Motion recognition using spatio-temporal random walks in sequence of 2D motion-related measurements

Ronan Fablet, Patrick Bouthemy

IN PROC. OF INTERNATIONAL CONFERENCE ON IMAGE PROCESSING, ICIP'2001, THESSALONIQUE, OCTOBER 2001

MOTION RECOGNITION USING SPATIO-TEMPORAL RANDOM WALKS IN SEQUENCE OF 2D MOTION-RELATED MEASUREMENTS

Ronan Fablet[1] *and Patrick Bouthemy*[2]

[1]IRISA/CNRS [2]IRISA/INRIA

Campus universitaire de Beaulieu 35042 Rennes Cedex, France

{rfablet,bouthemy}@irisa.fr – www.irisa.fr/vista/Vista.english.html

ABSTRACT

This paper describes an original approach for non parametric motion analysis in image sequences. It relies on a statistical modeling of distributions of local motion-related measurements, computed over image sequences, resulting from spatio-temporal random walks. It handles in a single probabilistic framework both spatial and temporal properties of motion content. The important feature of our method is to make feasible the exact computation of conditional likelihood functions. We have carried out motion recognition experiments over a large set of real image sequences comprising various motion types.

1. INTRODUCTION AND PROBLEM STATEMENT

As far as general dynamic image content recognition is concerned, the use of non parametric techniques as opposed to 2D parametric motion models appears quite relevant. In that context, the pioneering work of Nelson and Polana [4] introduced the notion of temporal textures which refer to complex motion types (such as moving crowds, river flows or wind blown trees), and relied on techniques originally developed for spatial texture analysis to characterize motion content. Global motion-based features computed from spatial cooccurrences of normal flow fields were indeed exploited to classify sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In [5], new spatial motion activity descriptors, still computed from normal flow fields, were added using Fourier spectrum and difference statistics. To handle temporal aspects of motion content, we have defined in [1] temporal motion-based features determined from temporal cooccurrences of local motion-related measurements.

Further investigating the analogy with texture analysis and exploiting the correspondence between cooccurrence measurements and probabilistic models [6], we have presented in [2] a non parametric statistical motion modeling framework using temporal Gibbs models. Compared to feature-based methods, statistical approaches appear more suited to properly formalize learning and classification stages and to cope with noise and uncertainty. One of the main advantages of our statistical scheme was to make feasible an exact computation of conditional likelihood functions, and then, to achieve in a simple and efficient manner model estimation. On the other hand, its main shortcoming was to discard spatial properties of motion information. To handle within a single statistical modeling framework both temporal and spatial aspects of dynamic content, we exploit spatio-temporal random walks within successive maps of motion-related measurements. It allows us to keep the crucial properties of temporal Gibbs models in terms of exact computation of conditional likelihood functions and in terms of model estimation efficiency.

The remainder of this paper is organized as follows. Section 2 presents the local motion-related measurements we utilize for non parametric motion activity modeling. In Section 3, the statistical modeling of motion information and the issue of estimating these models are addressed. Section 4 presents the application to motion classification and experimental results. Section 5 contains concluding remarks.

2. LOCAL MOTION-RELATED MEASUREMENTS

Our approach for non parametric motion analysis relies on a statistical modeling of distributions of local motion-related measurements. Dense optic flow field estimation remain difficult and costly, especially for complex dynamic scenes such as temporal textures. We prefer to consider local motion-related quantities directly evaluated from the spatio-temporal derivatives of the intensity function. The local motion-related measurement already used in [2] is given at pixel p by:

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

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

intensity function I, and ∇I its spatial gradient. Quantity $v_{obs}(p)$ is a weighted local average of the normal flow $v_n(p) = |I_t(p)|/\|\nabla I(p)\|$. It has proven more reliable than normal flow, used in [4, 5], which is known to be sensitive to noise attached to the computation of the spatio-temporal derivatives of the intensity function.

Obviously, when considering this local measure, we have lost any motion direction information. However, in the context of motion recognition, this is not a real shortcoming as stressed in previous work [1, 2], since we are rather interested in discriminating general motion types. Contrary to [4, 5], we do not exploit the direction information attached to normal flows. As a matter of fact, direction information rather reflects the spatial texture pattern present in the observed scene, whereas we aim at providing a general description of motion content almost independent of these spatial characteristics.

Our statistical modeling of motion activity requires to quantize the continuous motion-related measurements $\{v_{obs}(p)\}$. To cope with erroneous values, we apply a quantization on a predefined interval. It indeed appears relevant to introduce a limit beyond which measures are no more regarded as usable. In practice, we use 16 quantization levels in the interval $[0, 4]$. In the sequel, x will refer to a sequence of maps of motion-related measurements, and Λ to the space of the quantized motion-related measurements.

3. STATISTICAL MOTION ACTIVITY MODELING

3.1. spatio-temporal random walks

To characterize motion activity within video sequences, we exploit random walks as investigated in [3] for the characterization of the spatial color distributions of still images. In this paper, we rely on spatio-temporal random walks within sequences of motion-related measurement maps. A random walk is specified by a spatio-temporal neighborhood. Two different examples of neighborhood systems with 14 and 26 interactions are given in Fig.1. At instant t, the random walk goes from the current position $(k_t, p_t) \in \{1, .., K\} \times \mathcal{R}$ to a new location $(k_{t+1}, p_{t+1}) \in \{1, \ldots, K\} \times \mathcal{R}$, where K is the length of the sequence of motion-related measurement maps and $\mathcal{R}$ the image grid. The new location is chosen with a uniform probability within the spatio-temporal neighborhood of point (k_t, p_t). We randomly select the initial location (k_0, p_0). If an image border is reached, the random walk goes on from a new randomly chosen spatio-temporal location in $\{1, .., K\} \times \mathcal{R}$. Thus, at iteration T, we have defined a sequence of successive positions $S = \{(k_0, p_0), .., (k_T, p_T)\}$. The associated motion-related measurement sequence is denoted by $y = \{y_i\}_{i \in \{0,..,T\}}$ with $y_i = x_{k_i}(p_i)$. y can be considered as the realization of a random process Y, and the likelihood function $P_{\mathcal{M}}(Y)$ of

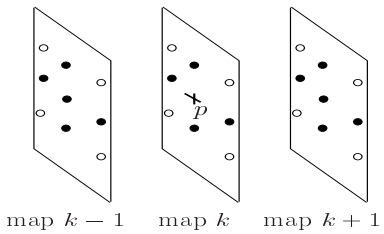

Fig. 1. *Two examples of spatio-temporal neighborhood for point p in frame k : 14-neighborhood (symbols $\bullet$) and 26-neighborhood (symbols $\bullet$ and $\circ$).*

Y is defined as :

$$P_{\mathcal{M}}(Y) = P_{\mathcal{M}}(Y_0) \prod_{i=0}^{i=T} P_{\mathcal{M}}(Y_i|Y_{i-1}) \qquad (2)$$

where $\mathcal{M}$ refers to the statistical model accounting for the distribution attached to the observed random process Y. The use of spatio-temporal random walks permits to evaluate spatio-temporal cooccurrences of motion-related measurements, which supplies a single statistical framework to handle both temporal and spatial properties of motion activity.

We assume that $P_{\mathcal{M}}(Y_0)$ follows a uniform law. Therefore, the knowledge of the transition matrix $\{P_{\mathcal{M}}(Y_i = \nu|Y_{i-1} = \nu')\}_{(\nu,\nu') \in \Lambda^2}$ entirely specifies the motion activity model $\mathcal{M}$. In order to supply an exponential formulation of $P_{\mathcal{M}}(Y)$, we introduce the Gibbsian potentials $\Psi_{\mathcal{M}} = \{\Psi_{\mathcal{M}}(\nu, \nu')\}_{(\nu,\nu') \in \Lambda^2}$ defined by :

$$P_{\mathcal{M}}(Y_i = \nu|Y_{i-1} = \nu') = \exp \Psi_{\mathcal{M}}(\nu, \nu') \qquad (3)$$

with $\sum_{\nu \in \Lambda} \exp \Psi_{\mathcal{M}}(\nu, \nu') = 1$. $P_{\mathcal{M}}(Y)$ can be rewritten according to an exponential expression:

$$P_{\mathcal{M}}(Y = y) = P_{\mathcal{M}}(Y_0 = y_0) \cdot \exp\left[\Psi_{\mathcal{M}} \bullet \Gamma(y)\right] \qquad (4)$$

where $\Psi_{\mathcal{M}} \bullet \Gamma(y)$ is a dot product between model potentials $\Psi_{\mathcal{M}}$ and cooccurrence measurements $\Gamma(y) = \{\Gamma(\nu, \nu'|y)\}_{(\nu,\nu') \in \Lambda^2}$, with:

$$\Gamma(\nu, \nu'|y) = \sum_{i=1}^{i=T} \delta(y_i - \nu) \cdot \delta(y_{i-1} - \nu') \qquad (5)$$

where δ is the Kronecker symbol. Then, the dot product $\Psi_{\mathcal{M}} \bullet \Gamma(y)$ can be simply expressed as:

$$\Psi_{\mathcal{M}} \bullet \Gamma(y) = \sum_{(\nu,\nu') \in \Lambda^2} \Psi_{\mathcal{M}}(\nu, \nu') \cdot \Gamma(\nu, \nu'|y) \qquad (6)$$

For a sequence y, the Maximum Likelihood (ML) model estimate is given by the empirical occurrences of the transitions observed in the sequence y. The potentials of the estimated ML model $\widehat{\mathcal{M}}$ w.r.t. y are then given by:

$$\Psi_{\widehat{\mathcal{M}}}(\nu, \nu') = \ln \left(\Gamma(\nu, \nu'|y) / \sum_{\nu'' \in \Lambda} \Gamma(\nu'', \nu'|y) \right) \qquad (7)$$

3.2. Statistical similarity measure of motion activity

Given two image sequences, we want to evaluate the degree of similarity between their respective motion activity levels. We note $\mathcal{M}^1$ and $\mathcal{M}^2$ two motion activity models issued from two different video sequences, and, Γ^1 and Γ^2 the associated cooccurrence distributions. $\mathcal{M}^1$ and $\mathcal{M}^2$ were estimated from two random walk realizations y^1 and y^2.

We introduce a similarity measure $D(\mathcal{M}^1, \mathcal{M}^2)$ relying on a symmetrical version of the Kullback-Leibler (KL) divergence which evaluates the distance between two probability distributions as the expectation of their log-ratio:

$$D(\mathcal{M}^1, \mathcal{M}^2) = \frac{1}{2}\left[KL(\mathcal{M}^1 \| \mathcal{M}^2) + KL(\mathcal{M}^2 \| \mathcal{M}^1)\right] \quad (8)$$

where $KL(\mathcal{M}^1 \| \mathcal{M}^2)$ denotes the KL divergence. Its approximation comes to compute a log-ratio of likelihoods evaluated on y^1 w.r.t. respectively model $\mathcal{M}^1$ and $\mathcal{M}^2$ [2]. Using the exponential formulation (4), we obtain:

$$KL(\mathcal{M}^2 \| \mathcal{M}^1) \approx \frac{1}{T}[\Psi_{\mathcal{M}^1} - \Psi_{\mathcal{M}^2}] \bullet \Gamma^1 \quad (9)$$

This expression quantifies the loss of information occurring when substituting $\mathcal{M}^2$ for $\mathcal{M}^1$ to account for the motion activity corresponding to Γ^1.

4. APPLICATION TO MOTION RECOGNITION

4.1. Experimental set of image sequences

We have carried out motion recognition experiments over an image sequence set[1] including eight motion classes. The video set contains four kinds of temporal texture: wind blown grass (A), gentle sea waves (B), rough water turbulence (C) and wind blown trees (D). We also introduce one class of static anchor shot from news program involving a weak motion activity (E) . In addition, two classes of rigid motion are included: sequences involving moving (descending or ascending) escalator (F) and traffic image sequences (G). The last class (H) is formed by sequences of a pedestrian walking either from the left to the right or from the right to the left. All these sequences have been acquired with a static camera. Moving camera can be handled as well [2].

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 the left to the right and five ones a pedestrian walking from right to the left). Fig.2, we display for each class in one image representative of each sequence (for class (H), we have selected three shots over the ten sequences belonging to this class).

[1]We thank INA, Département Innovation, Direction de la Recherche, for providing the news sequences, and, C.H. Peh and L.F. Cheong at National University of Singapore for providing temporal texture samples. The sequences of the video set can be viewed at http://www.irisa.fr/prive/rfablet/base_reco_mvt.english.html.

	A	B	C	D	E	F	G	H
A	**83.00** *74.00*		*9.00*	**11.00** *17.00*	**2.00**			
B		**91.0** *81.0*		**9.00** *19.00*				
C		**40.00** *40.00*	**60.00** *60.00*					
D			**2.00**	**96.0** *98.00*				**2.00** *2.00*
E	**2.50**	*5.00*			**95.0** *92.00*			**2.50** *3.00*
F		**2.00** *2.00*	*45.00*			**98.00** *53.00*		
G	**8.00** *10.00*		*14.00*		**12.00**		**80.00** *76.00*	*2.00*
H	**5.00** *2.50*				*2.50*			**95.00** *95.0*

Table 1. *Percentage of correct and false classification according to the different motion classes. For each class, we report results obtained using the RW method (bold type) and the TG method (italic type). For instance, the percentages of samples from class (A) assigned to class (A), (D) and (E) are resp.* 83%, 11.00% *and* 2.00% *using the RW method.*

4.2. Learning and classification stage

To cope with motion recognition, we first perform a supervised learning stage using a training set of image sequences, and, in a second step, we achieve motion recognition over a test set. These two sets are defined as follows. Each image sequence is divided into subsequences of six images to consider sequences of five maps of motion-related measurements. Thus, we obtain 57 samples in each class leading to a total set of 456 sequences. Then, the first ten samples of the first sequence of the seven first classes, (A) to (G), are used as the training data for the classifier. 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 $\mathcal{A}$ comprising 80 sequences, and a test set $\mathcal{B}$ including 376 sequences of five images.

For each element $a \in \mathcal{A}$, the learning stage consists in estimating and storing the associated statistical motion activity model $\mathcal{M}^a$ and cooccurrence measurements Γ^a. Each motion class is then described by a set of motion activity models. The classification stage resorts to a nearest-neighbor classification scheme. Given $b \in \mathcal{B}$, we determine the associated motion activity model $\mathcal{M}^b$ and cooccurrence distribution Γ^b. The recognition step comes to determine the closest motion class according to the statistical similarity measure D. We retrieve within the set of stored models $\{\mathcal{M}^a\}_{a\in\mathcal{A}}$, the model $\mathcal{M}^{\widehat{a}}$ the nearest of model $\mathcal{M}^b$, i.e. $\widehat{a} = \arg\min_{a\in\mathcal{A}} D(\mathcal{M}^b, \mathcal{M}^a)$ and b is finally stated as belonging to the motion class of $\widehat{a}$.

4.3. Classification experiments

In addition to experiments carried out using statistical models derived from spatio-temporal random walks of length $T = 3\times K\cdot|\mathcal{R}|$ with a 26-neighborhood (see Fig.1), we have considered the temporal Gibbsian modeling framework presen-

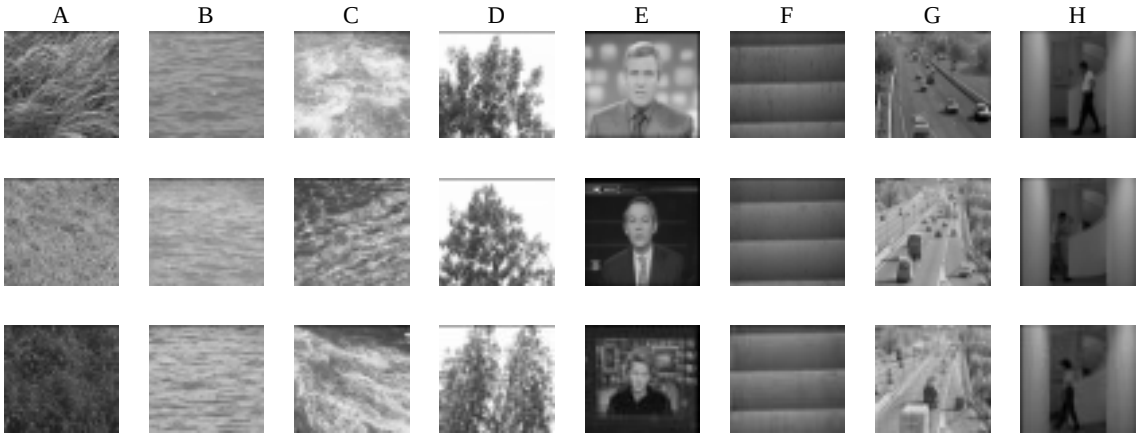

Fig. 2. *Experimental video set: for each of the eight motion classes (A to H), we display three images representative of each sequence of the motion class. The eight classes contain various dynamic contents: (A) wind blown grass, (B) gentle sea waves, (C) rough turbulent water, (D) wind blown trees, (E) anchor shot in news program (F) moving escalator, (G) car traffic and (H) pedestrian walking.*

ted in [2] for comparison purpose. In the sequel, we denote as the TG method the approach using Temporal Gibbsian models, and as the RW method the one based on spatio-temporal random walks.

As shown in Tab.1, the RW method succeeds in discriminating the eight motion classes defined in subsection 4.1 with a correct classification rate of 87.25% in average. The highest classification error corresponds to motion class (C). It is due to the misclassification of the elements issued from the second sequence of this class. Its motion activity is indeed at an intermediate level between motion activity of class (B) and the two other sequences of class (C). These results emphasize that the complete recovery of motion information by means of dense optic flow fields is not always necessary for issues such as motion recognition or classification. Furthermore, it suggests that the use of motion-related measurements which do not comprise direction information is sufficient to recover general motion types.

Tab.1 also shows that the RW method outperforms the TG method for almost all motion classes (the only exception is class (D) with a small difference, 96% vs. 98%). The mean rate of correct classification for the TG method is 78.62%, which is 8.6% less than the rate obtained with the RW method. These experiments highlight that the additional spatial characterization provided by the use of spatio-temporal random walks brings valuable and efficient supplementary information for recognition purpose. Besides, model complexity is identical for the two methods ($|\Lambda|^2$ model potentials in both cases). However, in terms of implementation, the RW method reveals more complex: it requires to process image sequences as a whole to generate spatio-temporal random walks, whereas temporal cooccurrences can be updated incrementally between successive maps of motion-related measurements.

5. CONCLUSION

We have presented a single non parametric statistical motion modeling framework which can capture both temporal and spatial aspects of motion activity. It relies on statistical models estimated from spatio-temporal random walks within sequences of maps of local motion-related motion measurements. It can be straightforwardly used for motion classification or recognition issues since it makes feasible and simple both ML model evaluation and the computation of the motion activity similarity measure. The motion recognition experiments carried out over a video set comprising various types of dynamic content (rigid motion, temporal texture, pedestrian walking) demonstrates the efficiency of our method for motion activity characterization.

6. REFERENCES

[1] R. Fablet and P. Bouthemy. Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval. In *Proc. of 3rd Int. Conf. on Visual Information Systems, VISUAL'99*, LNCS Vol 1614, pages 221–228, Amsterdam, June 1999. Springer.

[2] 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.

[3] D. De Menthon, L.J. Latecki, and A. Rosenfeld. Relevance ranking of video data using hidden Markov model distances and polygon simplification. In *Proc. of 4th Int. Conf. on Visual Information Systems, VISUAL'2000*, pages 49–61, Lyon, Nov. 2000.

[4] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP*, 56(1):78–99, 1992.

[5] 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.

[6] 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.