Non parametric motion recognition using temporal multiscale Gibbs models

R. Fablet[1] and P. Bouthemy[2]
[1]IRISA/CNRS [2]IRISA/INRIA
Campus universitaire de Beaulieu 35042 Rennes Cedex, France
e-mail: {rfablet,bouthemy}@irisa.fr

Abstract

We present an original approach for non parametric motion analysis in image sequences. It relies on the statistical modeling of distributions of local motion-related measurements computed over image sequences. Contrary to previously proposed methods, the use of temporal multiscale Gibbs models allows us to handle in a unified statistical framework both spatial and temporal aspects of motion content. The important feature of our probabilistic scheme is to make the exact computation of conditional likelihood functions feasible and simple. It enables us to straightforwardly achieve model estimation according to the ML criterion and to benefit from a statistical point of view for classification issues. 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.

1 Introduction

The interpretation of motion cues is at the core of visual perception [2]. In the field of computer vision, research devoted to motion analysis was initially dedicated to the complete recovery of motion information from image sequences and relied on the computation of dense optic flow fields, which is known to be an ill-posed problem [1, 10]. However, as emphasized in [6], it is not necessary to recover such complete information to further analyze the dynamic content in image sequences. The key point for given applications, such as motion classification [12] or action recognition [4], is rather to determine appropriate representation of motion information directly computed from images. In this paper, we follow this point of view and we tackle the motion recognition issue with no *a priori* knowledge on the content of the observed dynamic scenes. Our goal is then to design a general framework to provide a global characterization of motion content in image sequences.

As far as general dynamic content classification is concerned, the use of non parametric techniques as opposed to 2D parametric motion models appears attractive. In that context, Nelson and Polana [12] introduced the notion of temporal textures. They considered techniques originally developed for spatial texture analysis to characterize distributions of local motion-related measurements computed over image sequences. The resulting description of dynamic scenes can be interpreted in terms of motion activity. New developments in that direction have been proposed for motion-based video indexing and retrieval [5, 14].

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. We rely on temporal multiscale Gibbs models to represent distributions of local motion-related measurements. These statistical non parametric motion models are exploited for motion recognition and we handle 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 underlying our work. Section 3 presents the local motion-related measurements we use for non parametric motion modeling. In Section 4, the statistical modeling of motion information and the estimation of these models are addressed. Section 5 presents the application to motion classification and, Section 6 contains concluding remarks.

2 Problem statement

Proposed approaches for non parametric motion analysis mainly rely on techniques originally developed for texture analysis. For instance, motion-based features from spatial cooccurrences of normal flow fields were exploited in [12] to classify sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In [14], different motion-based descriptors still computed from normal flow fields were considered using other techniques developed for texture analysis (Fourier spectrum, difference statistics). In both cases, the extracted features only yield a global characterization on the spatial distribution of motion information in a given image (i.e., at a given instant). However, as far as the description of the dynamic content in image sequences is concerned, it is also crucial to handle the temporal properties of image motion distribution. Thus, it appears necessary to combine the characterizations of both spatial and temporal aspects of motion information to achieve motion recognition. This can for instance be performed using spatio-temporal Gabor filters applied to image intensities as explored in [17].

On the other hand, the introduction of probabilistic models, such as Gibbs random fields [7, 18], has led to important advances in texture analysis. In particular, statistical techniques appear more suited to properly formalize learning and classification issues. Therefore, we have further investigated the analogy between texture analysis and non parametric motion analysis, and we introduce statistical non parametric motion models. The use of probabilistic models for temporal texture synthesis has been investigated in [16]. However, the considered auto-regressive models cannot be applied to motion modeling and recognition.

We prefer to rely on Gibbs models since there is a straightforward relationship between cooccurrence measurements and Gibbs models [8, 18]. Nevertheless, the direct use of general Gibbs models for recognition and classification issues reveals impossible. Indeed, their associated likelihood function cannot be exactly computed due to the unknown partition function, and then, we cannot compare the conditional likelihoods of given observations w.r.t. two different models. To alleviate this problem, we consider Gibbs models associated to a causal formulation. It allows us to exactly and easily compute the corresponding likelihood functions. We finally introduce temporal multiscale Gibbs models specified over sequences of maps of motion-related quantities. This multiscale approach enables us to define causal models which handle within a single statistical framework both spatial and temporal aspects of image motion information.

3 Local motion-related quantities

3.1 Local motion-related measurement

Our approach for non parametric motion analysis relies on the statistical modeling of distributions of local motion-related measurements. As previously stressed, dense optic flow field estimation remains a difficult issue, especially for complex dynamic scenes such as temporal textures. As a consequence, we resort to local motion-related quantities directly computed from the spatio-temporal derivatives of the intensity function [12, 14]. The Optic Flow Constraint Equation (OFCE) relates these derivatives to the real displacement $\mathbf{w}(p)$ at point p by assuming brightness constancy along trajectories [10]:

$$\mathbf{w}(p) \cdot \nabla I(p) + I_t(p) = 0 \qquad (1)$$

where ∇I is the spatial gradient of the intensity function I and I_t its temporal derivative.

From equation (1), we can deduce the expression of the normal flow, $v_n(p) = -I_t(p)/\|\nabla I(p)\|$ which is exploited in [12, 14]. However, this quantity is known to be very sensitive to the noise attached to the computation of the intensity gradient ∇I. To overcome this problem, we consider a weighted average of normal flows within a local window. The weights are given by the spatial intensity gradient norms, which are a relevant measure of normal flow reliability as pointed in [13]. Thus, we compute the following local motion-related measurement which is more reliable than normal flow:

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

where $\mathcal{F}(p)$ is a 3×3 window centered on p, η^2 a predetermined constant related to the noise level (typically, $\eta = 5$).

Obviously, we have lost any direction information by considering the measure $v_{obs}(p)$. For instance, we will not be able to discriminate two translations with different directions. However, we are not interested in determining specific motion values, but we aim at supplying a global characterization of the dynamic content within image sequences with a view to evaluating similarity in terms of motion activity. On the other hand, contrary to [12, 14], we do not exploit the direction information attached to normal flows. These directions are rather descriptors of the spatial texture present in the observed scene whereas we are concerned with a general description of motion content independent of spatial scene characteristics.

Another important advantage of this motion-related quantity is the existence of confidence bounds to evaluate its reliability. Given a detection level of motion magnitude δ, there are two bounds $l_\delta(p)$ and $L_\delta(p)$ verifying the following properties. If the motion-related measurement $v_{obs}(p)$ is smaller than to $l_\delta(p)$, the magnitude of the real (unknown) displacement $\|\mathbf{w}(p)\|$ at point p is lower than δ. On the contrary, if $v_{obs}(p)$ is higher than $L_\delta(p)$, $\|\mathbf{w}(p)\|$ is greater than δ. The two bounds $l_\delta(p)$ and $L_\delta(p)$ are straightforwardly computed from the spatial first-order derivatives of the intensity function at point p. For details on the expression of these bounds, we let the reader refer to [13].

The OFCE (1) is known to present several shortcomings. First, it can only handle displacements of rather small magnitudes. Second, it is no longer valid in occlusion regions over motion discontinuities, and even on sharp intensity discontinuities. To cope with these limitations, we have settle a multiscale scheme based on the statistical test designed in [9] to evaluate the validity of the OFCE (1). We first build

Gaussian pyramids for the pair of successive images to be processed. Then, at each point p, we select the finest scale for which the OFCE (1) is valid, and we compute at that scale the measurement $v_{obs}(p)$ and 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 Robust Markovian quantization

Our approach can be viewed as an extension of texture modeling for grey level images, where local motion-related quantities play a role similar to grey levels for texture analysis. One of the main differences lies in the continuous nature of the motion measurements. Different reasons lead us to quantize them. First, even if we consider continuous values in our modeling framework, we will need in practice to cope with discrete states for model estimation and storage. Second, for motion recognition, the definition of a quantization range common to all processed image sequences is required to evaluate similarities between image sequences. Third, we can exploit the confidence bounds of the local motion-related measurements to define an efficient quantization scheme.

The motion quantization issue is stated as a Markovian labeling problem. Compared to a simple linear quantization of the motion-related measurements, it presents several interests. First, the resulting quantized motion-related measurements can be regarded as approximations of the magnitude of the real (unknown) displacements. Let Λ denote the set of values of quantized motion-related measurements. Let us define $\Lambda = \{\nu_0 = 0, \nu_1, \nu_2, \ldots, \nu_{|\Lambda|}\}$ with $0 < \nu_1 < \ldots < \nu_{|\Lambda|}$. The Markovian quantization comes to determine the interval $[\nu_{i-1}, \nu_i]$ within which the magnitude of the real (unknown) displacement at point p is the more likely to be. This is evaluated through a data-driven term involving the motion-related measurement $v_{obs}(p)$ and the associated confidence bounds $\{(l_{\nu_i}(p), L_{\nu_i}(p))\}$ described in subsection 3.1. In addition, the use of a contextual labeling technique enables us to cope with spurious local observations. Besides, experiments carried out on simple known motions (translation, rotation, divergence) have demonstrated that such a Markovian quantization provides us with quantized motion-related measurements closer to the magnitude of the real displacements, compared to a simple linear quantization. These comparisons were evaluated between the map of quantized motion-related measurements and the map of magnitudes of the real known displacements (ground-truth), in terms of mean square error and in terms of L_1 distance of the occurrence histograms.

Let $\mathcal{R}$ be the spatial image grid, $e = (e_p)_{p\in\mathcal{R}}$ the label field where each label takes its value in the set Λ, and $o = (v_{obs}(p))_{p\in\mathcal{R}}$ the observation field formed by the local motion-related measurements. To achieve the Markovian quantization, we adopt the MAP criterion. It comes to the minimization of a global energy function U [7]:

$$\begin{aligned} \widehat{e} &= \arg \min_{e\in\Lambda^{|\mathcal{R}|}} U(e,o) \\ &= \arg \min_{e\in\Lambda^{|\mathcal{R}|}} [U_1(e,o) + U_2(e)] \end{aligned} \quad (3)$$

where the energy function U is split 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 potentials V_1 and V_2:

$$\left\{ \begin{aligned} U_1(e,o) &= \sum_{p\in\mathcal{R}} V_1(e_p, v_{obs}(p)) \\ U_2(e) &= \sum_{(p,q)\in\mathcal{C}} \beta \cdot \rho(e_p - e_q) \end{aligned} \right. \quad (4)$$

where $\mathcal{C}$ denotes the set of binary cliques of the 4-connectivity neighborhood, β a positive coefficient setting the influence of the regularization (in practice, β is set to 2.0) and ρ a hard-redescending M-estimator, here Tukey's biweight function. It allows us to preserve the discontinuities present in the actual velocity field.

The potential function V_1 expresses how relevant a label is to describe a given motion quantity. Let us consider a quantization level ν_i with $i \in [\![1, |\Lambda|]\!]$, where $[\![1, |\Lambda|]\!]$ is the interval of discrete values comprised between 1 and $|\Lambda|$. The 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 defined as follows:

$$\begin{aligned} V_1(\nu_i, v_{obs}(p)) &= Sup_{L_{\nu_{i-1}}(p)}(v_{obs}(p)) \\ &+ Inf_{l_{\nu_i}(p)}(v_{obs}(p)) \end{aligned} \quad (5)$$

Sup_L is a continuous step function centered in L, and Inf_l is the opposite of a step function centered in l and rescaled to be in the interval $[0, 1]$.

The minimization issue (3) is achieved using a modified version of the ICM algorithm and the initialization is given by considering only the data-driven term in the minimization.

4 Statistical non parametric motion modeling

4.1 Temporal multiscale Gibbs models

In order to handle both the spatial and temporal aspects of the dynamic content of image sequences, we have designed a multiscale statistical framework. Given a sequence of maps of quantized motion-related measurements, we introduce at each point a vector of measurements computed at different scales instead of considering only one single value. Gibbs models are then specified on a sequence of maps of vectors of multiscale motion-related measurements. The proposed probabilistic models enable the exact and easy computation of the likelihood function attached to a given model. A direct model estimation scheme according to the Maximum Likelihood (ML) criterion can also be adopted.

Let $v = (v_0, v_1, \ldots, v_K)$ be a sequence of $K+1$ maps of quantized motion-related measurements issued from a sequence of $K+2$ frames. From this sequence v, we build a new sequence $x = (x_0, x_1, \ldots, x_K)$. For given instant $k \in [\![0, K]\!]$ and point p in the image support $\mathcal{R}$, $x_k(p)$ is defined as a vector of measures $(x_k^0(p), \ldots, x_k^L(p))$ at scales 0 to L which are computed by applying Gaussian filters of increasing variance to the map v_k at point p.

Our statistical modeling approach relies on the assumption that the sequence x is the realization of a first-order Markov chains $X = (X_0, \ldots, X_k)$ such that:

$$P_{\mathcal{M}}(x) = P_{\mathcal{M}}(x_0) \prod_{k=1}^{K} P_{\mathcal{M}}(x_k | x_{k-1}) \qquad (6)$$

$\mathcal{M}$ refers to the underlying motion model to be defined later. $P_{\mathcal{M}}(x_0)$ is the *a priori* distribution for the first image of the sequence. In practice, we will consider no specific *a priori*, i.e., $P_{\mathcal{M}}(x_0)$ is constant. Let $1/Z$ denote this constant. In order to design purely causal models, we assume that random variables $(X_k(p))_{p \in \mathcal{R}}$ at time k are independent conditionally to X_{k-1}. We further consider that, for given point p and instant k, $X_k(p)$ is also independent from $(X_{k-1}(q))_{q \in \mathcal{R} \setminus \{p\}}$ w.r.t. $X_{k-1}(p)$. Thus, $P_{\mathcal{M}}(x_k | x_{k-1})$ is given by:

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

For $(k, p) \in [\![1, K]\!] \times \mathcal{R}$, applying Bayes rule, we obtain:

$$\begin{gathered} P_{\mathcal{M}}\left(x_k(p) | x_{k-1}(p)\right) = \\ P_{\mathcal{M}}\left(x_k^0(p) | x_k^L(p), x_k^{L-1}(p), \ldots, x_k^1(p), x_{k-1}(p)\right) \\ \times \ldots \times P_{\mathcal{M}}\left(x_k^{L-1}(p) | x_k^L(p), x_{k-1}(p)\right) \\ \times P_{\mathcal{M}}\left(x_k^L(p) | x_{k-1}(p)\right) \end{gathered} \qquad (8)$$

Since $\{x_k^0(p), \ldots, x_k^L(p)\}$ are multiscale local motion-related measurements, accurate information is provided by quantities computed at the finest scales, whereas quantities attached to the coarsest levels convey more global and smooth information. In terms of conditional dependency, it leads to argue that, for any point p at instant k and scale $l \in [\![0, L-2]\!]$, $X_k^l(p)$ is independent from $X_k^{l+2}(p), ..., X_k^L(p)$ w.r.t. $X_k^{l+1}(p)$. Similarly, considering the conditional dependency of $X_k^l(p)$ w.r.t. $X_{k-1}(p) = \{X_{k-1}^0, \ldots, X_{k-1}^L(p)\}$, the most accurate information is supplied by the motion-related measurement $x_{k-1}^0(p)$ at scale 0. Thus, we also assume that $X_k^l(p)$ is conditionally independent of $\{X_{k-1}^1(p), \ldots, X_{k-1}^L(p)\}$ w.r.t. $X_{k-1}^0(p)$. Based on these two assumptions, expression (8) can be simplified as follows:

$$\begin{gathered} P_{\mathcal{M}}\left(x_k(p) | x_{k-1}(p)\right) = \\ P_{\mathcal{M}}\left(x_k^0(p) | x_k^1(p), x_{k-1}^0(p)\right) \\ \times \ldots \times P_{\mathcal{M}}\left(x_k^{L-1}(p) | x_k^L(p), x_{k-1}^0(p)\right) \\ \times P_{\mathcal{M}}\left(x_k^L(p) | x_{k-1}^0(p)\right) \end{gathered} \qquad (9)$$

This statistical setting involves the evaluation of “tri-occurrences”, which induces a high complexity to specify the model $\mathcal{M}$. Besides, we noticed in practice that scale cooccurrence distributions computed on pairs $\{(x_k^{l-1}(p), x_k^l(p))\}$ at two successive scales $l-1$ and l exhibit high values for the terms close to the diagonal, whereas temporal cooccurrence distributions computed on pairs $\{(x_k^l(p), x_{k-1}^0(p))\}$ are more widespread. As a consequence, temporal dependencies can be neglected w.r.t scale dependencies. The conditional likelihood $P_{\mathcal{M}}\left(x_k(p) | x_{k-1}(p)\right)$ is finally written as:

$$\begin{gathered} P_{\mathcal{M}}\left(x_k(p) | x_{k-1}(p)\right) = \\ P_{\mathcal{M}}\left(x_k^0(p) | x_k^1(p)\right) \times \ldots \times P_{\mathcal{M}}\left(x_k^{L-1}(p) | x_k^L(p)\right) \\ \times P_{\mathcal{M}}\left(x_k^L(p) | x_{k-1}^0(p)\right) \end{gathered} \qquad (10)$$

Thus, we only evaluate cooccurrences either computed at successive scales or at two successive instants between scales 0 and L. Let us point out that cooccurrence statistics computed between successive scales have proven interesting properties for texture analysis and synthesis [3, 11, 15].

In order to deliver an exponential formulation of the likelihood $P_{\mathcal{M}}(x)$, we introduce the following notations:

$$P_{\mathcal{M}}\left(x_k^L(p) | x_{k-1}^0(p)\right) \propto \exp \Psi_{\mathcal{M}}^L\left(x_k^L(p), x_{k-1}^0(p)\right) \qquad (11)$$

and $\forall l \in [\![0, L-1]\!]$:

$$P_{\mathcal{M}}\left(x_k^{L-1}(p) | x_k^L(p)\right) \propto \exp \Psi_{\mathcal{M}}^l\left(x_k^l(p), x_k^{l+1}(p)\right) \qquad (12)$$

where $\Psi_{\mathcal{M}} = \left\{\Psi_{\mathcal{M}}^l(\nu, \nu')\right\}_{(l, \nu, \nu') \in [\![0, L]\!] \times \Lambda^2}$ are the potentials which explicitly specify model $\mathcal{M}$. To guarantee the uniqueness of the potentials associated to $P_{\mathcal{M}}$, we impose the following normalization constraint:

$$\forall (l, \nu') \in [\![0, L]\!] \times \Lambda, \ \sum_{\nu \in \Lambda} \exp \Psi_{\mathcal{M}}^l(\nu, \nu') = 1 \qquad (13)$$

Using model potentials, $P_{\mathcal{M}}(x)$ is given by:

$$P_{\mathcal{M}}(x) = \frac{1}{Z}\exp\left[\sum_{k=1}^{K}\sum_{p\in\mathcal{R}}\Psi_{\mathcal{M}}(x_k(p), x_{k-1}(p))\right] \quad (14)$$

where $\Psi_{\mathcal{M}}(x_k(p), x_{k-1}(p))$ is the sum of temporal and scale potentials:

$$\begin{aligned}\Psi_{\mathcal{M}}(x_k(p), x_{k-1}(p)) = {} & \Psi^L_{\mathcal{M}}(x^L_k(p), x^0_{k-1}(p)) \\ & + \sum_{l=0}^{L-1}\Psi^l_{\mathcal{M}}(x^l_k(p), x^{l+1}_k(p))\end{aligned} \quad (15)$$

Specifying $\Psi_{\mathcal{M}}$ supplies the complete knowledge of $P_{\mathcal{M}}$. This provides us with a general statistical framework for motion recognition. Besides, we can argue from expression (14) that the introduced model $\mathcal{M}$ is a Gibbs random field for which the partition function is known and equals Z. Let us stress that Z is independent of the considered model.

We can rewrite the expression (14) using temporal and scale cooccurrences. We obtain a simple expression of the likelihood $P_{\mathcal{M}}(x)$ involving the computation of a dot product, $\Psi_{\mathcal{M}} \bullet \Gamma(x)$, between the potentials associated with the model $\mathcal{M}$ and the set of temporal and scale cooccurrence distributions $\Gamma(x)$ computed for the sequence of multiscale motion-related quantities x:

$$\begin{aligned}P_{\mathcal{M}}(x) &= \frac{1}{Z}\cdot\exp\Big[\Psi_{\mathcal{M}} \bullet \Gamma(x)\Big] \\ \text{with } \Psi_{\mathcal{M}} \bullet \Gamma(x) &= \sum_{l=0}^{l=L}\Psi^l_{\mathcal{M}} \bullet \Gamma^l(x)\end{aligned} \quad (16)$$

where $\Psi^l_{\mathcal{M}} \bullet \Gamma^l(x)$ is the dot product between the temporal ($l = L$) or scale ($l \in [\![0, L-1]\!]$) cooccurrence distributions and the potentials of model $\mathcal{M}$. The temporal cooccurrence distribution $\Gamma^L(x)$ is defined as: $\forall(\nu, \nu') \in \Lambda^2$,

$$\Gamma^L(\nu, \nu'|x) = \sum_{k=1}^{K}\sum_{p\in\mathcal{R}}\delta(\nu - x^L_k(p))\delta(\nu' - x^0_{k-1}(p)) \quad (17)$$

with δ the Kronecker symbol. The scale cooccurrence distribution $\Gamma^l(x)$ for $l \in [\![0, L-1]\!]$ is given by: $\forall(\nu, \nu') \in \Lambda^2$,

$$\Gamma^l(\nu, \nu'|x) = \sum_{k=1}^{K}\sum_{p\in\mathcal{R}}\delta(\nu - x^l_k(p))\delta(\nu' - x^{l+1}_k(p)) \quad (18)$$

For $l \in [\![0, L]\!]$, the dot product $\Psi^l_{\mathcal{M}} \bullet \Gamma^l(x)$ is expressed as:

$$\Psi^l_{\mathcal{M}} \bullet \Gamma^l(x) = \sum_{(\nu,\nu')\in\Lambda^2}\Psi^l_{\mathcal{M}}(\nu, \nu') \cdot \Gamma^l(\nu, \nu'|x) \quad (19)$$

The availability of an exponential formulation presents several interests. First, it makes the computation of the conditional likelihood $P_{\mathcal{M}}(x)$ for any sequence x and model $\mathcal{M}$ feasible and simple. Then, the use of these probabilistic models for recognition or classification issues based on ML or MAP criteria is straightforward. Second, all motion information exploited by these models is contained in the cooccurrence distributions. In particular, in order to evaluate the conditional likelihoods $\{P_{\mathcal{M}_i}(x)\}$ w.r.t. models $\{\mathcal{M}_i\}$ 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 and scale cooccurrence distributions $\Gamma(x)$. The evaluation of the conditional likelihoods $\{P_{\mathcal{M}_i}(x)\}$ is then simply achieved from the products $\{\Psi_{\mathcal{M}_i} \bullet \Gamma(x)\}$ using expression (16).

4.2 Maximum likelihood estimation

We now describe how we estimate the non parametric motion model $\mathcal{M}$ attached to a given image sequence. Given a sequence of multiscale motion-related measurements x, we estimate the potentials $\{\Psi^l_{\widehat{\mathcal{M}}}(\nu, \nu')\}_{(l,\nu,\nu')\in[\![0,L]\!]\times\Lambda^2}$ of the model $\widehat{\mathcal{M}}$ which best fits x. We resort to the ML criterion, which leads to solve for the following issue:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} P_{\mathcal{M}}(x) \quad (20)$$

Since the considered statistical formulation involves products of conditional likelihoods as given by relation (10), the ML model estimation only requires to compute them. The potentials of the ML model estimate $\widehat{\mathcal{M}}$ are given by: $\forall(l, \nu, \nu') \in [\![0, L]\!] \times \Lambda^2$,

$$\Psi^l_{\widehat{\mathcal{M}}}(\nu, \nu') = \log\left(\Gamma^l(\nu, \nu'|x)/\sum_{\nu''\in\Lambda}\Gamma^l(\nu'', \nu'|x)\right) \quad (21)$$

Thus, the ML estimation of the model associated with a sequence x is straightforward and directly results from the computation of the set of temporal and scale cooccurrence distributions $\Gamma(x)$. In addition, we can achieve model complexity reduction in order to supply an informative representation of the motion content while remaining parsimonious. After the ML estimation step, we select the relevant potentials by evaluating likelihood ratios as described in [5].

5 Application to motion-based image sequence classification

In order to demonstrate the ability of our non parametric statistical motion modeling framework to characterize and discriminate various motion types, we have carried out classification experiments over a set of image sequences involving a variety of motion contents (rigid motion, pedestrian walking, temporal textures).

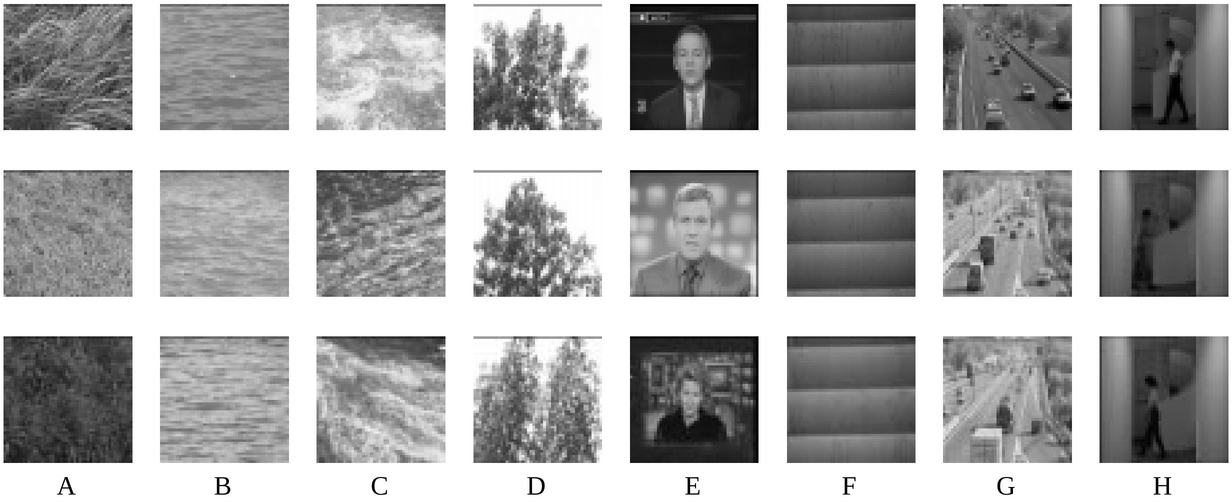

Figure 1: Experimental video set: for each of the eight motion classes (A) to (H), one image is displayed for each sequence of the motion class. These classes correspond to various dynamic contents: (A) wind blown grass, (B) gentle sea waves, (C) rough turbulent water, (D) wind blown trees, (E) anchor person, (F) moving escalator, (G) traffic, and (H) pedestrian walking.

5.1 Experimental set of image sequences

Our motion recognition experiments deal with eight motion classes. The set of video sequences comprises 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). In addition, a class of anchor shots (E) of low motion activity, and two classes of rather rigid motion situations, moving escalator shots (F) and traffic sequences (G), are included. The last class (H) refers to sequences of pedestrian walking either from left to right or from right to left.

Each motion class, except class (H), is represented by three sequences of one hundred frames. Class (H) includes ten sequences of thirty images (five shots involving a pedestrian moving from left to right and five ones for a pedestrian walking from right to left). Fig.1 contains one image representative of each sequence of each class (for class (I), we have selected three sequences).

5.2 Motion learning and recognition stages

Based on the eight motion classes, we first achieve a supervised learning stage using a training set of image sequences. Then, we carry out motion recognition experiments over a test set. These two sets are defined as follows.

Each image sequence of the set described above is divided into “micro-sequences” of six images. We obtain 57 samples in each motion class, which means that we consider a set of 456 micro-sequences. The first ten micro-sequences 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, we consider the first five subsequences of the first two sequences of this class. Finally, we obtain a training set comprising 80 micro-sequences, and a test set including 376 micro-sequences. Let $\mathcal{C}$ denote the set of eight motion classes, $\mathcal{A}_c$ the training set for a given class $c \in \mathcal{C}$, and $\mathcal{T}$ the set of test image sequences.

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 cooccurrence distributions $\Gamma(x^a)$. We then estimate the model $\mathcal{M}_c$ best fitting the observation set $\{x^a\}_{a\in\mathcal{A}_c}$ w.r.t. the ML criterion. We solve for:

$$\mathcal{M}_c = \arg\max_{\mathcal{M}} \left[\prod_{a\in\mathcal{A}_c} P_{\mathcal{M}}(x^a) \right] \tag{22}$$

Using the exponential formulation of $P_{\mathcal{M}}(x^a)$ given by relation (16), we obtain:

$$\mathcal{M}_c = \arg\max_{\mathcal{M}} \left[\sum_{a\in\mathcal{A}_c} \Psi_{\mathcal{M}} \bullet \Gamma(x^a) \right] \tag{23}$$

Since the dot product $\Psi_{\mathcal{M}} \bullet \Gamma(x^a)$ is linear w.r.t. the cooccurrence distribution $\Gamma(x^a)$, this expression leads to:

$$\mathcal{M}_c = \arg\max_{\mathcal{M}} \left[\Psi_{\mathcal{M}} \bullet \sum_{a\in\mathcal{A}_c} \Gamma(x^a) \right] \tag{24}$$

Thus, solving for (22) simply comes to determine the model best fitting the average cooccurrence distributions Γ_c over the set of cooccurrence distributions $\{\Gamma(x^a)\}_{a\in\mathcal{A}_c}$:

$$\mathcal{M}_c = \arg\max_{\mathcal{M}} \left[\Psi_{\mathcal{M}} \bullet \Gamma_c\right] \tag{25}$$

with: $\forall(l,\nu,\nu') \in [\![0,L]\!] \times \Lambda^2$,

$$\Gamma_c^l(\nu,\nu') = \sum_{a\in\mathcal{A}_c} \Gamma^l(\nu,\nu'|x^a) \tag{26}$$

Potentials $\Psi_{\mathcal{M}_c}$ are then directly computed from the cooccurrence distribution Γ_c using relation (21).

Using the set of statistical non parametric motion models $\{\mathcal{M}_c\}_{c\in\mathcal{C}}$, motion recognition is 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 cooccurrence distributions $\Gamma(x^t)$. To determine its motion class c^t, we again resort to the ML criterion:

$$\begin{aligned} c^t &= \arg\max_{c\in\mathcal{C}} P_{\mathcal{M}_c}(x^t) \\ &= \arg\max_{c\in\mathcal{C}} \left[\Psi_{\mathcal{M}_c} \bullet \Gamma(x^t)\right] \end{aligned} \tag{27}$$

This only involves the computation of eight dot products $\{\Psi_{\mathcal{M}_c} \bullet \Gamma(x^t)\}$ between model potentials $\{\Psi_{\mathcal{M}_c}\}_{c\in\mathcal{C}}$ and cooccurrence distributions $\Gamma(x^t)$.

5.3 Motion recognition experiments

All the experiments have been conducted using the following parameter setting. Quantization of motion-related measurements involve 64 levels within range $[0,8]$. We have considered different values of the number L of scale levels, from 0 to 4. The scheme used for model complexity reduction leads to keep only 10% to 20% of significant model potentials (over about 1000 potentials for each set of model potentials $\Psi_{\mathcal{M}}$).

Let us point out that no multiscale information is used if $L = 0$. In this case, no spatial aspect of motion content is captured. We will refer to these models with $L = 0$ as the Temporal Gibbs Models (TGM), whereas the models with $L \geq 1$ are called the Temporal Multiscale Gibbs Models (TMGM). In the sequel, the associated method for motion recognition are resp. denoted as the TGM method and the TMGM method. The comparison between these two methods will allow us to evaluate the interest of the combined characterization of spatial and temporal aspects of motion content through the considered multiscale modeling.

In Fig.2, we plot the average τ and the standard deviation $\Delta\tau$, over the eight motion classes, of the correct classification rate obtained for the elements of the test set $\mathcal{T}$. We report results obtained using TMG and TMGM with 1 to 4 scale levels. Average rate τ is greater than 95% using TMGM, whereas we get only 92.4% 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% with a standard deviation lower than 1. Thus, the explicit combination of spatial and temporal modeling of motion information through the proposed multiscale framework outperforms the TGM method. Besides, the average rate τ decreases when L is greater than to 3. This is due to the combination of two elements. First, the values of the terms close to the diagonal in scale cooccurrence distributions $\Gamma^l(x)$ become higher over scale. Second, the more the number L of scale levels increases, the less influential the motion information captured by the distribution of temporal cooccurrences $\Gamma^L(x)$ is.

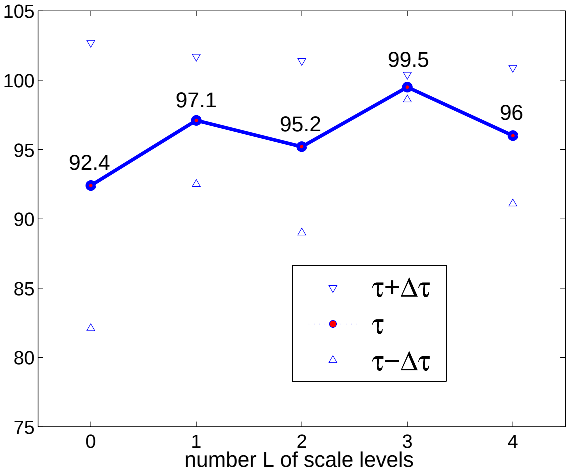

Figure 2: Motion recognition results for the video base presented in Fig. 1 using Temporal Multiscale Gibbs models (TMGM) with $L \in [\![1,4]\!]$ and Temporal Gibbs Models (TGM) ($L = 0$). We report the average τ and the standard deviation $\Delta\tau$ of the correct classification rate computed over the eight motion classes.

Table 1 provides a detailed evaluation of the recognition results obtained using the TGM method and the TMGM method with $L = 3$. In both cases, we report the percentage of correct and false classification for each motion class. The comparison of the results shows that the TMGM method outperforms the TGM method for all classes. The correct classification rate is indeed always greater than to 97% using the TMGM method, whereas it is comprised between 69.6% and 100% using the TGM method. The most significant improvements are obtained for classes (A) and (E), for which the correct classification rate increases respectively from 83% to 97.9% and from 69.6% to 100%. In the last case, 28.3% of test samples of class (E) are wrongly classified into class (D). Let us point out that micro-sequences 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 handling of spatial aspects of motion distribution using TMGM allows us to perfectly discriminate elements from classes (D) and (E).

	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*	
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*

Table 1: Percentage of correct and false classification for the eight considered motion. For each class, we report results obtained using TGM and TMGM 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 resp. 97.9% and 2.1% using TMGM), whereas experiments conducted with the TGM method are reported on the second line (italic type).

6 Conclusion

We have presented a unified non parametric statistical motion modeling framework which copes with both temporal and spatial aspects of dynamic scenes. It relies on temporal multiscale Gibbs models of distributions of local motion-related motion measurements. It can be straightforwardly exploited for motion recognition, since the complete evaluation of conditional likelihood functions is easy. Model estimation proceeds from ML criteria. We have shown that the designed statistical framework can be applied to perform supervised motion classification.

Our non parametric method is able to handle a wide range of dynamic contents, from rigid motion to temporal textures. Quite satisfactory results have been obtained in motion recognition over a representative set of image sequences. This demonstrates the interest of considering non parametric motion characterization. Furthermore, the use of temporal multiscale models allows us to capture in an easy and efficient way both spatial an temporal aspects of image motion structure.

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 National University of Singapore for supplying the sequences of temporal textures.

References

[1] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proc. of the IEEE*, 76(8):869–890, 1988.

[2] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Phil. Trans. Royal Society London B*, pages 1257–1265, 1997.

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

[4] J.W. Davis and A. Bobick. The representation and recognition of human movement using temporal templates. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'97*, pages 928–934, Porto-Rico, June 1997.

[5] R. Fablet, P. Bouthemy, and P. Pérez. Statistical motion-based video indexing and retrieval. In *Proc. of 6th Int. Conf. on Content-Based Multimedia Information Access, RIAO'2000*, pages 602–619, Paris, Apr. 2000.

[6] C. Fermuller and Y. Aloimonos. Vision and action. *Image and Vision Computing*, 13(10):725–744, 1995.

[7] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on PAMI*, 6(6):721–741, 1984.

[8] G.L. Gimel'Farb. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. on PAMI*, 18(11):1110–1114, 1996.

[9] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. on PAMI*, 15(2):1217–1232, 1993.

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

[11] J. Huang and D. Mumford. Statistics of natural images and models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'99*, pages 541–547, Fort Collins, June 1999.

[12] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP*, 56(1):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. In *Video Data Compression for Multimedia Computing*, chapter 8, pages 295–311. H. H. Li, S. Sun, and H. Derin, eds, Kluwer, 1997.

[14] C.-H. Peh and L.-F. Cheong. Exploring video content in extended spatio-temporal textures. In *Workshop on Content-Based Multimedia Indexing, CBMI'99*, pages 147–153, Toulouse, France, Oct. 1999.

[15] J. Portilla and E. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. Jal of Comp. Vis.*, 40(1):49–70, 2000.

[16] 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.

[17] R.P. Wildes and J.R. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *Proc. of 6th Eur. Conf. on Computer Vision, ECCV'2000*, pages 768–784, Dublin, June 2000.

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