, 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.

A. Dominant motion estimation

To model the transformation between two successive images, we consider a 2D affine motion model. A possible alternative is a 2D quadratic model involving eight parameters, i.e. corresponding to the 3D rigid motion of a planar surface. However, it is computationally more demanding, while not being significantly more suitable in most situations. The displacement $\mathbf{w}_\Theta(p)$, at pixel p , related to the affine motion model parameterized by Θ is given by:

$$\mathbf{w}_\Theta(p) = \begin{pmatrix} a_1 + a_2x + a_3y \\ a_4 + a_5x + a_6y \end{pmatrix} \quad (1)$$

with $p = (x, y)$ and $\Theta = [a_1 \ a_2 \ a_3 \ a_4 \ a_5 \ a_6]$. The estimation of the dominant parametric motion model is achieved with the gradient-based multi-resolution incremental method described in [35]. The following minimization problem is solved:

$$\hat{\Theta} = \arg \min_{\Theta} \sum_{p \in \mathcal{R}} \rho(DFD(p, \Theta)) \quad (2)$$

where $DFD(p, \Theta) = I_{t+1}(p + \mathbf{w}_\Theta(p)) - I_t(p)$, with I_t being the intensity function in the image at time t , is the “displaced frame difference”, \mathcal{R} is the image grid, and ρ is a robust M-estimator (here the Tukey biweight function). The use of a robust estimator ensures the dominant image motion estimation is not sensitive to secondary motions due to mobile objects in the scene. Minimization (2) is conducted by an iterative reweighted least-square technique embedded in a multiresolution framework and involving appropriate successive linearizations of the DFD expression [35].

B. 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. Dense optic flow fields provide such local information [32], [41] and have been exploited for feature-based video retrieval [2], [30]. However, as stressed above, the accuracy and relevance of the estimation cannot always be guaranteed in complex motion situations and, the computational load required 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 compensated sequence.

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 [26]:

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

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

the residual normal velocity $v_n^*(p)$ in the motion compensated sequence at pixel p by:

$$v_n^*(p) = \frac{-1}{\|\nabla I^*(p)\|} \frac{\partial I^*(p)}{\partial t}. \quad (4)$$

Temporal derivative $\frac{\partial I^*(p)}{\partial t}$ 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 related to the computation of the intensity derivatives.

As pointed out in [3], [36], the norm of the spatial image gradient $\|\nabla I^*(p)\|$ 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 can be used as a relevant motion-related quantity. More precisely, we consider the following expression :

$$v_{obs}(p) = \frac{\sum_{q \in \mathcal{F}(p)} \|\nabla I^*(q)\|^2 |v_n^*(q)|}{|\mathcal{F}(p)| \max(G^2, G_{moy}^2(p))} \quad (5)$$

where $\mathcal{F}(p)$ is a small window centered on p , $|\mathcal{F}(p)|$ its size and $G_{moy}(p)$ the square root of the average of the squared magnitude of the spatial gradient within window $\mathcal{F}(p)$:

$$G_{moy}(p) = \sqrt{\frac{1}{|\mathcal{F}(p)|} \sum_{q \in \mathcal{F}(p)} \|\nabla I^*(q)\|^2}. \quad (6)$$

G is 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 [19], [28], [36] and motion-based video indexing and retrieval using feature extraction [7], [17]

We also have to cope with the limitations of the gradient-based image motion constraint (3). This relation is not valid in occluded regions, over motion discontinuities, and even on sharp intensity discontinuities. In addition, it cannot handle large displacements. Therefore, we adopt a multiscale strategy to compute $v_{obs}(p)$ at a reliable scale and we use an appropriate test to validate its applicability. More precisely, we build a Gaussian pyramid of the video frame in consideration and the succeeding one. At each pixel p , we determine the lowest scale for which the image motion constraint (3) is valid using the statistical test described in [25]. 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 .

The expression described above for computing $v_{obs}(p)$ ignores information related to motion direction, which prevents us from discriminating, for instance, two opposite 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 that these continuous variables are quantized. By definition, the quantities

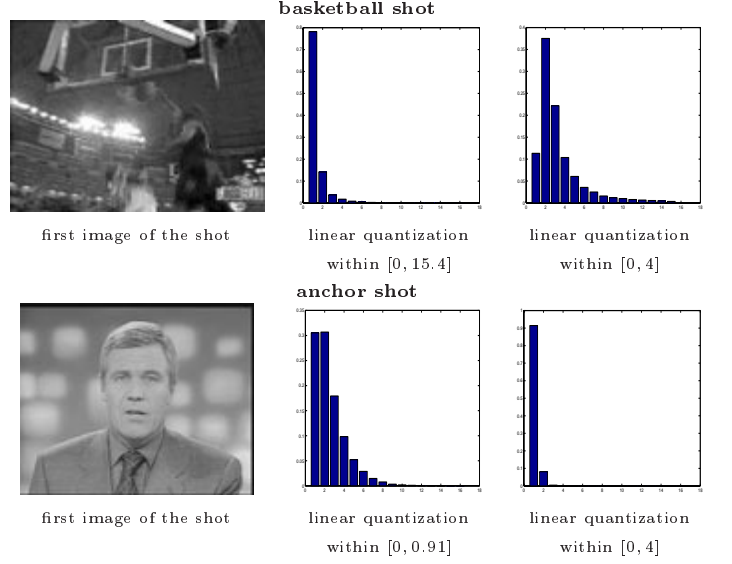


Fig. 1. **Quantization of motion-related measurements** $\{v_{obs}(p)\}_{p \in \mathcal{R}}$. We display two examples of quantization of the motion-related quantities for a basketball shot and an anchor shot. The first column depicts the first image of the processed shot; the middle one the histogram resulting from a linear quantization of $\{v_{obs}(p)\}_{p \in \mathcal{R}}$ 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.

$\{v_{obs}(p)\}_{p \in \mathcal{R}}$ are positive and, for a given pixel p , $v_{obs}(p)$ is theoretically less or equal 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 $v_{obs}^{max} = \max_{p \in \mathcal{R}} v_{obs}(p)$. However, we

would face two main problems. First, since we aim at evaluating content similarity between video shots, a range of quantized motion-related quantities common 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 used, because 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 in specific situations where the validity test happens to fail. Although $v_{obs}^{max} = 15.4$ in the first sequence of Fig.1, it appears that the really informative part of the histogram is retained within the range $[0, 4]$. Therefore, we prefer to consider a linear quantization within a predefined interval $[0, V_{max}]$. Applying this quantization scheme, the direct comparison 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 Λ the discretized range of variations for $\{v_{obs}(p)\}_{p \in \mathcal{R}}$. Henceforth, we denote x_k the set of the quantized motion-related measurements for the k th frame of the video sequence.

IV. CAUSAL SPATIO-TEMPORAL GIBBS MODELS

A. Causal Gibbs random fields

This Section is concerned with the description of 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 II, we consider Gibbs models expressed in terms of cooccurrences. We previously exploited cooccurrence statistics for video indexing in [7], [17]. We have investigated causal probabilistic models for two reasons. Firstly, the corresponding likelihood functions can be exactly computed (including normalization constants), which in turn allows us to properly define a motion-based video similarity measure. Whereas the exact computation of likelihood functions is generally intractable with classical spatial Markov random fields [23] due to the unknown partition function, it can be readily obtained with most causal models. Secondly, we are concerned with the characterization of sequences of maps of motion-related measurements. The evolution of the content of such maps is by nature causal along the time axis. Therefore, it seems pertinent to design a temporally causal modeling of motion information. It enables to handle temporal non-stationarities while being sufficient to discriminate motion classes of interest.

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

$$P_{\mathcal{M}}(x) = P_{\mathcal{M}}(x_0) \prod_{k=1}^K P_{\mathcal{M}}(x_k | x_{k-1}) \quad (7)$$

where \mathcal{M} refers to the underlying model to be explicitly defined later. $P_{\mathcal{M}}(x_0)$ represents the *a priori* distribution for the first map of the sequence. In practice, we will consider no specific prior, i.e., $P_{\mathcal{M}}(x_0)$ is uniform. In addition, we assume that the random variables $(X_k(p))_{p \in \mathcal{R}}$ at time k are conditionally independent given X_{k-1} and that for each of them, the conditioning w.r.t. X_{k-1} reduces to a small subset of measurements around the location under concern. Thus, we assume that conditional probabilities $P_{\mathcal{M}}(x_k | x_{k-1})$ factorize as:

$$\begin{aligned} P_{\mathcal{M}}(x_k | x_{k-1}) &= \prod_{p \in \mathcal{R}} P_{\mathcal{M}}(x_k(p) | x_{k-1}) \\ &= \prod_{p \in \mathcal{R}} P_{\mathcal{M}}(x_k(p) | x_{k-1}(\eta_p)) \end{aligned} \quad (8)$$

where \mathcal{R} is the image grid, and η_p designates the set of sites in image $k-1$ which interact with site p in image k . η_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 \mathcal{A} denote the set of the nine possible polar coordinates a corresponding to the temporal pairs defined in Fig.2. Henceforth, we use the term clique to designate a temporal pair. In practice, we consider three different neighborhoods η^1 , η^5 and η^9 (Fig.2). The simplest case η^1 is just the temporal clique defined by $a = (0, 0)$ whereas η^5 and η^9 refer to the cases with 5 cliques and 9 cliques respectively.

We also assume that $P_{\mathcal{M}}(x_k(p) | x_{k-1}(\eta_p))$ is expressed as the exponential of a sum of local Gibbsian potentials. It can be written as follows:

$$P_{\mathcal{M}}(x_k(p) | x_{k-1}(\eta_p)) = \frac{\exp \left[\sum_{a \in \mathcal{A}} \Psi_{\mathcal{M}}^a(x_k(p), x_{k-1}(p_a)) \right]}{Z_{\mathcal{M}}(k, p, x_{k-1}(\eta_p))} \quad (9)$$

where $\Psi_{\mathcal{M}}^a = \{\Psi_{\mathcal{M}}^a(\nu, \nu')\}_{(\nu, \nu') \in \Lambda^2}$ is the potential for the temporal clique a . Model \mathcal{M} is then defined by $|\mathcal{A}| \cdot |\Lambda|^2$ potential values $\{\Psi_{\mathcal{M}}^a(\nu, \nu')\}_{(\nu, \nu') \in \Lambda^2, a \in \mathcal{A}}$. pixel p_a is the temporal

neighbor of p for clique a in the considered neighborhood η_p (see Fig.2), and $Z_{\mathcal{M}}(k, p, x_{k-1}(\eta_p))$ designates the local normalization constant. This normalization is given by:

$$Z_{\mathcal{M}}(k, p, x_{k-1}(\eta_p)) = \sum_{\nu \in \Lambda} \exp \left[\sum_{a \in \mathcal{A}} \Psi_{\mathcal{M}}^a(\nu, x_{k-1}(p_a)) \right]. \quad (10)$$

The considered statistical models will be referred as “causal spatio-temporal Gibbs” models. Let us point out that they are not usual Gibbs models, which are equivalent to Markov models [23], since the considered neighborhood configuration is not symmetric.

For notation convenience, $\Psi_{\mathcal{M}}^a \equiv \mathbf{0}$ will denote the constant potential for a given clique a . Similarly, the uniform model \mathcal{M} for which $P_{\mathcal{M}}(x) \propto 1$ is specified by the constant potential function for all cliques denoted $\Psi_{\mathcal{M}} \equiv \mathbf{0}$.

Contrary to the case of general Markov random fields [23], such a causal modeling provides an exact expression of the joint distribution $P_{\mathcal{M}}(x)$ as a product of local transition probabilities:

$$P_{\mathcal{M}}(x) = P_{\mathcal{M}}(x_0) \prod_{k=1}^K \prod_{p \in \mathcal{R}} \frac{\exp \left[\sum_{a \in \mathcal{A}} \Psi_{\mathcal{M}}^a(x_k(p), x_{k-1}(p_a)) \right]}{Z_{\mathcal{M}}(k, p, x_{k-1}(\eta_p))}. \quad (11)$$

Thus, for given $P_{\mathcal{M}}(x_0)$ and potentials $\Psi_{\mathcal{M}}$, $P_{\mathcal{M}}$ is entirely known, which provides us with a general statistical framework for motion-based video classification and retrieval as described in Section VI. Following [24], [48], we can now rewrite the causal expression (11) using the temporal cooccurrence measurements attached to the clique a as follows:

$$P_{\mathcal{M}}(x) = P_{\mathcal{M}}(x_0) \frac{\exp \left[\sum_{a \in \mathcal{A}} \Psi_{\mathcal{M}}^a \bullet \Gamma_a(x) \right]}{Z_{\mathcal{M}}(x)}, \quad (12)$$

where $Z_{\mathcal{M}}(x)$ is the global normalization factor given by:

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

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

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

where $\delta()$ denotes the Kronecker delta function. The dot product between cooccurrence matrix Γ_a and potentials $\Psi_{\mathcal{M}}^a$ is defined as follows:

$$\Psi_{\mathcal{M}}^a \bullet \Gamma_a(x) = \sum_{(\nu, \nu') \in \Lambda^2} \Psi_{\mathcal{M}}^a(\nu, \nu') \Gamma_a(\nu, \nu' | x). \quad (15)$$

This statistical framework for motion information modeling in image sequences can be claimed as non-parametric in two ways. Firstly, from a statistical point of view, our approach is non-parametric in the sense that the conditional likelihood $P_{\mathcal{M}}(x_k(p) | x_{k-1}(\eta_p))$ is not assumed to follow a known parametric law (e.g., Gaussian). Secondly, from a measurement point of view, the definition of quantities x_k does not refer to 2D parametric motion model. We think these quantities thus capture motion information in generic enough way to characterize the motion activity.

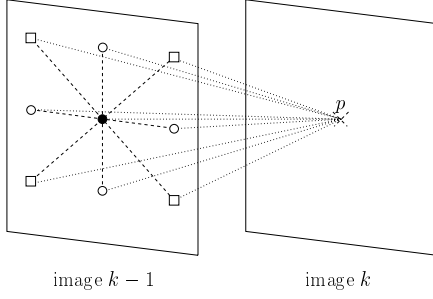
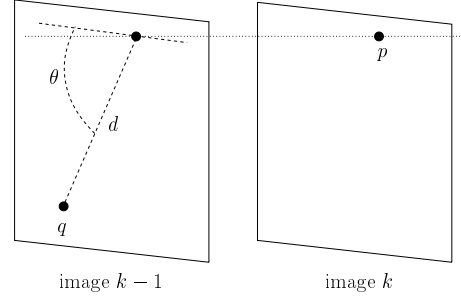
Causal temporal neighborhoods of pixel p Parameterization of the temporal pair (p, q) using polar coordinates $a = (d, \theta)$

Fig. 2. **Causal temporal neighborhood comprising up to 9 pairs.** Given a pixel p in image k , we denote η_p^1 the temporal neighborhood formed by the single site \bullet at same location as p in image $k-1$, η_p^5 the set of the 5 sites represented by symbols \bullet and \circ , and η_p^9 the whole set of the 9 neighbors of p (symbols \bullet , \circ and \square). Each kind of neighbor pair is parameterized using polar coordinates as illustrated on the right.

B. Maximum likelihood estimation of potentials

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

$$\widehat{\mathcal{M}} = \arg \max_{\mathcal{M}} LF_{\mathcal{M}}(x) \quad (16)$$

where the log-likelihood function is given by:

$$LF_{\mathcal{M}}(x) = \ln(P_{\mathcal{M}}(x)). \quad (17)$$

We hereafter assume that $P_{\mathcal{M}}(x_0)$ is uniform. From (12), we get:

$$\widehat{\mathcal{M}} = \arg \max_{\mathcal{M}} \sum_{a \in \mathcal{A}} \Psi_{\mathcal{M}}^a \bullet \Gamma_a(x) - \ln Z_{\mathcal{M}}(x). \quad (18)$$

Setting to zero the first-order derivatives of the log-likelihood function w.r.t. potential values $\{\Psi_{\mathcal{M}}^a(\nu, \nu')\}$ provides the following equations to be simultaneously solved by the ML model estimate:

$$\forall (a, \nu, \nu') \in \mathcal{A} \times \Lambda^2, \quad \sum_{(k, p) \in S_{a\nu\nu'}} P_{\widehat{\mathcal{M}}}(x_k(p) = \nu | x_{k-1}(\eta_p)) = \Gamma_a(\nu, \nu' | x) \quad (19)$$

with $S_{a\nu\nu'} = \{(k, p) \in \{1, \dots, K\} \times \mathcal{R} / x_{k-1}(p_a) = \nu'\}$.

This naturally confirms that significant potentials of model $\widehat{\mathcal{M}}$ correspond to high cooccurrence values, as the model $\widehat{\mathcal{M}}$ will give the highest probabilities to the configurations associated to the greatest cooccurrence values. We exploit this property to reduce the model complexity in subsection V-C. In practice, the maximization in (16) is carried out using a classical conjugate gradient procedure as detailed in Algorithm 1.

It is worth mentioning that the log-likelihood function $LF_{\mathcal{M}}(x)$ may have several local minima w.r.t. $\Psi_{\mathcal{M}}$, whereas the existence of a unique global minimum is guaranteed in the case of exponential models [24]. Hence, it is important to define an appropriate optimization scheme. As described in the next Section, we have adopted an incremental strategy, in terms of model complexity, which has proven robust and accurate enough.

V. MODEL ESTIMATION

This Section details how the potential values which explicitly specify the causal Gibbs models are estimated. Besides, we describe a scheme to reduce model complexity after potential estimation.

Algorithm 1 - Maximum likelihood estimation of model potentials $\Psi_{\mathcal{M}} = (\Psi_{\mathcal{M}}^a(\nu, \nu'))_{a \in \mathcal{A}, (\nu, \nu') \in \Lambda^2}$ by applying a conjugate gradient technique to criterion (16)

• **Step 1:** Initialization

1. $k = 0$
2. Initialize the Gibbs model \mathcal{M}^0
3. Initialize the ascent direction $\mathbf{d}_0 \equiv \mathbf{0}$

• **Step 2:**

1. $k \leftarrow k + 1$
2. Compute the gradient $\nabla LF_{\mathcal{M}^{k-1}}(x)$
3. Update the ascent direction \mathbf{d}_k :

$$\mathbf{d}_k = \nabla LF_{\mathcal{M}^{k-1}}(x) + \frac{\|\nabla LF_{\mathcal{M}^{k-1}}(x)\|^2}{\|\nabla LF_{\mathcal{M}^{k-2}}(x)\|^2} \mathbf{d}_{k-1}$$

4. Search for the coefficient λ_k which verifies:

$$\lambda_k = \arg \min_{\lambda} LF_{\mathcal{M}^{\lambda}}(x)$$

where \mathcal{M}^{λ} stands for the model with potentials $\Psi_{\mathcal{M}^{\lambda}} = \Psi_{\mathcal{M}^k} + \lambda \mathbf{d}_k$. Update model potential:

$$\Psi_{\mathcal{M}^{k+1}} = \Psi_{\mathcal{M}^k} + \lambda_k \mathbf{d}_k$$

- **Step 3:** repeat step 2 until: $\|\nabla LF_{\mathcal{M}^k}(x)\|_{\infty} < \gamma$ where γ is a predefined constant.

A. Estimation of the simple temporal model

When using the simple temporal clique model for which $\mathcal{A} = \{a_0\} = \{(0, 0)\}$, the model under consideration is in fact equivalent to a product of $|\mathcal{R}|$ independent Markov chains. If the unique potential is constrained to verify:

$$\sum_{\nu \in \Lambda} \exp \Psi_{\mathcal{M}}^{(0,0)}(\nu, \nu') = 1, \quad \forall \nu' \in \Lambda, \quad (20)$$

then the transition probabilities amount to:

$$P_{\mathcal{M}}(x_k(p) | x_{k-1}(p)) = \exp \left[\Psi_{\mathcal{M}}^{(0,0)}(x_k(p), x_{k-1}(p)) \right] \quad (21)$$

Thus, this simple temporal model provides a characterization of the temporal aspects of motion content whereas spatial aspects, captured by the more complex spatio-temporal causal Gibbs models, are not explicitly modeled. However, the use of only one clique makes easier the computation and the maximization of the likelihood function

For simple temporal model, the likelihood function is then simply given by:

$$P_{\mathcal{M}}(x) = \exp[\Psi_{\mathcal{M}} \bullet \Gamma(x)] \quad (22)$$

where, for sake of concision, the mention of the unique clique type $a = (0, 0)$ is dropped (e.g., $\Psi_{\mathcal{M}}$ stands for $\Psi_{\mathcal{M}}^{(0,0)}$). The availability of this simple exponential formulation presents several interests. First, it makes the computation of the likelihood $P_{\mathcal{M}}(x)$ for any sequence x and model \mathcal{M} for which $\mathcal{A} = (0, 0)$ feasible and simple. Second, all motion information exploited by these models is contained in the cooccurrence distributions. In particular, in order to evaluate the likelihoods $\{P_{\mathcal{M}_i}(x)\}$ w.r.t. different models $\{\mathcal{M}_i\}$'s for a given sequence x , it is not necessary to store the entire sequence x . We only need to compute and store the related temporal cooccurrence distributions $\Gamma(x)$. The evaluation of the conditional likelihoods $\{P_{\mathcal{M}_i}(x)\}$ is then simply achieved by exponentiating products $\{\Psi_{\mathcal{M}_i} \bullet \Gamma(x)\}$, whereas, in the general case, it is required to store the sequence of maps x to compute the normalization constant $Z_{\mathcal{M}}(x)$.

From equation (19), we get the following ML estimate of the only temporal model for a given sequence x of local motion-related quantities:

$$\Psi_{\widehat{\mathcal{M}}}(\nu, \nu') = \ln \left(\Gamma(\nu, \nu' | x) / \sum_{\nu \in \Lambda} \Gamma(\nu, \nu' | x) \right). \quad (23)$$

Similarly to the computation of the likelihood $P_{\mathcal{M}}(x)$, the ML model estimation only requires the evaluation of the temporal cooccurrence distribution $\Gamma(x)$.

B. Estimation of the extended temporal models

Let us now consider the case of the extended temporal neighborhoods η^5 or η^9 (see Fig.2). To perform the ML estimation, we adopt an incremental strategy, sketched in Algorithm 2. 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 \mathcal{M}^a with potentials set as constant for all cliques b other than a :

$$\widehat{\mathcal{M}}^a = \arg \max_{\mathcal{M}} LF_{\mathcal{M}}(x). \quad (24)$$

$$\forall b \neq a, \Psi_{\mathcal{M}}^b \equiv 0$$

Exactly as for the ML estimation of $\Psi_{\mathcal{M}}^{(0,0)}$ only in previous subsection, the ML estimated potential $\Psi_{\widehat{\mathcal{M}}^a}^a$ is given by: $\forall (\nu, \nu') \in \Lambda^2$,

$$\Psi_{\widehat{\mathcal{M}}^a}^a(\nu, \nu') = \ln \left(\Gamma_a(\nu, \nu' | x) / \sum_{\nu \in \Lambda} \Gamma_a(\nu, \nu' | x) \right), \quad (25)$$

under normalizing constraint:

$$\forall a \in \mathcal{A}, \quad \forall \nu' \in \Lambda, \quad \sum_{\nu \in \Lambda} \exp \Psi_{\widehat{\mathcal{M}}^a}^a(\nu, \nu') = 1. \quad (26)$$

We can rank cliques $a \in \mathcal{A}$ according to the values of the likelihoods of the sequence of motion-related quantities x w.r.t. $\widehat{\mathcal{M}}^a$:

$$\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\} \text{ with } LF_{\widehat{\mathcal{M}}^{a_1}}(x) \geq \dots \geq LF_{\widehat{\mathcal{M}}^{a_{|\mathcal{A}|}}}(x). \quad (27)$$

The incremental estimation of the model \mathcal{M} is then carried out as follows. At step l from 1 to $|\mathcal{A}|$, it consists in estimating the

model $\widehat{\mathcal{M}}^l$ that maximizes the likelihood $LF_{\mathcal{M}}(x)$ under the constraint:

$$\forall b \in \{a_{l+1}, \dots, a_{|\mathcal{A}|}\}, \quad \Psi_{\mathcal{M}}^b \equiv 0. \quad (28)$$

This maximization is achieved using the conjugate gradient ascent described in Algorithm 1 with initialization $\Psi_{\widehat{\mathcal{M}}^{l-1},0} = \Psi_{\widehat{\mathcal{M}}^{l-1}}$. Finally, at iteration $|\mathcal{A}|$, we obtain the ML estimate $\widehat{\mathcal{M}}$ defined on the whole temporal neighborhood structure under consideration.

Algorithm 2 - Incremental strategy for model potential estimation

• **Step 1:** Initialization

1. $l = 1$

2. Sort clique set $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$ according to relation (27)

3. Estimate model potentials $\Psi_{\widehat{\mathcal{M}}^1}$ considering only the first clique a_1 (relation (25))

• **Step 2:**

1. $l \leftarrow l + 1$

2. Introduce the new clique a_l

3. Initialize model potentials $\Psi_{\widehat{\mathcal{M}}^l}$ with $\Psi_{\widehat{\mathcal{M}}^{l-1}}$

4. Use the conjugate gradient procedure (detailed in Algorithm 1) to estimate potentials $\Psi_{\widehat{\mathcal{M}}^l}$ with cliques a_1, \dots, a_l

• **Step 3:** repeat step 2 until $l = |\mathcal{A}|$

C. Model complexity reduction

When considering n cliques (i.e., $|\mathcal{A}| = n$) with N levels of quantization (i.e., $|\Lambda| = N$) for the local motion-related measurements, $n \times N^2$ potential values $\{\Psi_{\mathcal{M}}^a(\nu, \nu')\}_{a \in \mathcal{A}, (\nu, \nu') \in \Lambda^2}$ have to be estimated. Typically, $N = 16$ and $n = 1, 5$, or 9. The number of potential values 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 retaining the most pertinent information in the selected model. Two aspects are considered.

C.1 modification of the range of Λ

Some quantization levels may seldom appear in the sequence of local motion-related quantities x . In that case, the potentials associated with these quantization levels are less important 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 ν^0 with an occurrence number lower than a given threshold, potential values $\{\Psi_{\mathcal{M}}^a(\nu_0, \nu), \Psi_{\mathcal{M}}^a(\nu, \nu_0)\}_{(a, \nu) \in \mathcal{A} \times \Lambda}$ are set to $-\infty$ (a very low value in practice), which corresponds to a null probability of the local configurations including measurement ν_0 . These potentials are let unchanged in the whole estimation process.

C.2 Selection of informative ML potential values

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 potential values of the global estimated model $\widehat{\mathcal{M}}$ while setting the other ones to a constant value. Second, we eliminate cliques that bring negligible information. This model complexity reduction can be regarded as a pruning procedure applied to the set of potential values of the ML estimate of the causal Gibbs model $\widehat{\mathcal{M}}$. To achieve this, we resort to likelihood ratio tests to specify the amount of

information to be kept. For both aspects of complexity reduction, we compute the ratio of the likelihood of sequence x w.r.t. a proposed reduced model \mathcal{M}^* over the likelihood of x w.r.t. $\widehat{\mathcal{M}}$:

$$LR_x(\mathcal{M}^*, \widehat{\mathcal{M}}) = P_{\mathcal{M}^*}(x) / P_{\widehat{\mathcal{M}}}(x). \quad (29)$$

This ratio is compared to a user-specified threshold λ_{LR} . This threshold allows us to specify the tolerated error between the ML estimate of the Gibbs model and the reduced model actually stored. $LR_x(\mathcal{M}^*, \widehat{\mathcal{M}})$ can be viewed as an evaluation of the precision loss occurring if we substitute \mathcal{M}^* for $\widehat{\mathcal{M}}$.

We now describe in more detail the incremental complexity reduction strategy. It is composed of two successive steps: a first step to select the informative model potentials for each clique as described in Algorithm 3, and a second step to select the pertinent cliques as detailed in Algorithm 4.

Algorithm 3 - Selection of the informative potentials of ML estimated model $\widehat{\mathcal{M}}$ for a given clique a

- **Step 1:** Initialization
 1. Sort estimated potential values $\{\Psi_{\widehat{\mathcal{M}}}^a(\nu, \nu')\}_{(\nu, \nu') \in \Lambda^2}$ w.r.t. cooccurrence values $\{\Gamma^a(\nu, \nu' | x)\}_{(\nu, \nu') \in \Lambda^2}$
 2. Initialize the potentials of the reduced model \mathcal{M}^* : $\Psi_{\mathcal{M}^*} \equiv \mathbf{0}$
 - **Step 2:**
 1. Introduce one-by-one sorted potential values $\Psi_{\widehat{\mathcal{M}}}^a(\nu, \nu')$ in the potentials $\Psi_{\mathcal{M}^*}^a$
 2. Compute the likelihood ratio $LR_x(\mathcal{M}^*, \widehat{\mathcal{M}})$ by relation (29) considering only the simple temporal model with clique a
 - **Step 3:** repeat Step 2 while $LR_x(\mathcal{M}^*, \widehat{\mathcal{M}}) < \lambda_{LR}$
-

Concerning model potential selection for a given clique a , equation (19) shows that the largest potential values of ML estimate $\Psi_{\widehat{\mathcal{M}}}^a$ correspond to high cooccurrence values. For a given clique a , potential values $\Psi_{\widehat{\mathcal{M}}}^a(\nu, \nu')$ are one-by-one introduced in a model \mathcal{M}^* (initially, $\Psi_{\mathcal{M}^*} \equiv \mathbf{0}$), according to their corresponding value $\Gamma_a(\nu, \nu' | x)$ in the cooccurrence matrix with the highest values being introduced first. At each step, we compute the likelihood ratio (29). As soon as this ratio exceeds λ_{LR} , we consider the selected potential values as representative of the ML potential estimate $\Psi_{\widehat{\mathcal{M}}}^a$ associated to the sequence x . Let $\widehat{\mathcal{M}}$ denote the reduced model consisting of the potentials selected after this procedure has been applied to each clique a .

Algorithm 4 - Selection of the informative cliques for the ML estimated model $\widehat{\mathcal{M}}$

- **Step 1:** Initialization
 1. $l = 0$
 2. Sort clique set $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$ according to relation (27)
 3. Compute the reduced model potentials $\widehat{\mathcal{M}}$ by selecting the informative potentials for each clique a (see Algorithm 3)
 - **Step 2:**
 1. $l \leftarrow l + 1$
 2. Define the reduced model $\widehat{\mathcal{M}}^l$ using relation (30)
 3. Compute the likelihood ratio $LR_x(\widehat{\mathcal{M}}^l, \widehat{\mathcal{M}})$ using relation (29)
 - **Step 3:** repeat step 2 while $LR_x(\widehat{\mathcal{M}}^l, \widehat{\mathcal{M}}) < \lambda_{LR}$
-

Then, the selection of the representative cliques, as detailed in Algorithm 4, relies on the ranking $a_1, \dots, a_{|\mathcal{A}|}$ defined in

subsection V-B. We consider the different reduced models $(\widehat{\mathcal{M}}^k)_{k \in \{1, \dots, |\mathcal{A}|\}}$ such that:

$$\begin{cases} \Psi_{\widehat{\mathcal{M}}^k}^a = \Psi_{\widehat{\mathcal{M}}}^a, & \forall a \in \{a_1, \dots, a_k\} \\ \Psi_{\widehat{\mathcal{M}}^k}^a \equiv \mathbf{0}, & \forall a \in \{a_{k+1}, \dots, a_{|\mathcal{A}|}\}. \end{cases} \quad (30)$$

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

VI. MOTION-BASED VIDEO CLASSIFICATION AND RETRIEVAL

We now discuss the application of our modeling framework to motion-based video classification and retrieval. Considering a set of video sequences, we are interested in retrieving examples in this database similar, in terms of motion content or more generally of motion activity, to a given video query. 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 Mahalanobis distances, [30], [31]. In our case, we first benefit from our statistical modeling of motion 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 task.

A. Statistical similarity measure related to motion activity

Given video shots characterized by statistical models of motion activity, we have to evaluate the degree of similarity of their contents. We have defined a similarity measure based on the Kullback-Leibler (KL) divergence [4], [6]. Let x_{n_1} and x_{n_2} be the two sequences of motion-related measurements associated to videos n_1 and n_2 , and, \mathcal{M}^{n_1} and \mathcal{M}^{n_2} the two estimated models on them using the method from the previous section.

Considering an approximation of the KL divergence detailed in Appendix A, and using the exponential form (12) of the likelihood function $P_{\mathcal{M}}$, the KL divergence $KL(\mathcal{M}^{n_1} \parallel \mathcal{M}^{n_2})$ of law $P_{\mathcal{M}^{n_2}}$ w.r.t. law $P_{\mathcal{M}^{n_1}}$ can be approximated by:

$$\begin{aligned} KL(\mathcal{M}^{n_2} \parallel \mathcal{M}^{n_1}) &\approx \sum_{a \in \mathcal{A}} [\Psi_{\mathcal{M}^{n_1}}^a - \Psi_{\mathcal{M}^{n_2}}^a] \bullet \frac{1}{K|\mathcal{R}|} \Gamma_a(x^{n_1}) \\ &\quad - \frac{1}{K|\mathcal{R}|} [\ln Z_{\mathcal{M}^{n_1}}(x^{n_1}) - \ln Z_{\mathcal{M}^{n_2}}(x^{n_1})]. \end{aligned} \quad (31)$$

Expression (31) quantifies the loss of information occurring when considering \mathcal{M}^{n_2} instead of \mathcal{M}^{n_1} to model the motion distribution attached to n_1 .

In the case of the simple temporal model defined in Subsection V-A, this reduces to:

$$KL(\mathcal{M}^{n_2} \parallel \mathcal{M}^{n_1}) \approx [\Psi_{\mathcal{M}^{n_1}} - \Psi_{\mathcal{M}^{n_2}}] \bullet \frac{1}{K|\mathcal{R}|} \Gamma(x^{n_1}). \quad (32)$$

In order to deal with a symmetric similarity measure, the similarity measure $D_{KL}(n_1, n_2)$ between elements n_1 and n_2 is defined by:

$$D_{KL}(n_1, n_2) = \frac{1}{2} [KL(\mathcal{M}^{n_1} \parallel \mathcal{M}^{n_2}) + KL(\mathcal{M}^{n_2} \parallel \mathcal{M}^{n_1})]. \quad (33)$$

Note that this similarity measure is not a metric since it does not satisfy the triangular inequality. However, it can be easily computed and interpreted, since it involves logarithms of likelihood ratios.

B. Hierarchical motion-based indexing and retrieval

For efficient retrieval in large databases, it is necessary to structure the target video set beforehand. We focus here on hierarchical representations that have been successfully exploited for browsing or retrieval in still image database citeChen99b,Milanesi96,Schweitzer99,Yeung96. Such indexing structures rely on binary trees. The tree nodes will correspond to subsets of shots of the processed video database. To achieve this hierarchical structuring, either top-down [40] or bottom-up [31] strategies can be adopted. As pointed out in [11], 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 (AHC) procedure, [15].

We also need to define the similarity measure D_{AHC} between clusters of videos used in the ascendant hierarchical classification scheme. For two clusters C^1 and C^2 , D_{AHC} is defined by:

$$D_{AHC}(C^1, C^2) = \max_{(n_1, n_2) \in C^1 \times C^2} D_{KL}(n_1, n_2). \quad (34)$$

We can now construct an ascendant hierarchical classification based on D_{AHC} . It proceeds incrementally as follows. At a given iteration, a pair is formed by merging the closest clusters according to D_{AHC} . If a cluster C is too far from all the others, i.e., $\min_{C' \neq C} D_{AHC}(C, C') > D_{max}$, it is kept alone to form a single cluster. D_{max} is a given threshold. For two clusters C_1 and C_2 , $\exp[-D_{AHC}(C_1, C_2)]$ can be expressed as the product of two likelihood ratios and is comprised in $[0, 1]$ (relations (33) and (34)). Therefore, we set $D_{max} = -\ln \tau$ where τ is a threshold in $[0, 1]$. This Threshold 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 starts at the level of individual shots, which form the leaves of the tree, and is iterated until no new cluster can be built.

For retrieval purposes, a motion activity model has to be attached to each newly created cluster. In the case of the simple temporal Gibbs model, since it is directly determined from temporal cooccurrence measurements, the activity model associated with the cluster formed by merging two clusters can be straightforwardly estimated using relation (23). Indeed, for the set of sequences comprised in the new cluster, the corresponding cooccurrence measurements can be directly determined as the sum of the cooccurrence measurements computed for each sequence of the new 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 then, exploiting relation (23), we estimate the potentials of the Gibbs model associated with the new cluster formed by the union of C_1 and C_2 . On the other hand, such a simple updating is no longer possible for the extended temporal Gibbs models. We could use the incremental estimation scheme described in Subsection V-B. However, it would be computationally demanding when handling large hierarchical structures with numerous nodes. Therefore, we prefer not to estimate the model associated with the union of two clusters to save computation, and rather to select either \mathcal{M}^{C^1} or \mathcal{M}^{C^2} as the model representative of the new cluster resulting from the merged nodes C_1 and C_2 . We select the model that maximizes the likelihood computed

for the motion-related quantity sequence issued from the union of all the sequences from of the two clusters C_1 and C_2 . Even if we thus do not compute the exact model for the new cluster, we believe that the selected model still provides a pertinent characterization of the motion content of the new cluster. Indeed, the two merged clusters are supposed to be similar in terms of motion content.

C. Probabilistic retrieval

As in [43], the retrieval process is formulated as a Bayesian inference issue. Given a video query q , we aim at determining the best match n^* in the stored set \mathcal{D} of video sequences according to the MAP criterion:

$$n^* = \arg \max_{n \in \mathcal{D}} P(n|q) = \arg \max_{n \in \mathcal{D}} P(q|n)P(n). \quad (35)$$

The distribution $P(n)$ allows us to formulate *a priori* knowledge of the video content relevance over the database. It can be inferred from semantic descriptions attached to each type of video sequence. This distribution could also be learned from relevance feedback during the retrieval process [33]. 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 however, we will in fact incorporate no *a priori* (distribution $P(n)$ is taken uniform, i.e. $P(n) \propto 1$).

Furthermore, criterion (35) also supplies a ranking of the elements $\{n\}_{n \in \mathcal{D}}$ according to $P(q|n)P(n)$, which quantifies how relevant the selection of n w.r.t. the motion content of query q is. In our case, to each element n of the database, a causal Gibbsian model \mathcal{M}^n is attached. We compute the sequence of motion-related measurements x^q for video query q and the likelihood $P(q|n)$ is expressed using $P_{\mathcal{M}^n}$. Then, we obtain:

$$n^* = \arg \max_n P_{\mathcal{M}^n}(x^q). \quad (36)$$

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

In addition, we can take advantage of the hierarchical representation of the video database mentioned in the previous section to satisfy a video query. When dealing with large databases, solving criterion (36) exhaustively is quite time consuming. Therefore, we exploit the constructed binary tree to obtain a suboptimal but efficient solution of criterion (36). If obtaining 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. To initialize, we select the best node C^0 at the root \mathcal{T}_{root} of the search tree according to:

$$C^0 = \arg \max_{C \in \mathcal{T}_{root}} P_{\mathcal{M}^C}(x^q). \quad (37)$$

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

$$C^{k+1} = \arg \max_{C \in C^k} P_{\mathcal{M}^C}(x^q). \quad (38)$$

This procedure is iterated until a given maximal number of elements in the selected cluster is reached.

VII. RESULTS

We have evaluated the whole proposed framework for motion activity modeling, content-based video indexing and content-based video retrieval, on a database containing samples of real videos. We have paid particular care to choose examples that



Fig. 3. Set of the 20 video shots used in the motion-based hierarchical classification of Fig.4. For each video, we display the median image of the shot.

are representative of various motion situations. The database includes temporal textures (samples of fire and sequences of river), video shots exhibiting significant motion activity such as sports videos (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 derived from 70 video shots (elements from the same video shot are not temporally adjacent). A sample of frames from this video set is provided in Fig. 3.

The experiments reported in this section have been performed using parameter values set as follows. In the motion-related measurement stage, we set $V_{max} = 4.0$ and $|\Lambda| = 16$. These values seem to be suitable based on previous work [7], [17], [20]. For the model complexity reduction stage, we set $\lambda_{LR} = 0.99$. Finally, we set $D_{max} = 2.3$ (i.e., $\tau = 0.1$), in the hierarchical structuration of the database.

The goal of this section is to illustrate the interest of the designed statistical motion modeling with different neighborhood structures for video indexing and retrieval. We do not aim at supplying a complete experimental comparison of the different versions of the motion activity modeling introduced in this paper. We report an example of motion-based hierarchical classification using neighborhood η^5 , and we describe different retrieval operations with query by example using neighborhood η^1 .

A. Model complexity reduction

In a first step, we have estimated the causal Gibbs model attached to each element of the database for the neighborhood η^5 (see Section IV). For the processed database, we finally only kept from 5% to 20% of the 1280 potential values (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 reduc-

tion phase. We report here two examples of model complexity reduction respectively for shots *anchor1* and *basket1* with the temporal neighborhood η^5 . The median images of these two sequences are displayed in Fig.3. Video *anchor1* is a static shot of an anchor person in a news program. The motion activity is very low and only potential values related to low values of motion magnitude are kept. This leads to 5% of the estimated ML potential values being retained and only one clique out of the five initial ones. The second example *basketball3* involves substantial motion activity. The stored Gibbs model is more complex, with two selected cliques and 10% of the estimated potential values being retained.

B. Statistical hierarchical motion-based classification

To provide a comprehensive visualization of the statistical hierarchical motion-based classification described in Section VI, we have performed a classification on the subset of 20 sequences displayed in Fig.3. It contains: two shots of anchor person in news programs, *anchor1* and *anchor2*, with a very weak motion activity; two other examples of low motion activity, *hall* and *Concorde*; four examples of rigid motion situations corresponding to road traffic sequences, *highway1* and *highway2*, and airport sequences, *landing* and *take-off*; ten sport video sequences involving shots of rugby games, *rugby1* and *rugby2*, hockey games, *hockey1*, *hockey2*, and *hockey3*, basketball games, *basketball1*, *basketball2* and *basketball3*, and windsurfing, *windsurfing1* and *windsurfing2*; finally, two samples of temporal textures with high motion activity, *fire* and *river*.

For this experiment, we exploit extended temporal models corresponding to η^5 . The unsupervised hierarchical classification obtained, shown in Fig.4, correctly separates the different kinds of dynamic contents. Traffic sequences, *road1* and *road2*,

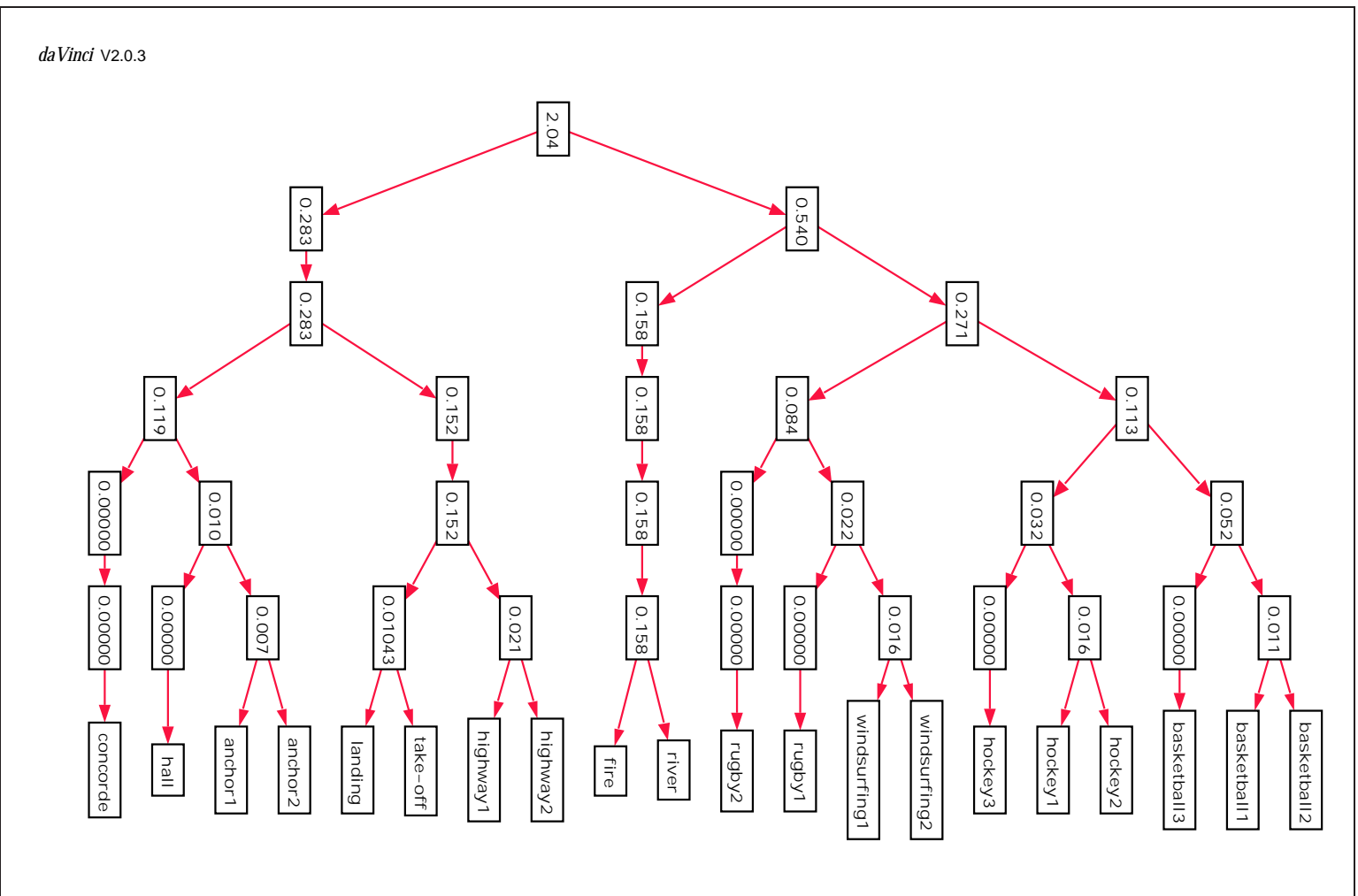


Fig. 4. **Motion-based statistical classification:** obtained motion-based hierarchical classification for the set of 20 video sequences presented in Fig.3 with $D_{max} = 2.3$ ($\tau = 0.1$). At each leaf of the tree, we report the name of the video sequence. For the other nodes of the tree, we display the maximum intra-cluster distance evaluated using expression D_{ANC} of relation (34).

airport videos, *landing* and *take-off*, and low motion activity situations, *anchor₁*, *anchor₂*, *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.

C. 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 η^1 . Fig.5 describes the results of four experiments of retrieval operations with the query-by-video example. The first query is a news program which consists of a static shot of an anchor person. A rigid motion situation (airplane take-off) is proposed as the second query. The third and fourth retrieval operations concern sport videos. The third query is a global view of the game field, whereas a close-in shot of a basketball player tracked by the camera constitutes the last example. We locate the three best replies according to the computed log-likelihood values $P_{\mathcal{M}^c}(x^q)$ (as given in relation (38)). For all the considered queries, the retrieval process supplies 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 evaluate *a posteriori* the relevance of the replies, we have also estimated the model \mathcal{M}^q associated with the query q and we report the values of the distance $D_{KL}(q, n)$ given by relation (33) between \mathcal{M}^q and the different retrieved models \mathcal{M}^n . The ranking supplied by log-likelihood values is confirmed by the values of distance D_{KL} 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 database 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.4. More precisely, class (I) refers to low motion 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 the evaluation of retrieval performances w.r.t. semantic classes is necessarily 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 database and not by way of to an exhaustive search. For the processed video database, the first retrieved answer is the query shot 76% of the time. 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* motion activity classes described above. The results obtained with simple temporal Gibbs models are given in Table I. For classes (I), (II), (III) and (IV), the rate of correct classification is mostly within the range 89% to 100%. 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 11% for class (II) which involves rigid motion situations. The corresponding video shots do actually involve rigid objects, but these are 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 illustrated in Fig.4. However, this evaluation should be considered as a first validation of our approach. We plan to evaluate it on a larger database.

D. Discussion

Promising results have been obtained using the statistical non-parametric motion models introduced in this paper both for motion classification and for motion-based video retrieval on a video database involving various types of motion activity. We have not tackled the issue of selecting the causal spatio-temporal neighborhood structure. As far, note that there are unfortunately no labeled video databases and protocols available in order to carry out objective performance comparisons between different methods in the field of content-based video indexing and retrieval.

However, in order to evaluate the influence of the choice of the neighborhood structure on the achieved global motion characterization, we addressed a different motion recognition task in another work [16]. In that case, a ground-truth was available to compute rates of correct and false classification, and we were able to compare statistical motion activity models associated with different neighborhood structures. We refer the reader to [16] for further details on these experiments. For the considered motion recognition task, spatio-temporal neighborhoods η^5 and η^9 did not bring substantial improvements compared to the simple temporal model, while the latter is far less complex and time consuming regarding the computation of likelihood functions and the ML model estimation.

VIII. CONCLUSION

We have described an original method for the global characterization of motion content in video sequences, which is able to handle a very large range of dynamic scene contents. We rely on a statistical modeling of the distribution of local motion-related measurements using non-parametric causal Gibbs distribution fitted at the ML sense. In addition, 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.

In future work, we plan to validate our approach on a larger video database. In that context, as pointed out in [11], 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 using this novel approach of motion modeling and characterization to automatically segment entities of interest in the shot and to satisfy partial queries [18]. It could also be useful to extract shots of interest in video sequences with a view to creating video summaries.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of 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*.



Fig. 5. **Results of retrieval operations involving three replies.** For each reply n , we give the value LF of the log-likelihood $\ln(P_{\mathcal{M}^n}(x^q))$ corresponding to video query q . To evaluate a posteriori the relevance of the replies, we have also estimated the model \mathcal{M}^q associated to the query q and we report the values of the distance $D_{KL}(n, q)$, given by relation (33) between \mathcal{M}^q and the different retrieved models \mathcal{M}^n .

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

TABLE I

TAB. I. 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 low motion 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 database as a query. For instance, within the 18 elements of class II, 89% and 11% were respectively assigned to classes (II) and (IV).

APPENDIX

I. APPROXIMATION OF KULLBACK-LEIBLER DIVERGENCE

In this appendix, we give details of a Monte-Carlo approximation of the KL divergence (39). Considering two probability distributions μ and μ' , the Kullback-Leibler (KL) divergence $KL(\mu\|\mu')$ is defined by:

$$KL(\mu\|\mu') = \int \ln \frac{\mu}{\mu'} d\mu. \quad (39)$$

It can be viewed as the expectation of the log-likelihood ratio $\ln(\mu/\mu')$ w.r.t. distribution μ . In our case, if we consider an element n of the processed video database, the sequence of motion-related quantities x^n represents a sample associated with the distribution modeled by \mathcal{M}^n . More precisely, for each $(k, p) \in [1, K] \times \mathcal{R}$, the transition probability from $x_{k-1}^n(\eta_p)$ to $x_k^n(p)$ is governed by the causal Gibbs model \mathcal{M}^n . If we consider two elements n_1 and n_2 of the video database, their associated models \mathcal{M}^{n_1} and \mathcal{M}^{n_2} , and the sequences of computed motion-related quantities x^{n_1} and x^{n_2} , then, the KL divergence $KL(\mathcal{M}^{n_1}\|\mathcal{M}^{n_2})$ is approximated as the average of the log-ratio of the transitions probabilities from $x_{k-1}^{n_1}(\eta_p)$ to $x_k^{n_1}(p)$ computed respectively w.r.t. \mathcal{M}^{n_1} and \mathcal{M}^{n_2} :

$$KL(\mathcal{M}^{n_1}\|\mathcal{M}^{n_2}) \approx \frac{1}{K|\mathcal{R}|} \sum_{k=1}^K \sum_{p \in \mathcal{R}} \ln \left(\frac{P_{\mathcal{M}^{n_1}}(x_k^{n_1}(p)|x_{k-1}^{n_1}(\eta_p))}{P_{\mathcal{M}^{n_2}}(x_k^{n_1}(p)|x_{k-1}^{n_1}(\eta_p))} \right). \quad (40)$$

Due to the causal nature of the model, it comes to approximate the KL divergence $KL(\mathcal{M}^{n_1}\|\mathcal{M}^{n_2})$ by the log-ratios of the likelihoods of the sequence of motion-related quantities x^{n_1} under models \mathcal{M}^{n_1} and \mathcal{M}^{n_2} respectively:

$$KL(\mathcal{M}^{n_1}\|\mathcal{M}^{n_2}) \approx \frac{1}{K|\mathcal{R}|} \ln \left(\frac{P_{\mathcal{M}^{n_1}}(x^{n_1})}{P_{\mathcal{M}^{n_2}}(x^{n_1})} \right) \quad (41)$$

Using the exponential formulation of law $P_{\mathcal{M}}$, we then obtain relation (31).

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, 1996.
- [2] E. Ardizzone and M. La Cascia. Automatic video database indexing and retrieval. *Multimedia Tools and Applications*, 4(1):29–56, 1997.
- [3] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *Int. J. of Comp. Vis.*, 12(1):43–77, 1994.
- [4] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989.
- [5] A. Del Bimbo. *Visual information retrieval*. Morgan Kaufmann, 1999.
- [6] 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.
- [7] P. Boutheimy and R. Fablet. Motion characterization from temporal co-occurrences of local motion-based measures for video indexing. In *Proc. of 14th Int. Conf. on Pattern Recognition, ICPR'98*, pages 905–908, Brisbane, Aug. 1998.
- [8] P. Boutheimy, C. Garcia, R. Ronfard, G. Tziritas, E. Veneau, and D. Zujaj. Scene segmentation and image feature extraction for video indexing and retrieval. In LNCS Vol 1614, editor, *Proc. of 3rd Int. Conf. on Visual Information Systems, VISUAL'99*, pages 245–252. Springer, 1999.
- [9] P. Boutheimy, 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.
- [10] R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *J. of Vis. Comm. and Im. Repr.*, 10(2):78–112, 1999.
- [11] 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.
- [12] J.-Y. Chen, C. Taskiran, E. J. Delp, and C. A. Bouman. ViBE: A new paradigm for video database browsing and search. In *Workshop on Content-Based Access of Image and Video Libraries, CVPR'98*, pages 96–100, Santa-Barbara, June 1998.
- [13] J.D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–625, 1997.
- [14] 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.
- [15] 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.
- [16] R. Fablet. Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images : application à l'indexation vidéo. *Phd thesis, University of Rennes 1, Irisa No. 2526*, 2001.
- [17] R. Fablet and P. Boutheimy. 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.
- [18] R. Fablet and P. Boutheimy. Non-parametric motion activity analysis for statistical retrieval with partial query. *Journal of Mathematical Imaging and Vision*, 14(3):257–270, 2001.
- [19] R. Fablet, P. Boutheimy, 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, Oct. 1999.
- [20] R. Fablet, P. Boutheimy, 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, Apr. 2000.
- [21] 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, Oct. 1998.
- [22] M. Gelgon and P. Boutheimy. 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.
- [23] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on PAMI*, 6(6):721–741, 1984.
- [24] G.L. Gimel'Farb. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. on PAMI*, 18(11):1110–1114, 1996.
- [25] F. Heitz and P. Boutheimy. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. on PAMI*, 15(2):1217–1232, 1993.
- [26] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.
- [27] M. Irani and P. Anandan. Video indexing based on mosaic representation. *Proc. of the IEEE*, 86(5):905–921, 1998.
- [28] 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.
- [29] A. Jain and D. Zongker. Feature selection: evaluation, application and small sample performance. *IEEE Trans. on PAMI*, 19(2):153–158, 1997.
- [30] A.K. Jain, A. Vailaya, and W. Xiong. Query by video clip. *Multimedia Systems*, 7(5):369–384, 1999.
- [31] 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, Sept. 1996.
- [32] A. Mitche and P. Boutheimy. Computation and analysis of image motion: a synopsis of current problems and methods. *Int. J. of Comp. Vis.*, 19(1):29–55, 1996.
- [33] 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.
- [34] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP*, 56(1):78–99, 1992.
- [35] J.M. Odobez and P. Boutheimy. Robust multiresolution estimation of parametric motion models. *J. of Vis. Comm. and Im. Repr.*, 6(4):348–365, 1995.
- [36] J.M. Odobez and P. Boutheimy. 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.
- [37] 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, Aug. 1998.
- [38] Y. Rui, T. Huang, and S. Mehrotra. Constructing table-of-content for videos. *Multimedia Systems*, 5(7):359–368, 1999.
- [39] S. Santini and R. Jain. Similarity measures. *IEEE Trans. on PAMI*, 21(9):871–883, 1999.
- [40] H. Schweitzer. Organizing image databases as visual-content search trees. *Image and Vision Computing*, 17:501–511, 1999.
- [41] C. Stiller and J. Konrad. Estimating motion in image sequences. *IEEE Signal Processing Magazine*, 16(4):70–91, 1999.

- [42] 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, Sept. 1996.
- [43] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'2000*, pages 216–221, Hilton Head, June 2000.
- [44] 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.
- [45] M. M. Yeung, B.-L. Yeo, and B. Liu. Extracting story units from long programs for video browsing and navigation. In *Proc. 3rd IEEE Int. Conf. on Multimedia Computing and Systems, ICMCS'96*, pages 296–305, Hiroshima, Japan, June 1996.
- [46] L. Younes. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Poincaré*, 24(2):269–294, 1988.
- [47] 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, 1997.
- [48] S.C. Zhu, T. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling. *Int. J. of Comp. Vis.*, 27(2):107–126, 1998.