Statistical modeling for motion-based video classification and retrieval

Ronan Fablet, Patrick Bouthemy

STATISTICAL MODELING FOR MOTION-BASED VIDEO CLASSIFICATION AND RETRIEVAL

R. Fablet

IRISA/CNRS
Campus universitaire de Beaulieu
35042 Rennes Cedex, France
rfablet@irisa.fr

P. Bouthemy

IRISA/INRIA
Campus universitaire de Beaulieu
35042 Rennes Cedex, France
bouthemy@irisa.fr

ABSTRACT

We have developed an original approach for content-based video indexing and retrieval. By introducing a causal Gibbsian modeling of the spatio-temporal distribution of appropriate local motion-related measurements, we have designed a general and efficient statistical framework for non parametric motion modeling, motion recognition and classification, and motion segmentation. It is exploited for motion-based video indexing and video retrieval for both global and partial queries by example.

1. INTRODUCTION AND RELATED WORK

Video archives are at the core of various application fields. The efficient use of these databases requires to offer reliable and relevant access to visual information. In particular, it implies to index and retrieve visual documents by their content. A great deal of research amount is currently devoted to image and video database management, [1, 11, 13, 20]. Nevertheless, it remains hard to easily identify the relevant information for a given query, due to the complexity of image and sequence interpretation.

Content-based video editing, indexing, browsing, and retrieval primarily require to recover the elementary shots of the video and to recognize typical forms of video shooting such as static shot, traveling, zooming and panning [1, 20][1]. At a second stage, it appears necessary to provide an interpretation and a representation of the shot content. Dynamic content analysis is of particular interest and its combination with static content descriptions should offer new functionalities for video navigation, browsing or retrieval. A first class of approaches, which rely on parametric or dense motion field estimation, includes image mosaicing [13], segmentation, tracking and characterization of moving elements in order to determine a spatio-temporal representation of the video shot [10, 11]. However, these techniques turn out to be unadapted to certain classes of sequences with complex dynamic contents such as motion of rivers, flames, foliages in the wind, or crowds, *etc*. Therefore, in the context of video indexing, it seems also relevant to adopt a global point of view that avoids any explicit motion segmentation step.

We have developed an original approach for motion-based video indexing and retrieval. It aims at interpreting dynamic contents without any prior motion segmentation and without any complete motion estimation in terms of parametric models or optical flow fields. Preliminary researches in that direction have defined extraction techniques of "temporal texture" features, [16, 19]. Motions of rivers, foliages, flames, or crowds, for instance, can indeed be regarded as temporal textures. In [16], features issued from spatial cooccurrences of normal flows are exploited to classify sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In previous work, we have found global features extracted from temporal cooccurrence distributions of local motion-related measurements more appropriate than the use of normal velocity fields [2].

This paper is organized as follows. Section 2 outlines the general ideas underlying our work. Section 3 describes the local non-parametric motion-related information that we consider. In Section 4, we introduce the statistical modeling of the temporal distribution of motion-related quantities computed from a video sequence and the associated estimation scheme. Section 6 presents several applications of our statistical non-parametric motion modeling framework to motion-based video indexing and recognition.

2. PROBLEM STATEMENT

To cope with video databases involving various dynamic contents, it is necessary to determine an optimal set of features and the associated similarity measure [18]. These issues can be tackled using Principal Component Analysis [15] or some feature selection techniques [14]. Nevertheless, feature space is usually of high dimension, and the

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

considered distance is likely not to capture properly uncertainty attached to feature measurements. Therefore, statistical methods appear more appropriate in that context than numerical feature values. Our aim is to supply within the same framework a global description of the dynamic content as well as efficient probabilistic tools for video database classification and retrieval with query by example.

To this end, we adopt a motion-classification approach for video indexing which relies on a statistical analysis of the spatio-temporal distribution of local non-parametric motion-related measurements. We aim at identifying probabilistic models corresponding to different dynamic content types to be discriminated. In [3, 12], a correspondence between cooccurrence distributions and Markov Random Field (MRF) models are established in the context of spatial texture analysis. We propose an extension to temporal textures. We have introduced causal models allowing us to compute the exact expression of the corresponding likelihood functions. We have thus developed a general statistical framework exploitable for video indexing and retrieval.

3. LOCAL MOTION-RELATED MEASUREMENTS

The first step is to define relevant local motion-related measurements whose spatio-temporal distributions will be modeled within a causal Gibbsian framework described in Section 4. Since our goal is to characterize the actual dynamic content of the scene, we first cancel camera motion. As a consequence, we estimate the dominant image motion between two successive images, which is assumed to be due to camera motion. Then, to cancel it, we wrap the successive images to the first image of the shot by combining the elementary dominant motions successively estimated over the consecutive image pairs.

3.1. Dominant motion estimation

To model the transformation between two successive images, we consider a 2D affine motion model. The estimation of the six affine motion parameters is achieved with the robust gradient-based incremental estimation method described in [17]. The use of a robust estimator allows the motion estimation not to be sensitive to secondary motions due to mobile objects in the scene. The minimization is performed by means of an iterative reweighted least-square technique embedded in a multiresolution framework.

3.2. Local motion-related quantity

In order to characterize the residual motion information in the motion-compensated image sequence, we consider the following local motion-related quantity :

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

where $I^*(p)$ is the intensity function at point p in the wrapped image, $\mathcal{F}(p)$ is a small window centered on p, η^2 a predetermined constant related to the noise level in uniform areas, and I_t^* is the temporal derivative of the intensity function I^*. $I_t^*(p)$ is approximated by a simple finite difference. Whereas the normal flow measure $\frac{I_t^*(p)}{\|\nabla I^*(p)\|}$ turns out to be very sensitive to noise attached to the computation of spatio-temporal derivatives of the intensity function, the considered motion-related measurement forms a more reliable quantity, yet simply computed from the intensity function and its derivatives. Obviously, the information relative to motion direction has been lost, which prevents from discriminating for instance two opposed translations with the same magnitude. However, this is not a real shortcoming, since we are interested in interpreting the type of dynamic situations observed in the considered video shot and not in identifying a specific motion value.

4. STATISTICAL NON-PARAMETRIC MODELING OF MOTION CONTENT

4.1. Temporal Gibbsian modeling

In order to characterize the motion activity, we exploit the probabilistic framework presented in [9] which relies on non-parametric motion models. We briefly outline it hereafter.

Let $x = \{x_0, ..., x_K\}$ be a sequence of $K+1$ maps of quantized motion-related quantities and $\mathcal{R}$ the spatial support of these maps. $\mathcal{R}$ can be the entire image, a given region, or a block, depending on the global or local nature of the targeted characterization. We first evaluate the associated distribution of temporal cooccurrences $\Gamma(x)$ defined by: $\forall(\nu, \nu') \in \Lambda^2$

$$\Gamma(\nu, \nu'|x) = \sum_{k=1}^{K} \sum_{r \in \mathcal{R}} \delta(\nu - x_k(r)) \cdot \delta(\nu' - x_{k-1}(r)) \quad (2)$$

where Λ is the set of possible discrete values of the quantized motion-related measurements.

We assume that sequence x is the realization of a random process X such that the conditional likelihood $P_{\mathcal{M}}(x)$ of sequence x w.r.t. to a model $\mathcal{M}$ is expressed as:

$$P_{\mathcal{M}}(x) = \frac{1}{Z} \exp\Big[\Psi_{\mathcal{M}} \bullet \Gamma(x)\Big] \quad (3)$$

where Z is a known normalization constant independent of $\mathcal{M}$ and x. $\Psi_{\mathcal{M}}$ is the set of potentials $\{\Psi_{\mathcal{M}}(\nu, \nu')\}_{(\nu,\nu') \in \Lambda^2}$

which explicitly specifies model $\mathcal{M}$. $\Psi_{\mathcal{M}} \bullet \Gamma(x)$ denotes the dot product between model potentials $\Psi_{\mathcal{M}}$ and cooccurrence distribution $\Gamma(x)$ given by:

$$\Psi_{\mathcal{M}} \bullet \Gamma(x) = \sum_{(\nu,\nu') \in \Lambda^2} \Psi_{\mathcal{M}}(\nu,\nu') \cdot \Gamma(\nu,\nu'|x) \quad (4)$$

The availability of such 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 cooccurrence distribution $\Gamma(x)$. 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 (3).

This modeling approach is non-parametric in two ways. First, it does not correspond to 2D parametric (affine or quadratic) motion models [17]. Second, from a statistical point of view, it does not involve parametric distributions (i.e., Gaussian) to model the law $P_{\mathcal{M}}(\nu|\nu')$. In some way, it is learnt from data.

Besides, the ML (Maximum Likelihood) estimation of model $\mathcal{M}$ best fitting the motion distribution attached to a given sequence x reveals straightforward. The potentials of the model $\widehat{\mathcal{M}}$, which verifies $\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} P_{\mathcal{M}}(x)$, are readily given by [9]:

$$\Psi_{\widehat{\mathcal{M}}}(\nu,\nu') = \ln\left(\Gamma(\nu,\nu'|x) / \sum_{\vartheta \in \Lambda} \Gamma(\vartheta,\nu'|x)\right) \quad (5)$$

4.2. Spatio-temporal Gibbsian modeling

The statistical modeling of motion information described above does not explicitly characterize spatial aspects (spatial patterns) of motion information, since it only relies on the evaluation of temporal cooccurrences of motion-related measurements. In order to combine within a single statistical framework the characterization of both spatial and temporal aspects of motion information, we have investigated three different alternatives:

- First, we have used causal spatio-temporal neighborhood instead of the use of a single temporal neighborhood [8]. The main drawback of this approach is that the computation of the likelihood function $P_{\mathcal{M}}$ and ML model estimation becomes more complicate;
- Second, we have exploited spatio-temporal random walks through the sequence of maps of motion-related measurements [5]. Similarly to temporal Gibbs model, the likelihood function $P_{\mathcal{M}}$ has an exponential form involving the dot product between model potentials and spatio-temporal cooccurrence measurements. Furthermore, ML model estimation is also readily determined from cooccurrence distributions;
- Third, we have investigated the characterization of spatial aspects of motion distribution through a multiscale strategy [6]. In fact, instead of using a single motion-related measurement at each location in the sequence, we consider a vector of motion-related quantities computed at different scales. The statistical modeling of motion information is then based on the computation of temporal and scale cooccurrences, while sharing characteristics similar to temporal Gibbs models, both in terms of computation of the likelihood function $P_{\mathcal{M}}$(x) and in terms of ML model estimation.

5. SIMILARITY MEASURE OF MOTION CONTENT

Designing a similarity measure between dynamic contents is of key interest for motion-based video indexing and retrieval. We can exploit the statistical motion modeling introduced in the previous section to define such a similarity measure.

Given two video shots n_1 and n_2, the associated sequences of maps of motion-related measurements, x^{n_1} and x^{n_2}, and statistical motion models, $\mathcal{M}^1$ and $\mathcal{M}^2$, the motion-based similarity measure $D_{KL}(n_1,n_2)$ is defined as:

$$D_{KL}(n_1,n_2) = \frac{1}{2}\Big[KL(\mathcal{M}^2\|\mathcal{M}^1) + KL(\mathcal{M}^1\|\mathcal{M}^2)\Big] \quad (6)$$

where $KL(\mathcal{M}^2\|\mathcal{M}^1)$ is the Kullback-Leibler divergence. When using temporal Gibbs model, $KL(\mathcal{M}^2\|\mathcal{M}^1)$ can be approximated by [9]:

$$KL(\mathcal{M}^2\|\mathcal{M}^1) = \Big[\Psi_{\mathcal{M}^1} - \Psi_{\mathcal{M}^2}\Big] \bullet \Gamma(x^{n_1}) \quad (7)$$

6. APPLICATION TO MOTION-BASED VIDEO INDEXING AND RETRIEVAL

We have exploited these statistical motion models for various applications related to motion recognition and classification.

In [5, 6], we have tackled the motion recognition issue for a video base containing different classes of motion content: temporal texture (rivers, grass motion, trees in the wind,...), sequences of pedestrians, traffic video shots. Based

on the computation of the motion similarity measure given by relation (6), we have obtained promising results with a mean recognition rate higher than 90%.

We have addressed the classification of video databases in [9]. Still using relation (6), we can build a binary tree which expresses in a hierarchical way video similarities in terms of motion content. This kind of hierarchical structure can be useful for retrieval or browsing in video databases.

We have also dealt with the extraction of entities of interest in images w.r.t. motion information in [7]. Based on the general characterization of motion information in terms of motion activity, statistical motion models are used to merge blocks of an initial partition of the image w.r.t. motion content similarity. Our approach enables to handle a wide range of dynamic entities of interest: single moving entity or group of entities in sport video shots, entities formed by a moving texture,... In addition, simultaneously to the extraction, we can provide a characterization of the movement of the entity of interest by means of the estimated statistical motion model.

Finally, the designed statistical motion models can also be used for motion-based retrieval with query by example. The retrieval process is in fact viewed as a Bayesian inference issue based on the MAP criterion. It first comes to build a database of statistical motion models. Then, the reply to a given query is determined by finding in the database the model best fitting the query. It involves the use of relation (3). We can handle both global query involving a whole video sequence [4, 9] as well as partial query (the query is concerned by a given area in the image) [4, 7]. In the later case, from a given database of video shots, we first build the database of statistical motion models attached to entities of interest extracted from the keyframe of each video shot.

7. REFERENCES

[1] P. Aigrain, H-J. Zhang, and D. Petkovic. Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia Tools and Applications*, 3(3):179–202, 1996.

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

[3] I.M. Elfadel and R.W. Pickard. Gibbs random fields, cooccurrences and texture modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(1):24–37, 1994.

[4] R. Fablet. Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images ; application à l'indexation vidéo. *Phd Thesis, University of Rennes 1, Irisa, No. 2526*, July 2001.

[5] R. Fablet and P. Bouthemy. Motion recognition using spatio-temporal random walks in sequence of 2D motion-related measurements. In *Proc. of 8th IEEE Int. Conf. on Image Processing, ICIP'2001*, Thessaloniki, Oct. 2001.

[6] R. Fablet and P. Bouthemy. Non parametric motion recognition using temporal multiscale Gibbs models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'2001*, Kauai, Dec. 2001.

[7] R. Fablet and P. Bouthemy. Non-parametric motion activity analysis for statistical retrieval with partial query. *Journal of Mathematical Imaging and Vision*, 14(3):257-270, May 2001.

[8] R. Fablet, P. Bouthemy, and P. Pérez. Non parametric statistical analysis of scene activity for motion-based video indexing and retrieval. Technical Report 4005, INRIA, 2000.

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

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

[11] M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. of 5th Eur. Conf. on Computer Vision, ECCV'98*, Freiburg, June 1998, LNCS Vol.1406, pages 595–609, Springer.

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

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

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

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

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

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

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

[19] 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, septembre 1996.

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