HAL
open science

Motion recognition using non parametric image motion models estimated from temporal and multiscale cooccurrence statistics

Ronan Fablet, Patrick Bouthemy

To cite this version:

Ronan Fablet, Patrick Bouthemy. Motion recognition using non parametric image motion models estimated from temporal and multiscale cooccurrence statistics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25 (12), pp.1619 - 1624. 10.1109/TPAMI.2003.1251155 . hal-02341756

HAL Id: hal-02341756

https://hal.science/hal-02341756v1

Submitted on 25 Nov 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.

Motion Recognition Using Nonparametric Image Motion Models Estimated from Temporal and Multiscale Co-Occurrence Statistics

R. Fablet and P. Bouthemy

Abstract—A new approach for motion characterization in image sequences is presented. It relies on the probabilistic modeling of temporal and scale co-occurrence distributions of local motion-related measurements directly computed over image sequences. Temporal multiscale Gibbs models allow us to handle both spatial and temporal aspects of image motion content within a unified statistical framework. Since this modeling mainly involves the scalar product between co-occurrence values and Gibbs potentials, we can formulate and address several fundamental issues: model estimation according to the ML criterion (hence, model training and learning) and motion classification. We have conducted motion recognition experiments over a large set of real image sequences comprising various motion types such as temporal texture samples, human motion examples, and rigid motion situations.

Index Terms—Nonparametric motion analysis, motion recognition, multiscale analysis, Gibbs models, co-occurrences, ML criterion.

1 Introduction

MOTION information is a crucial cue for visual perception. The well-known MLD (Moving Light Display) experiments carried out in the early seventies [11] demonstrated that human beings were able to recognize activities such as walking or getting up simply by perceiving moving dot lights appropriately placed on the body. Initially, research in motion interpretation by computer vision was dedicated to the recovery of 3D motion information from image sequences and usually relied on the computation of dense optic flow fields known to be an ill-posed problem [9]. We believe that the complete recovery of motion information is not always required to achieve a useful interpretation of motion content. The key point for applications such as motion classification [12] or action recognition [19] is rather to determine appropriate (possibly partial) representation of motion information which can be easily computed from images while enabling further interpretation. We adopt this point of view to address motion recognition with no a priori knowledge on the content of the observed dynamic scenes. Our goal is to design a general framework to provide a global characterization of motion content within image sequences. It will, in particular, involve the design of an appropriate motion-based similarity measure between image sequences.

If the classification of general motion content is sought, the use of nonparametric motion models as opposed to 2D parametric motion models, e.g., affine or quadratic motion models, appears necessary. Indeed, the latter cannot account for a large range of motion types. In that context, Nelson and Polana [12] introduced the notion of temporal textures which refers to nonstructured dynamic scenes such as fluttering leaves, or river scenes. They followed an approach originally developed for spatial texture analysis to characterize distributions of local motion-related measurements in images from co-occurrence statistics. The resulting description can be interpreted in terms of motion activity. New developments in that direction have more recently been proposed for motion-based video indexing and retrieval [5], [6], [14], [19].

We further investigate such an approach and we introduce new probabilistic motion models with a view to handling both spatial and temporal properties of image motion content within a unified statistical framework. The proposed temporal multiscale Gibbs models are exploited for motion recognition while considering a wide range of motion types, from rigid motion situations to temporal texture samples. The remainder of this paper is organized as follows: Section 2 outlines the general ideas supporting our work. Section 3 describes the considered local motion-related measurements we use for nonparametric motion modeling. In Section 4, the statistical modeling of motion information and the estimation of the introduced models according to the Maximum Likelihood (ML) criterion are addressed. Section 5 is specifically concerned with the motion recognition issue. Experiments are reported in Section 6 and Section 7 contains concluding remarks.

2 Related Work

Global motion characterization can be based on different types of motion measurements from image sequences. Motion-based features can be extracted from dense velocity fields computed using optic flow methods as in [1] or issued from MPEG motion vectors as in [3], if MPEG-coded videos are processed. Dense optic flow estimation remains a difficult task especially for complex dynamic scenes such as temporal textures. Besides, the accuracy of MPEG motion vectors is often poor and extremely dependent on the MPEG encoder. Furthermore, the physical reliability of the MPEG motion vectors is questionable since their computation rely on image coding criteria. Motion-based features extracted from dense velocity fields usually correspond to first-order statistics (histogram [3], mean values in different directions [1]). In this case, considering higher order statistics, as in texture analysis, appears too complex since velocity measurements are two-dimensional vectors. If co-occurrence statistics were computed on velocity fields with N levels of quantization over the vertical and horizontal directions, it would result in a co-occurrence matrix of size N^4. Therefore, scalar motion-related measurements are more suitable, even if they only convey partial motion information.

Let us point out that motion characterization should account for temporal and spatial properties of image motion content, as illustrated in Fig. 1. By spatial properties, we mean the spatial pattern formed by the local motion measurements in a given image. Specific patterns can be easily figured out in case of global translation, rotation, or zooming, for instance, but spatial patterns can be associated with more complex situations, too. Temporal properties can be analyzed on Eulerian or Lagrangian basis (i.e., at a fixed image grid location or along the trajectory of a given point). The former is easier to consider and supplies information relative to the temporal variability of the motion content.

Spatial aspects of motion content are the main focus in [12], as motion-based features extracted from spatial co-occurrences of normal flow fields are exploited to classify sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In [14], different spatial motion-based descriptors, also computed from normal flow fields, are considered using other techniques developed for texture analysis. To take into account the temporal properties of image motion distribution, we previously proposed to extract global motion-based features from temporal co-occurrences of local motion-related measurements in [5]. In [19], the focus is also given to temporal features by means of histograms of local quantities (spatial and temporal intensity derivatives) computed at different temporal scales. It nevertheless appears more relevant to combine spatial and temporal motion information to successfully achieve motion characterization. Such an attempt has been investigated by using spatio-temporal Gabor filters applied to image intensities [18].

- *R. Fablet is with IFREMER/LASAA, BP 70, 29280 Plouzané, France. E-mail: rfablet@ifremer.fr.*
- *P. Bouthemy is with IRISA/INRIA, Campus de Beaulieu, 35042 Rennes, France. E-mail: bouthemy@irisa.fr.*

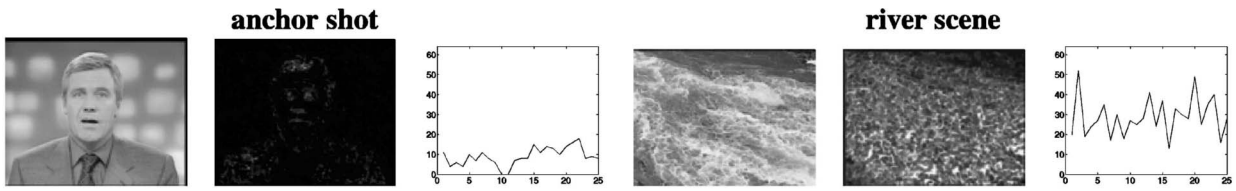

Fig. 1. Illustration of spatial and temporal properties of apparent motion within image sequences. For two sequences involving very different motion contents, we display the first frame of the image sequence, a map of quantized local motion-related measurements as defined in Section 3 and the temporal evolution, over 25 successive frames, of the local motion-related quantity computed at the image center.

However, the use of only numerical (global) motion-based features is limited. Statistical modeling has already proven its ability to supply flexible, general, and efficient frameworks for classification and recognition issues. In particular, the introduction of probabilistic models, such as Gibbs random fields, has led to important advances for texture analysis [2], [15], [20]. Probabilistic models have also been considered for temporal texture synthesis and recognition [7], [16], [17], but the developed solutions are suited for appearance change modeling and not really for general motion information analysis. We have also further investigated the analogy between texture analysis and nonparametric motion analysis and statistical nonparametric motion models have been introduced [6]. We have made use of Gibbs models since there is an explicit relationship between co-occurrence measurements and Gibbs models [20]. The probabilistic models designed in [6] adopt a causal formulation of the problem, since the computation of the normalizing constant of the conditional likelihood is then easy and even explicit. Such a property is of key importance for recognition or classification issues, as they require the comparison of the conditional likelihoods of the observations (here, local motion-related measurements) for different motion models. Here, we extend our previous work [6] in three ways. First, while keeping the key characteristics of the temporal Gibbs models involving the exact derivation (including the normalizing factor) and low-cost computation of the associated likelihood function and the direct ML estimation of the models, we propose temporal and multiscale Gibbs models to represent both temporal and spatial properties of image motion content. Second, we have improved the computation and the quantization steps of the local motion-related measurements, which is of key importance to the evaluation of reliable co-occurrence statistics. Third, we deal with motion recognition, whereas in [6], we were concerned with video indexing and (query-by-example) retrieval issues

3 Motion Data Computation

3.1 Local Motion-Related Measurements

The Optic Flow Constraint Equation (OFCE) relates the intensity derivatives to the displacement $\mathbf{w}(p)$ at point p by assuming brightness constancy along trajectories [9]: $\mathbf{w}(p) \cdot \nabla I(p) + I_t(p) = 0$, where ∇I is the spatial gradient of the intensity function I and I_t its temporal derivative. We can then infer the expression of the normal flow, $v_n(p) = -I_t(p)/\|\nabla I(p)\|$ exploited in [12], [14]. However, the latter is known to be very sensitive to the noise attached to the computation of the intensity gradient ∇I. A weighted average of the magnitude of normal flows within a local window forms a more reliable measurement. The weights are given by the spatial intensity gradient norms, as proposed, for instance, in [13]:

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

where $\mathcal{F}(p)$ is a 3×3 window centered on image point p, and η^2 a predetermined constant related to the noise level (typically, $\eta = 5$). This measure $v_{obs}(p)$ conveys no information on motion direction. However, we are not interested in determining specific motion values or actions, but we aim at supplying a global characterization of the dynamic content of image sequences into relevant "qualitative" motion classes or events. Contrary to [12], [14], we do not exploit the direction information attached to the normal flow, which are defined by the spatial image gradient. In fact, they only reveal the spatial texture present in the image, whereas we are concerned with a general description of motion content independently of the texture (or color) of the surfaces of the objects moving in the depicted scene.

An interesting feature of the motion-related measure given by (1) is that we can exhibit bounds to evaluate the reliability of the information it conveys. Given a motion magnitude Δ, one can derive two bounds, $l_\Delta(p)$ and $L_\Delta(p)$, verifying the following properties. If the motion-related measurement $v_{obs}(p)$ is smaller than $l_\Delta(p)$, the magnitude of the real (unknown) displacement $\mathbf{w}(p)$ at point p is necessarily lower than Δ. Conversely, if $v_{obs}(p)$ is higher than $L_\Delta(p)$, $\|\mathbf{w}(p)\|$ is necessarily greater than Δ. The two bounds $l_\Delta(p)$ and $L_\Delta(p)$ are directly computed from the spatial derivatives of the intensity function within a small window centered on point p [13]. Another issue is to cope with the shortcomings of the OFCE. The OFCE is known to be exploitable only for small displacements and to become invalid in occlusion areas, over motion discontinuities and even on sharp intensity discontinuities. Hence, we have now settled a multiscale scheme and used the likelihood test designed in [8] to evaluate the validity of the OFCE. We build Gaussian pyramids for the processed pair of successive images. Then, at each point p, we select the finest scale for which the OFCE is valid and we compute at that scale the motion-related measurement $v_{obs}(p)$ and associated bounds $l_\Delta(p)$ and $L_\Delta(p)$. If the OFCE remains invalid at all scales, we do not compute any motion measurement at point p.

3.2 Markovian Quantization of the Local Motion-Related Measurements

To evaluate co-occurrences of the motion-related measurements $v_{obs}(p)$, a quantization step is first needed. It is also preferable to cope with discrete states for model estimation and storage issues. Furthermore, the definition of a quantization range common to all the processed image sequences is a requirement to properly evaluate similarities between image sequences. A straightforward and low-cost solution is to linearly quantize the motion-related measurements $v_{obs}(p)$ as initially done in [5], [6]. However, as shown in [4], it may not be adapted to all motion contents and the introduction of contextual information can be beneficial. We have exploited the confidence bounds associated to the local motion-related measurements to define an efficient quantization scheme stated as a Markovian labeling issue. A major result is that the quantized values can now be considered as approximations of the magnitudes of the real (unknown) displacements. Given Λ, the set of quantized values of the motion-related measurements, $\Lambda = \{\nu_0 = 0, \nu_1, \nu_2, \ldots, \nu_{|\Lambda|}\}$ with $0 < \nu_1 < \ldots < \nu_{|\Lambda|}$, we determine the interval $[\nu_{i-1}, \nu_i]$ within which the magnitude of the real (unknown) displacement at point p is most likely to fall. In addition, our Markovian labeling scheme allows us to cope with spurious local measurements in a well-formalized way. As demonstrated by

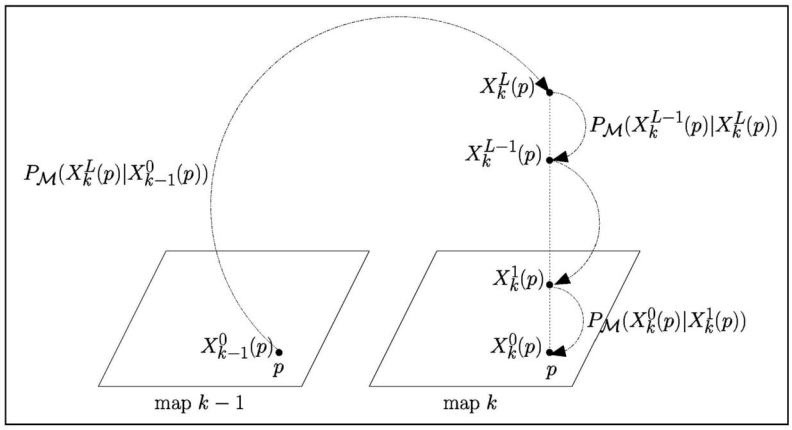

Fig. 2. Illustration of the conditional dependencies involved in the temporal multiscale Gibbs models defined from temporal and scale co-occurrence statistics. Given a point p at instant k, $X_k(p) = \{X_k^0(p), \ldots, X_k^L(p)\}$ is the vector of motion variables corresponding to scales 0 to L. $P_{\mathcal{M}}(X_k^L(p)|X_{k-1}^0(p))$ and $\{P_{\mathcal{M}}(X_k^l(p)|X_k^{l+1}(p))\}_{l\in[[0,\ldots,L-1]]}$ are conditional likelihoods of the transitions involved by the temporal multiscale Gibbs model $\mathcal{M}$.

experiments reported in [4] considering pure known motions (translation, rotation, zooming), our Markovian quantization yields quantized values closer to the magnitude of the real displacements than a simple linear quantization. Objective evaluation was performed between the maps of quantized values derived by the two quantization methods and the map of magnitudes of the real displacements (ground-truth), using the global mean square error and L_1 distance between motion histograms.

Let us formally describe our Markovian quantization. Let $\mathcal{R}$ be the spatial image grid, $e = (e_p)_{p\in\mathcal{R}}$ the quantization label field where each label takes its value in Λ, and $o = (v_{obs}(p))_{p\in\mathcal{R}}$ the map of local motion-related measurements. The solution is given by the minimization of a global energy function U as follows:

$$\widehat{e} = \arg\min_{e\in\Lambda^{|\mathcal{R}|}} U(e,o) = \arg\min_{e\in\Lambda^{|\mathcal{R}|}} [U_1(e,o) + U_2(e)], \quad (2)$$

where the energy function U splits into a data-driven term $U_1(e,o)$ and a regularization term $U_2(e)$. In addition, U_1 and U_2 are expressed as the sum of local potentials: $U_1(e,o) = \sum_{p\in\mathcal{R}} V_1(e_p, v_{obs}(p))$ and $U_2(e) = \sum_{(p,q)\in\mathcal{C}} \beta\cdot\rho(e_p - e_q)$. $\mathcal{C}$ denotes the set of binary cliques of the 4-connectivity neighborhood and β is a positive factor setting the influence of the regularization term (in practice, β is set to 2.0). ρ designates a hard-redescending M-estimator, here Tukey's biweight function [10], and allows us to preserve the (unknown) discontinuities present in the motion field. Given a quantization level ν_i with $i \in \{1,\ldots,|\Lambda|\}$, potential $V_1(\nu_i, v_{obs}(p))$ evaluates how likely the magnitude of the real (unknown) displacement at point p is to be within the interval $[\nu_{i-1}, \nu_i]$. It is given by: $V_1(\nu_i, v_{obs}(p)) = Sup_{L_{\nu_{i-1}}(p)}(v_{obs}(p)) + Inf_{l_{\nu_i}(p)}(v_{obs}(p))$. Sup_L is a continuous step function centered on L, and Inf_l is the opposite of a step function centered in l, both shifted to be within $[0,1]$. Minimization (2) is solved using an ICM-style algorithm, and the initialization is given by considering only the data-driven term in the minimization. Two examples of maps of quantized motion-related measurements are shown in Fig. 1. We consider 64 quantization levels within $[0, 8[$.

4 Statistical Nonparametric Motion Modeling

4.1 Temporal Multiscale Gibbs Models

In order to account for both the spatial and temporal properties of the image motion content, we have designed a multiscale statistical framework. Given a sequence x of maps of quantized motion-related measurements, we form at each image point a vector of motion measurements computed at different scales, instead of considering only one single value. Let us consider a sequence of $K+1$ maps of quantized motion-related measurements $v = (v_0, v_1, \ldots, v_K)$ computed from a sequence of $K+2$ frames. From the sequence v, we build a new sequence $x = (x_0, x_1, \ldots, x_K)$. For $k \in [[0,K]]$, $x_k(p)$ is defined as the vector of quantized measures $(x_k^0(p) = v_k(p), \ldots, x_k^L(p))$ computed at successive scales 0 to L, by applying L Gaussian filters of increasing variance to the map v_k at point p.

Similarly to recent work on texture analysis [2], [15], we exploit scale co-occurrences to characterize spatial image motion properties. For $l \in [[0, L-1]]$, the scale co-occurrence distribution $\Gamma^l(x)$ is given by:

$$\forall(\nu,\nu') \in \Lambda^2, \quad \Gamma^l(\nu,\nu'|x) = \sum_{k=1}^{K}\sum_{p\in\mathcal{R}} \delta(\nu - x_k^l(p))\delta(\nu' - x_k^{l+1}(p)), \quad (3)$$

with δ the Kronecker symbol. To account for temporal motion content, we consider temporal co-occurrences as in our previous work [5], [6], but the temporal co-occurrence distribution $\Gamma^L(x)$ is now defined as:

$$\forall(\nu,\nu') \in \Lambda^2, \quad \Gamma^L(\nu,\nu'|x) = \sum_{k=1}^{K}\sum_{p\in\mathcal{R}} \delta(\nu - x_k^L(p))\delta(\nu' - x_{k-1}^0(p)). \quad (4)$$

To maintain a causal global formulation, the temporal co-occurrences intervene between the motion quantities, respectively, at scale 0 at time $k-1$ and at scale L at time k. The underlying model is graphically formulated in Fig. 2, which exhibits the stated local conditional dependencies. To define a probabilistic motion model capturing these co-occurrence statistics, we consider the Maximum Entropy (ME) principle, also used in [20] for texture synthesis. The solution is a Gibbs distribution which can expressed in the following exponential form (see [4], [20] for details):

$$P_{\mathcal{M}}(x) = \frac{1}{Z_{\mathcal{M}}}\exp\Big[\Psi_{\mathcal{M}} \bullet \Gamma(x)\Big] \quad \text{with}\Psi_{\mathcal{M}} \bullet \Gamma(x) = \sum_{l=0}^{l=L} \Psi_{\mathcal{M}}^l \bullet \Gamma^l(x), \quad (5)$$

where $\Psi_{\mathcal{M}}^l \bullet \Gamma^l(x)$ is the dot product between the temporal (for $l = L$ by convention) or scale ($l \in [[0, L-1]]$) co-occurrence distribution $\Gamma^l(x)$ and potentials $\Psi_{\mathcal{M}}^l$ of the model $\mathcal{M}$: $\Psi_{\mathcal{M}}^l \bullet \Gamma^l(x) = \sum_{(\nu,\nu')\in\Lambda^2} \Psi_{\mathcal{M}}^l(\nu,\nu')\cdot\Gamma^l(\nu,\nu'|x)$. Potentials $\Psi_{\mathcal{M}} = \{\Psi_{\mathcal{M}}^l\} = \{\Psi_{\mathcal{M}}^l(\nu,\nu')\}$ explicitly specify the distribution $P_{\mathcal{M}}(x)$ associated to model $\mathcal{M}$. As given by Fig. 2, the key point is that this probabilistic modeling has an equivalent causal formulation in terms of products of conditional likelihoods [4]. Setting the following constraint on model potentials: $\forall(l,\nu') \in [[0,L]]\times\Lambda$, $\sum_{\nu\in\Lambda}\exp\Psi_{\mathcal{M}}^l(\nu,\nu') = 1$, $P_{\mathcal{M}}(x)$ is exactly given by:

$$P_{\mathcal{M}}(x) = \frac{1}{Z}\exp\Big[\Psi_{\mathcal{M}} \bullet \Gamma(x)\Big], \quad (6)$$

where the normalizing factor Z equals $|\mathcal{R}|^L$ and is finally independent on model $\mathcal{M}$.

Such a complete exponential formulation presents several interesting features. It makes the computation of the conditional likelihood $P_{\mathcal{M}}(x)$ for any sequence x and model $\mathcal{M}$ feasible and simple as explained below. It is not necessary to store the entire sequence x to evaluate the conditional likelihoods $\{P_{\mathcal{M}_i}(x)\}$ with regard to models $\{\mathcal{M}_i\}$ for a given sequence x. We only need to compute and store the corresponding temporal and scale co-occurrence distributions $\Gamma(x)$, and the evaluation of the likelihoods $\{P_{\mathcal{M}_i}(x)\}$ only requires the computation of $\{\Psi_{\mathcal{M}_i} \bullet \Gamma(x)\}$. Besides, motion recognition or classification can be straightforwardly formulated using the ML or MAP criterion. In addition, the nonpredefined parametric form of the resulting motion models allows us to characterize complex multimodal motion distribution.

4.2 ML Estimation of the Motion Models

Given a sequence of observations x, we estimate the potentials $\{\Psi^l_{\widehat{\mathcal{M}}}(\nu,\nu')\}_{(l,\nu,\nu')\in[[0,L]]\times\Lambda^2}$ of the motion model $\widehat{\mathcal{M}}$ which best fits x. We adopt the ML criterion, that is: $\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} P_{\mathcal{M}}(x)$. As detailed in [4] computing the ML model estimate merely involves the evaluation of the empirical mean of the observed temporal and scale transitions. This leads to:

$$\forall (l,\nu,\nu') \in [[0,L]] \times \Lambda^2,$$
$$\Psi^l_{\widehat{\mathcal{M}}}(\nu,\nu') = \ln \left. \Gamma^l(\nu,\nu'|x) / \sum_{\nu''\in\Lambda} \Gamma^l(\nu'',\nu'|x) \right). \quad (7)$$

The ML estimation of the nonparametric motion model associated to a sequence x is then directly derived from the computation of the set of temporal and scale co-occurrence distributions $\Gamma(x)$. For a given discrete state space Λ and a number L of scale levels, the complexity of the temporal and multiscale Gibbs models in terms of number of coefficients is equal to $(L+1)\cdot|\Lambda|^2$. For instance, if we consider 64 quantization levels for the motion-related measurements and three scale levels, each motion model will comprise 16,128 potentials. To reduce the model complexity and select the most informative potentials, we have adapted the technique proposed in [6] for temporal Gibbs models. It relies on the comparison of the ratios of the likelihood function corresponding to the full ML model and the reduced model to a predefined precision threshold λ. It proceeds by progressively introducing ML potentials in the reduced model in an appropriate order [6].

5 Motion Recognition

5.1 Training Stage

We can consider a supervised recognition task. We assume to be provided with a set $\mathcal{C}$ of qualitative motion classes, represented by different image sequences, including a training set $\mathcal{A}_c$ for each class $c \in \mathcal{C}$ and a test set $\mathcal{T}$. Given a class $c \in \mathcal{C}$, the learning stage consists in estimating the associated statistical motion model $\mathcal{M}_c$. For each element $a \in \mathcal{A}_c$, we compute the sequence of maps of multiscale motion-related measurements x^a and the related set of temporal and scale co-occurrence distributions $\Gamma(x^a)$. We then estimate the ML model $\mathcal{M}_c$ that best fits the observation set $\{x^a\}_{a\in\mathcal{A}_c}$. We solve for: $\mathcal{M}_c = \arg\max_{\mathcal{M}} \left[\prod_{a\in\mathcal{A}_c} P_{\mathcal{M}}(x^a)\right]$. Using the exponential expression of $P_{\mathcal{M}}(x^a)$ given by relation (6), we obtain:

$$\mathcal{M}_c = \arg\max_{\mathcal{M}} \left[\sum_{a\in\mathcal{A}_c} \Psi_{\mathcal{M}} \bullet \Gamma(x^a)\right] = \arg\max_{\mathcal{M}} \left[\Psi_{\mathcal{M}} \bullet \sum_{a\in\mathcal{A}_c} \Gamma(x^a)\right]. \quad (8)$$

Solving for (8) leads to the computation of the mean co-occurrence statistics over the set of co-occurrence matrices. If we denote by $\Gamma^l_c(\nu,\nu') = \sum_{a\in\mathcal{A}_c} \Gamma^l(\nu,\nu'|x^a)$, potentials are directly estimated from the mean co-occurrence matrix Γ_c using (7).

5.2 Classification Stage

Using the set of models $\{\mathcal{M}_c\}_{c\in\mathcal{C}}$, the motion recognition problem can be stated as a statistical inference issue based on the ML criterion. Given t in the test set $\mathcal{T}$, we compute its sequence of maps of multiscale motion-related measurements x^t and the associated temporal and scale co-occurrence distributions $\Gamma(x^t)$. To determine its motion class c^t, we again resort to the ML criterion: $c^t = \arg\max_{c\in\mathcal{C}} P_{\mathcal{M}_c}(x^t) = \arg\max_{c\in\mathcal{C}} [\Psi_{\mathcal{M}_c} \bullet \Gamma(x^t)]$. Let us again stress that this classification step only involves the computation of $|\mathcal{C}|$ dot products $\{\Psi_{\mathcal{M}_c} \bullet \Gamma(x^t)\}$ between model potentials $\{\Psi_{\mathcal{M}_c}\}_{c\in\mathcal{C}}$ and co-occurrence matrices $\Gamma(x^t)$.

6 Experiments

6.1 Experimental Set of Image Sequences

The motion recognition experiment we have conducted on real image sequences comprises eight classes. The set of processed video sequences involves different temporal textures, rigid motion situations, and human motion samples. More precisely, it contains four kinds of temporal textures: wind blown grass (a), gentle sea waves (b), rough water turbulence (c), and wind blown trees (d). A class of anchor shots (e) of low motion activity, and two classes related to rigid motion situations, moving escalator shots (f), and traffic sequences (g) are also added. The last class (h) refers to sequences of a pedestrian walking either from left to right or from right to left. Each motion class, except class (h), is represented by three sequences of 100 frames. Class (h) includes 10 sequences of 30 images (five shots involving a pedestrian moving from left to right and five ones with a pedestrian walking from right to left). Fig. 3 contains one image representative of every sequence of each class (for class (h), we have selected three sequences among 10).[1] We believe that, even if this motion recognition experiment can be regarded as somewhat "artificial," the provided set of real examples provides a realistic (while controlled for objective evaluation purpose) and convincing benchmark involving a wide range of motion content of varying difficulty with classes not easy to discriminate (from their motion content, as the semantic interpretation of the scene hast not been considered here). It could be compared to a certain extent to the Brodatz image set used for the evaluation of texture analysis methods.

Each image sequence of the video set described above is divided into "microsequences" of six images. We thus obtain 57 samples in each motion class, which means that we consider a set of 456 microsequences. The first 10 microsequences of the first sequence of each class (a) to (g) are used as the training data. For class (h), since the sequences contain only 30 frames, the first five subsequences of the first two sequences of this class are included in the training set. Finally, we obtain a training set formed by 80 microsequences, and a test set including 376 microsequences.

6.2 Motion Recognition Results

All the experiments have been conducted with the same parameter settings. The Markovian quantization of motion-related measurements involve 64 levels within range $[0,8]$. The model complexity reduction step, with λ set to 0.99, results in the retention of only about 10 percent of significant model potentials over about 1,000 coefficients for each set of model potentials $\Psi_{\mathcal{M}}$. In terms of computational time, the quantization step requires about one second to process a pair of 256×256 images using a 500MHz workstation, while the computation of co-occurrence statistics and model estimation are completed in less than one second. Let us point

1. The authors are grateful to INA, Departement Innovation, Direction de la Recherche, for providing the news sequences, C.H. Peh and L.F. Cheong at the National University of Singapore for providing temporal texture samples, and E. Bruno and D. Pellerin from INPG/LIS for providing human motion sequences.

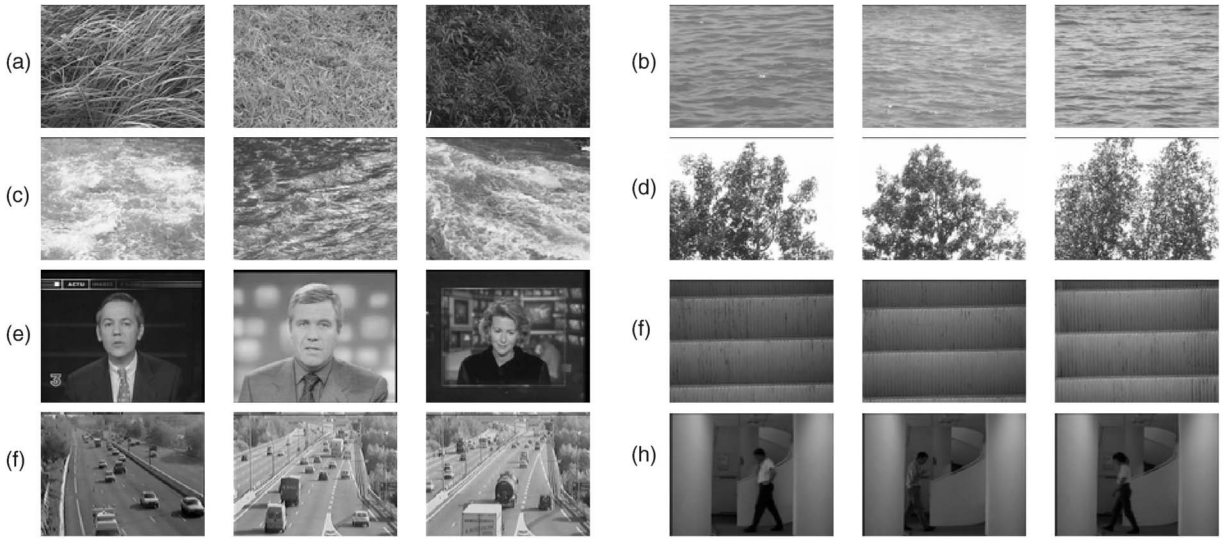

Fig. 3. Experimental video set: for every motion class, (a)-(h), one image is displayed for each sequence of the motion class. The eight classes correspond to various dynamic contents: (a) wind blown grass, (b) gentle sea waves, (c) rough turbulent water, (d) wind blown trees, (e) anchor person, (f) moving escalator, (g) traffic scene, and (h) pedestrian walking.

out that no multiscale information is used if $L = 0$. In that case, no spatial aspect of motion content is captured. It corresponds to the simple Temporal Gibbs Models (TGM) introduced in [6]. In these experiments, the motion recognition method is referred to as the TMGM method when considering the Temporal Multiscale Gibbs Models with $L \geq 1$, and as the TGM method otherwise. The comparison between the two methods will judge the improvement brought by the combined characterization of both spatial and temporal aspects of motion content conveyed by the temporal multiscale modeling.

The plot in Fig. 4b depicts the average τ and the standard deviation $\Delta\tau$, over the eight motion classes, of the correct classification rate for the elements of the test set $\mathcal{T}$ using the TGM method and the TMGM method for 1 to 4 scale levels with reduced models ($\lambda = 0.99$). The average recognition rate τ is greater than 95 percent using TMGM, whereas we get only 92.4 percent of correct classification using TGM. The best results are obtained using TMGM with $L = 3$ for which the mean classification rate is higher than 99 percent with a standard deviation lower than 1. Exploiting both spatial and temporal properties of motion content with the proposed multiscale framework outperforms the TGM method. In addition, the average rate τ decreases when L is greater than 3. This is due to the combination of two factors. First, the values of the elements close to the diagonal in the scale co-occurrence matrices become higher over scale. Second, the more the number L of scale levels, the less influential the motion information captured by the temporal co-occurrences.

Table 1 supplies a detailed evaluation of the recognition results obtained with the TGM method and the TMGM method for $L = 3$. For both methods, we report the percentage of correct and false classification for every motion class. The comparison of the results demonstrates that the TMGM method outperforms the TGM method for all classes. The correct classification rate is indeed always greater than 97 percent for the TMGM method, whereas it is between 69.6 percent and 100 percent using the TGM method. The most significant improvements are obtained for classes (A) and (E), for

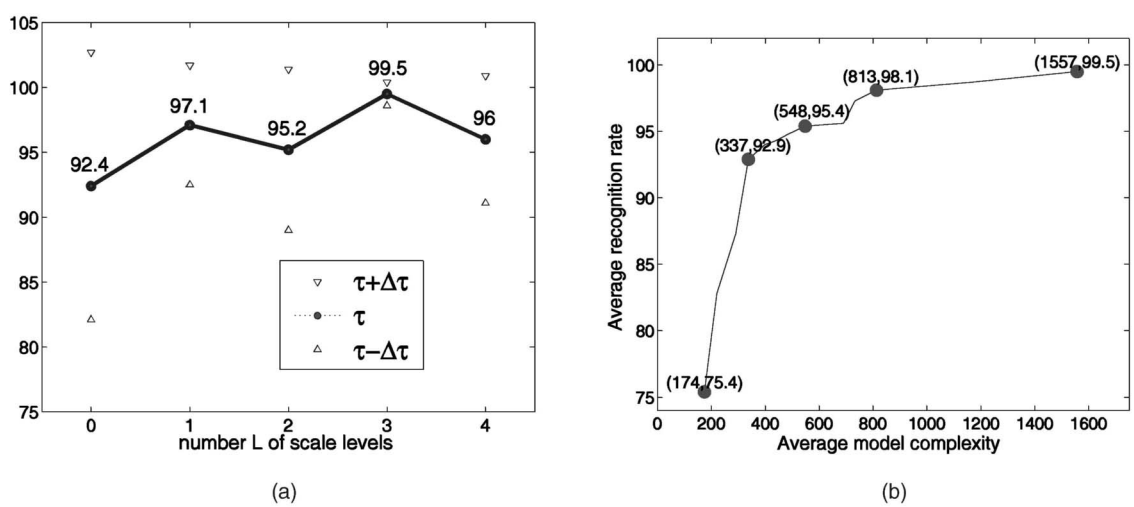

Fig. 4. Motion recognition results for the video set presented in Fig. 3. (a) Motion recognition using Temporal Multicale Gibbs Models (TMGM) with $L \in [[1, 4]]$ and Temporal Gibbs Models (TGM) ($L = 0$). We use reduced models with $\lambda = 0.99$. (We report the average τ and the standard deviation $\Delta\tau$ of the correct classification rate computed over the eight motion classes). (b) Motion recognition using reduced models. (We plot the average recognition rate versus the average model complexity; model complexity reduction is achieved using the technique described in Section 4.2 for different values of precision parameter λ within $[0.5, 0.99]$).

TABLE 1
Percentage of Correct and False Classification for the Eight Considered Motion Classes

	A	B	C	D	E	F	G	H
A	**97.9**		**2.1**					
	83.0	*4.3*						*12.7*
B		**100.**						
		100.						
C			**100.**					
			100.					
D				**97.9**				**2.1**
				91.5	*2.1*		*6.4*	

	A	B	C	D	E	F	G	H
E					**100.0**			
	2.1			*28.3*	*69.6*			
F						**100.**		
		2.1				*97.9*		
G							**100.**	
							100.0	
H								**100.0**
					2.4			*97.6*

For each class, we report results obtained using TGM and TMGM methods with $L = 3$. For each class, the first line (bold type) refers to the TMGM method (for instance, for class (A), the percentage of samples assigned to class (A) and (C) were, respectively, 97.9 percent and 2.1 percent using TMGM), while experiments conducted with the TGM method are reported on the second line (italic type).

which the correct classification rate increases, respectively, from 83 percent to 97.9 percent and from 69.6 percent to 100 percent. 28.3 percent of test samples of class (E) are wrongly classified into class (D) with the TGM method. Let us point out that microsequences of class (E) involve a low motion activity with small displacements of the anchor person and the tree sequences of class (D) include fluttering leaves with motion of rather low magnitudes. The spatial aspects of motion content captured by the TMGM method allows us to perfectly discriminate elements from classes (D) and (E).

The plot in Fig. 4a presents motion recognition results when considering different model reduction rates. We have plotted the average recognition rate versus the average model complexity. Different model reductions were achieved with values of the precision threshold λ in the range $[0.5, 0.99]$. Not surprisingly, the higher the model reduction, the weaker the average recognition rate. It is demonstrated that keeping about 1,600 model potentials is sufficient to obtain recognition results equivalent to those obtained with the ML models (i.e., a mean recognition rate of 99.5 percent in both cases). It represents an important model reduction since it corresponds to select only 10 percent of the ML model potentials. We have thus designed an efficient yet parsimonious motion models that successfully achieve motion recognition in quite general situations.

These experiments focus on the evaluation of TMGM versus TGM for motion recognition. In [4], we have carried out complementary experiments which favorably compare TMGM to other approaches (distance between global features extracted from temporal co-occurrences, spatio-temporal random walks evaluating both spatial and temporal co-occurrences). Gaussian mixture models applied to co-occurrence distributions were also investigated. While exhibiting a very low model complexity (about 50 parameters), the use of Gaussian mixtures also significantly degrade recognition results (mean rate lower than 90 percent).

7 Conclusion

We have presented a unified nonparametric statistical motion modeling framework in order to characterize motion content within image sequences. The introduction of temporal multiscale Gibbs models specified from co-occurrence statistics of properly quantized local motion-related measurements, computed over the processed image sequence, allows us to properly handle both spatial and temporal aspects of the underlying motion configuration. In addition, our probabilistic approach makes the computation of likelihood functions and ML model estimation for motion classification and recognition issues feasible and simple, which results in an efficient and low cost implementation.

Our method is able to successfully handle a wide range of dynamic contents, from rigid motion to temporal textures. Satisfactory results have been obtained concerning motion recognition over a representative set of real image sequences, demonstrating the interest of considering nonparametric motion characterization.

Acknowledgments

The authors would like to thank the reviewers for their helpful comments to improve the quality of the paper and R. Stufford for checking the syntax of the paper. The work was carried out while the R. Fablet was with IRISA and was supported by a grant supplied by CNRS and Brittany Council.

References

[1] E. Ardizzone and M. La Cascia, "Video Indexing Using Optical Flow Field," *Proc. Third IEEE Int'l Conf. Image Processing*, pp. 831-834, Sept. 1996.

[2] J.S. De Bonet and P. Viola, "Texture Recognition Using a Non-Parametric Multi-Scale Statistical Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 641-647, June 1998.

[3] Y. Deng and B.S. Manjunath, "Content-Based Search of Video Using Color, Texture and Motion," *Proc. Fourth IEEE Int'l Conf. Image Processing*, pp. 543-547, Oct. 1997.

[4] R. Fablet, "Modélisation Statistique non Paramétrique et Reconnaissance du Mouvement dans des Sequences d'Images; Application à l'Indexation Vidéo," PhD thesis, Univ. of Rennes 1, Irisa no. 2526, July 2001.

[5] R. Fablet and P. Bouthemy, "Motion-Based Feature Extraction and Ascendant Hierarchical Classification for Video Indexing and Retrieval," *Proc. Third Int'l Conf. Visual Information Systems*, pp 221-228, June 1999.

[6] R. Fablet, P. Bouthemy, and P. Perez, "Non Parametric Motion Characterization Using Causal Probabilistic Models for Video Indexing and Retrieval," *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 393-407, 2002.

[7] A. Fitzgibbon, "Stochastic Rigidity: Image Registration for Nowhere-Static Scenes," *Proc. Seventh IEEE Int'l Conf. Computer Vision*, pp. 662-670, July 2001.

[8] F. Heitz and P. Bouthemy, "Multimodal Estimation of Discontinuous Optical Flow Using Markov Random Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1217-1232, Dec. 1993.

[9] B. Horn and B. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, nos. 1-3, pp. 185-203, 1981.

[10] P. Huber, *Robust Statistics*. John Wiley & Sons, 1981.

[11] G. Johansonn, "Visual Perception of Biological Motion and a Model for Its Analysis," *Perception and Physics*, vol. 14, pp. 201-211, 1973.

[12] R. Nelson and R. Polana, "Qualitative Recognition of Motion Using Temporal Texture," *Computer Vision, Graphics, and Image Processing*, vol. 56, no. 1, pp. 78-99, 1992.

[13] J.M. Odobez and P. Bouthemy, "Separation of Moving Regions from Background in an Image Sequence Acquired with a Mobile Camera," *Video Data Compression for Multimedia Computing*, H.H. Li, S. Sun, and H. Derin, eds., chapter 8, pp. 295-311, Kluwer, 1997.

[14] C.-H. Peh and L.-F. Cheong, "Exploring Video Content in Extended Spatio-Temporal Textures," *Proc. Workshop Content-Based Multimedia Indexing*, pp. 147-153, Oct. 1999.

[15] J. Portilla and E. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *Int'l J. Computer Vision*, vol. 40, no. 1, pp. 49-70, 2000.

[16] P. Saisan, G. Doretto, Y. Wu, and S. Soatto, "Dynamic Texture Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Dec. 2001.

[17] M. Szummer and R.W. Picard, "Temporal Texture Modeling," *Proc. Third IEEE Int'l Conf. Image Processing*, pp. 823-826, Sept. 1996.

[18] R.P. Wildes and J.R. Bergen, "Qualitative Spatiotemporal Analysis Using an Oriented Energy Representation," *Proc. Sixth European Conf. Computer Vision*, pp. 768-784, June 2000.

[19] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Dec. 2001.

[20] S.C. Zhu, Y. Wu, and D. Mumford, "Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling," *Int'l J. Computer Vision*, vol. 27, no. 2, pp. 107-126, 1998.