Motion Characterization from Temporal Cooccurrences of Local Motion-based Measures for Video Indexing

Patrick Bouthemy and Ronan Fablet
IRISA / INRIA
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France
E-mail : bouthemy@irisa.fr

14th Int. Conf. on Pattern Recognition, ICPR'98, Brisbane, August 1998

Abstract

This paper describes an original approach for motion interpretation with a view to content-based video indexing. We exploit a statistical analysis of the temporal distribution of appropriate local motion-based measures to perform a global motion characterization. We consider motion features extracted from temporal cooccurrence matrices, and related to properties of homogeneity, acceleration or complexity. Results on various real video sequences are reported and provide a first validation of the approach.

1. Introduction

Multimedia databases are of growing importance in various application fields, such as television archives (movies, documentaries, news, ...), multimedia publishing, road traffic surveillance, medical imaging, remote sensing and meteorology (satellite images), ... Then, efficient use of these databases requires to identify pertinent information related to their content to satisfy a given query or to access particular pieces of information. There is obviously a real need of indexing and retrieving multimedia documents by their content in an automatic way (at least partly). Several pioneering systems already exist for still images, [1, 5], and a large research effort is currently undertaken to handle image and video databases, [1, 7, 9, 13, 16]. Nevertheless, due to the complexity of image interpretation and dynamic scene analysis, several important issues remain to be further investigated.

As far as video sequences are concerned, content-based video indexing, browsing, editing, or retrieval, have motivated specific investigations focusing first on the structuration of the video in elementary shots, [1, 3, 7, 16], and concentrating more recently on image mosaicing [10], on image layering [2], on object motion characterization in case of a static camera [4], or on segmentation and tracking of moving elements [6]. Motion segmentation methods usually rely on 2D parametric motion models, and aim at localizing the different types of motions present in a scene. However, they turn out to be unadapted to certain classes of sequences, particularly in the case of unstructured motions of rivers, flames, foliages in the wind, or crowds, ..., (see Fig.1). Besides, it seems pertinent to provide a direct global characterization without any prior motion segmentation or without any complete motion estimation in terms of parametric models or optical flow fields. These remarks emphasize the need for the design of new low-level approaches in order to supply a direct global motion description, [11, 14, 15].

We follow this point of view and we propose an original method for video indexing with respect to motion content. It uses local non-parametric motion-related information, and extracts, from temporal cooccurrence statistics, global motion features relative to motion complexity, coherence or acceleration. In Section 2, we outline the analogy between our approach and texture analysis. Section 3 describes the local motion-related information used. In Section 4, we introduce temporal cooccurrence matrices and the extracted global motion features. Section 5 contains results obtained on various real video sequences and concluding remarks.

2. Problem statement

Our approach consists in analyzing, within each shot previously extracted from the processed video sequence, the whole spatio-temporal motion distribution, as spatial grey level distribution in texture analysis. In particular, we aim at adapting in that context cooccurrence measurements which supply a texture characterization in terms of homogeneity, contrast or coarseness [8].

Preliminary work in that direction, developed by

Polana and Nelson, [11, 14], introduces the notion of temporal texture, related to fluid motions. Indeed, motions of rivers, foliages, flames, or crowds, ..., can be regarded as temporal textures.

a)  b) 

Figure 1. *Examples of temporal textures : a) foliage b) fire (by courtesy of* MIT*).*

Maps of local motion-related measures along the image sequence, required as input of cooccurrence measurements, could be provided by dense optical flow fields. However, first, it is really time consuming, and second, the quality of the estimated displacement fields cannot be ensured in the case of temporal content corresponding to such complex dynamic scenes. Therefore, we prefer to consider local motion-related information, easily computed from the spatio-temporal gradients of the intensity. Contrary to [11], where the normal velocity is considered, we make use of a more reliable information as explained in the next section.

3. Local motion-related measures

By assuming intensity constancy along 2D motion trajectories, the image motion constraint relating the 2D apparent motion and the spatio-temporal derivatives of the intensity function can be expressed by :

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

where $\mathbf{v}$ is the 2D motion vector in the image, $I(p)$ the intensity function at point p, $\nabla I = (I_x, I_y)$ the intensity spatial gradient, and $I_t(p)$ the intensity partial temporal derivative. Then, we can infer the normal velocity v_n :

$$v_n(p) = \frac{-I_t}{\|\nabla I(p)\|} \qquad (2)$$

This quantity v_n can in fact be null whatever the motion magnitude, if the motion direction is perpendicular to the spatial intensity gradient. v_n is also very sensitive to noise attached to the computation of the intensity derivatives. However, if the spatial intensity gradient is sufficiently distributed in terms of direction in the vicinity of point p, an appropriately weighted average of v_n in a given neighbourhood forms a more relevant motion-related quantity :

$$v_{obs}(p) = \frac{\sum_{s \in \mathcal{F}(p)} \|\nabla I(s)\|^2 \cdot |v_n(s)|}{\max(\eta^2, \sum_{s \in \mathcal{F}(p)} \|\nabla I(s)\|^2)} \qquad (3)$$

where $\mathcal{F}(p)$ is a 3×3 window centered on p. η^2 is a predetermined constant, related to the noise level in uniform areas, which prevents from dividing by zero or by a very low value. In [12], this motion-related measure v_{obs} has been successfully used to process sequences compensated by the estimated dominant motion with a view to detecting moving objects in a scene, and confidence bounds, depending on the intensity gradient distribution within $\mathcal{F}(p)$, have been derived to assess its reliability.

Thus, v_{obs} provides us with a local motion measure, easily computed and reliably exploitable. The loss of the information relative to motion direction is not a real shortcoming, since we are interested in interpreting the general type of dynamic situations observed in a given video shot.

4. Extraction of global motion features

4.1. Quantifying the motion quantities

The computation of cooccurrence matrices requires a quantization of the continuous variables v_{obs}. The simplest way would consist in applying a linear quantization within the interval $[\inf_p v_{obs}(p); \sup_p v_{obs}(p)]$, which would keep the whole structure of the distribution.

a) 

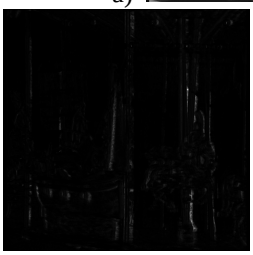
b)  c) 

Figure 2. *Motion-related information for the video shot Merry-go-round : (a) first image of the sequence, (b)-(c) maps of motion quantities v_{obs} with a linear quantization (b) and with the introduction of a-priori bounds (c)*

Nevertheless, in practice, it turns out to be irrelevant due to the spreading out of motion quantities, as shown in Figure 2. The sequence *Merry-go-round* is a static camera shot with, in the foreground, a merry-go-round which undergoes a rotation, and, in the background, walking persons and a bus just leaving the square.

Therefore, we introduce bounds which define an interval where measures are regarded as pertinent. Since the motion quantities are positive, 0 is taken as the lower bound. Moreover, in motion estimation, it is generally considered that a single resolution analysis is un-

able to correctly estimate displacements of large magnitude. It appears relevant to introduce a limit beyond which measures are no more regarded as usable. In the experiments,practice, sampling within $[0, 4]$ on 16 levels proves accurate, as shown in Figure 2.

4.2. Motion cooccurrences

Polana and Nelson have combined spatial cooccurrence distribution with normal flow fields to classify processed examples in pure motions (rotational, divergent) or in temporal textures (river, foliage), [11]. However, temporal evolution cannot be handled in that way , since studied interactions are purely spatial, and only stationary motions can be characterized. Moreover, considering spatial cooccurrences is highly time-consuming, since matrices relative to several configurations of spatial interactions have to be computed. That is the reason why we have looked for transferring cooccurrence to the temporal domain.

The temporal cooccurrence for the pair of quantified motion quantities (i, j) at the temporal distance d_t is defined as follows :

$$P_{d_t}(i,j) = \frac{\sharp\{(r,s) \in C_{d_t} / obs(r) = i, obs(s) = j\}}{|C_{d_t}|}$$

where obs holds for the quantified version of v_{obs}, and $C_{d_t} = \{(r, s)$ at the same spatial position in the image grid $/\exists t, r \in \text{image}(t)$ and $s \in \text{image}(t - d_t)\}$. These temporal cooccurrences are evaluated over all the images of the given shot of the video sequence (typically, about 20 images for the examples reported below).

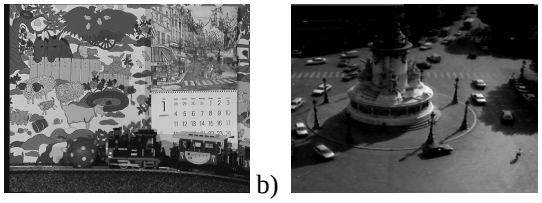
4.3. Global motion features

From cooccurrence matrices, we can now extract global motion features similar as those defined in [8] :

$$\begin{cases} \text{Average} : A = \sum_{(i,j)} i P_{d_t}(i,j) \\ \text{Variance} : \sigma^2 = \sum_{(i,j)} (i - A)^2 P_{d_t}(i,j) \\ \text{Dirac} : \delta = A^2/\sigma^2 \\ \text{Angular Second Moment} : ASM = \sum_{(i,j)} P_{d_t}(i,j)^2 \\ \text{Contrast} : Cont = \sum_{(i,j)} (i-j)^2 P_{d_t}(i,j) \end{cases}$$

The average feature indicates the importance of the observed motion, whereas the variance and the Dirac features express the degree of spreading out of the motion distribution. The contrast feature is related to the average acceleration. The ASM feature quantifies the temporal coherence. It varies within $\left[\frac{1}{(N+1)^3}, 1\right]$; it is equal to 1, if motion is close to uniformity in space and time. On the contrary, if the motion coherence falls, it tends to $\frac{1}{(N+1)^3}$.

This set of global motion features is computed over all the image grid. This could also be achieved either on predefined blocks or on extracted regions resulting from a spatial segmentation. This ensures feasible cooperations with other methods for video content analysis.



a) b)

Figure 3. *Video shots: a) Mobi, b) Concorde.*

5. Results and concluding remarks

This approach has been validated by experiments with several kinds of real video sequences. We report here results obtained on four video sequences representative of different classes of dynamic situations. The first sequence, called *Fire* (Fig.1), is temporal texture with a poorly structured motion in space and time. The second processed example is the sequence *Merry-go-round*, shown in Figure 2, which involves quite an important motion activity. In the third sequence *Mobi* (Fig. 3a), several rigid motions are combined ; along with the camera panning, the train and the calendar undergo a translation respectively from right to left, and towards the top of the scene, whereas the ball rolls towards the left. The fourth sequence is a static shot of the Concorde Place in Paris (Fig. 3b), with a weak motion activity resulting from the presence of cars around the Obelisk.

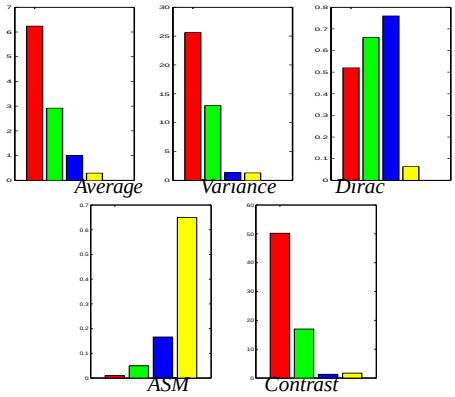


Figure 4. *Global motion features : from left to right and from top to bottom, A, σ^2, δ, ASM, and Cont, computed with $d_t = 1$, for four sequences, Fire, Merry-go-round, Mobi and Concorde (from left to right within each sub-figure).*

The results reported in Figure 4 show that the computed motion features can discriminate between the different motion configurations and quantify their degree of homogeneity, spatial and temporal uniformity, as well

a)  b)

d_t	A	σ^2	δ	ASM	$Cont$
1	1.22	1.07	1.4	0.14	1.4
4	1.22	1.05	1.4	0.12	1.6
8	1.22	1.07	1.4	0.12	1.7

Figure 5. *Influence of the temporal distance on the global motion features. (a) One image of the shot* zoom*, (b) Table of values computed for the shot* zoom

as complexity. The characterization of secondary motions in presence of camera motion seems also possible. In the case of sequence *Mobi*, the camera is panning the scene. Nevertheless, our global motion characterization method accesses directly to the content of the filmed scene. Moreover, other experiments prove that this analysis is quite independent of subsampling in space or time.

Results given in Figure 5 corresponds to another example involving a camera zooming on a curtain. They show that the four features relative to coherence and homogeneity, i.e., average, variance, Dirac, and ASM features, are almost independent of the chosen temporal distance d_t. Consequently, we can compute these four motion features for only one cooccurrence parameter value, which greatly saves calculation time. Concerning the contrast feature, its evolution with respect to d_t yields a characterization of the motion acceleration. Indeed, in the video shot *zoom*, the acceleration is positive and the contrast feature increases with d_t.

We have described in this paper an original and efficient method of global motion characterization for content-based video indexing. It relies on a second-order statistical analysis of temporal distributions of relevant, local, and quantified motion-related information. It exploits global motion features extracted from temporal cooccurrence matrices. This approach allows us to deal with various kinds of motion configurations, and to describe properties of the image dynamic content as complexity, uniformity and homogeneity. It does not require parametric motion models nor the computation of optical flow fields. Obtained results on a reasonable range of real scenes demonstrate that these global motion features can provide us with a set of pertinent and discriminating indexes with a view to video indexing by the dynamic content. In future work, a first step will be to determine optimal sets of global features corresponding to specific video databases, and to design a complete classification scheme. Moreover, we hope to benefit from the flexibility of our method by introducing a multi-scale approach in order to refine the analysis of the motion structure.

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, Nov. 1996.

[2] S. Ayer and H. Sawhney. Compact representations of videos through dominant and multiple motion estimation. *IEEE Trans. on PAMI*, 18(8):814-830, Aug. 1996.

[3] P. Bouthemy and F. Ganansia. Video partioning and camera motion characterization for content-based video indexing. In *Proc. ICIP'96*, Lausanne, Sept. 1996.

[4] J. D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recog.*, 30(4):607-625, Apr. 1997.

[5] M. Flicker *et al.* Query by image and video content : the QBIC system. *IEEE Computer*, pp.23-32, Sept. 1995.

[6] M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. 5th European Conf. on Computer Vision, ECCV'98*, Freiburg, June 1998.

[7] A. Gupta, A. Hampaper, M. Gorkani, and R. Jain. On summarization of video. In *Proc. IEEE 4th Int Conf. on Image Processing, ICIP'97*, Santa-Barbara, Oct. 1997.

[8] R. M. Haralick, K. Shanmugan, and I. Dinstein. Textural features for image classification. *IEEE Trans. on Systems, Man and Cybernetics*, 3(6):610-621, Nov. 1973.

[9] F. Idris and S. Pandranathan. Review of image indexing techniques. *Jal of Visual Communication and Image Representation*, 8(2):146-166, June 1997.

[10] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Efficient representations of video sequences and their applications. *Signal Processing : Image Communication*, 8:327-351, 1996.

[11] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP : Image Understanding*, 56(1):78-89, July 1992.

[12] J. M. Odobez and P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In H. H. Li, S. Sun, and H. Derin, eds, *Video Data Compression for Multimedia Computing*, ch. 8, pp 283-311. Kluwer Academic Publ., 1997.

[13] R. W. Pickard and T. O. Minka. Vision texture for annotation. *Multimedia Systems*, 3(3):3-14, Feb. 1995.

[14] R. Polana and R. Nelson. Detecting activities. In *Proc. Conf. Computer Vision and Pattern Recog., CVPR'93*, New-York, 1993.

[15] M. Szummer and R. Picard. Temporal texture modeling. In *Proc. ICIP'96*, Lausanne, Sept. 1996.

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