Motion-Based Feature Extraction and Ascendant Hierarchical Classification for Video Indexing and Retrieval

Ronan Fablet[1] and Patrick Bouthemy[2]

[1]IRISA / CNRS
[2]IRISA / INRIA
Campus universitaire de Beaulieu,
35042 Rennes Cedex, France
e-mail : rfablet@irisa.fr, bouthemy@irisa.fr
Tel : (33) 2.99.84.25.23 Fax : (33) 2.99.84.71.71

3rd Int. Conf. on Visual Information Systems, VISual'99, Amsterdam, June 1999

Abstract This paper describes an original approach for motion characterization with a view to content-based video indexing and retrieval. A statistical analysis of temporal cooccurrence distributions of relevant local motion-based measures is exploited to compute global motion descriptors, which allows to handle diverse motion situations. These features are used in an ascendant hierarchical classification procedure to supply a meaningful hierarchy from a set of sequences. Results of classification and retrieval on a database of video sequences are reported.

1 Introduction

Image databases are at the core of various application fields, either concerned with professional use (remote sensing and meteorology from satellite images, road traffic surveillance from video sequences, medical imaging, ...) or targeted at a more general public (television archives including movies, documentaries, news, ...; multimedia publishing,...). Reliable and convenient access to visual information is of major interest for an efficient use of these databases. Thus, it exists a real need for indexing and retrieving visual documents by their content. A large research amount is currently devoted to image and video database management, [1, 7, 8, 17]. Nevertheless, due to the complexity of image interpretation and dynamic scene analysis, it remains hard to easily identify relevant information with regards to a given query.

As far as image sequences are concerned, content-based video indexing, browsing, editing, or 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, 3, 8, 16, 17]. These issues also motivate studies concentrating on image mosaicing [10], on object motion characterization in case of a static camera [4], or on segmentation and tracking of moving elements, [6]. These methods generally exploit motion segmentation relying either on 2D parametric motion models or on dense optical flow field estimation. They aim at determining a partition of a given scene into regions attached to different types of motions with a view to extracting relevant moving objects. Nevertheless, 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 Figure 1). Moreover, providing a global interpretation of motion along a sequence, without any prior motion segmentation or without any complete motion estimation in terms of parametric models or optical flow fields, seems in the context of video indexing attractive and achievable to discriminate general types of motion situations. These remarks emphasize the need of designing new low-level approaches in order to supply a direct global motion description, [2, 12, 14, 15].

We propose an original approach to video indexing and retrieval according to the motion content. It relies on the global motion-based features presented in our previous work [2]. They are extracted using a statistical analysis of temporal cooccurrences of local non-parametric motion-related information. These motion indexes are introduced in a flexible ascendant hierarchical classification scheme to determine a meaningful hierarchy from a large video sequence set. It expresses similarities based on some metrics in the feature space. We can easily exploit the computed hierarchy for efficient retrieval with query by example.

This paper is organized as follows. In Section 2, we outline the general ideas leading to our work. Section 3 briefly describes the motion-based feature extraction. In Section 4, we introduce the indexing structure and the retrieval procedure. Section 5 contains classification results and retrieval examples, obtained on a large set of video sequences, and Section 6 contains concluding remarks.

2 Problem statement and related work

Video sequences are first processed to extract elementary shots with the technique presented in [3] (note that in the following we may use the term of sequence to deal with an elementary shot). Then, for each previously extracted shot, we intend to characterize the whole spatio-temporal motion distribution in order to build a motion-based indexing and retrieval system.

Let us note that, in the same manner, texture analysis methods study the spatial grey-level distribution. In particular, cooccurrence measurements provide efficient tools for texture description in terms of homogeneity, contrast or coarseness, [9]. Therefore, we aim at adapting cooccurrence-based features in the context of motion analysis. Preliminary research in that direction was developed by Polana and Nelson for activity recognition [12]. As part of their work, they introduce the notion of temporal texture, opposed to periodic activities or rigid motions, and associated to fluid motions. Indeed, motions of rivers, foliages, flames, or crowds, ..., can be regarded as temporal textures (see Figure 1).

In [15], temporal texture synthesis examples close to the original sequences are reported. However, this work is devoted to these particular cases of dynamic scenes, and cannot be extended to rigid motions or periodic activities. In [14], temporal texture features are extracted based on the description of spatio-

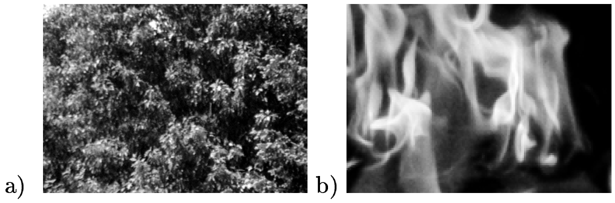


Figure1. *Examples of temporal textures : a) foliage b) fire (by courtesy of* MIT*).*

temporal trajectories. However, it relies on detection of moving contours by a simple thresholding of the pixel-based frame differences, which are known to be noisy.

In the subsequent, maps of local motion measures along the image sequence are required as input of cooccurrence measurements. As dense optical flow field estimation is time-consuming and unreliable in case of complex dynamic scenes, we prefer to consider local motion-related information, easily computed from the spatio-temporal derivatives of the intensity. Rather than the normal velocity used in [12], a more reliable information is exploited as explained in the next section. Besides, we intend to design a new video indexing and retrieval approach using the global motion-based features extracted from the temporal cooccurrences statistics. Thus, we first need to determine a meaningful indexing structure on a large dataset. Among all the clustering methods, we focus on ascendant hierarchical classification (AHC), [5, 11]. It exploits a Euclidean norm on the motion-based feature space and aims at minimizing the within-class variances. The obtained hierarchical representation is directly exploited for efficient retrieval with query by example.

3 Extraction of global motion-based features

3.1 Local motion-related measures

By assuming intensity constancy along 2D motion trajectories, the well-known image motion constraint relates the 2D apparent motion and the spatio-temporal derivatives of the intensity function, and the normal velocity v_n at a point p is given by : $v_n(p) = \frac{-I_t(p)}{\|\nabla I(p)\|}$ where $I(p)$ is the intensity function, $\nabla I = (I_x, I_y)$ the intensity spatial gradient, and $I_t(p)$ the intensity partial temporal derivative.

If the motion direction is orthogonal to the spatial intensity gradient, this quantity v_n can in fact be null whatever the motion magnitude. v_n is also very sensitive to noise attached to the computation of the intensity derivatives. Nevertheless, an appropriately weighted average of v_n in a given neighborhood forms a more relevant motion-related quantity as shown in [13] :

$$v_{obs}(p) = \frac{\sum_{s \in \mathcal{F}(p)} \|\nabla I(s)\|^2 \cdot |v_n(s)|}{\max(\eta^2, \sum_{s \in \mathcal{F}(p)} \|\nabla I(s)\|^2)} \tag{1}$$

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. 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.

The computation of cooccurrence matrices can not be achieved on a set of continuous variables. Due to the spreading out of the measures v_{obs}, a simple linear quantization within the interval $[\inf_p v_{obs}(p); \sup_p v_{obs}(p)]$ is not pertinent. Since it is generally assessed in motion analysis that large displacements can not be handled through a single resolution analysis, we set a limit beyond which measures are no more regarded as reliable. Thus, in practice, we quantize linearly the motion quantities within $[0, 4]$ on 16 levels.

3.2 Global motion features

In [12], spatial cooccurrence distributions are evaluated on normal flow fields to classify processed examples in pure motion (rotational, divergent) or in temporal texture (river, foliage). In that case, since studied interactions are spatial, only motions which are stationary along the time axis can be characterized. Moreover, to recover the spatial structure of motion, several configurations corresponding to different spatial interactions have to be computed, which is highly time-consuming. Consequently, we focus on temporal cooccurrences defined for a pair of quantized motion quantities (i, j) at the temporal distance d_t by :

$$P_{d_t}(i,j) = \frac{\sharp\{(r,s) \in C_{d_t} / obs(r) = i, obs(s) = j\}}{|C_{d_t}|} \tag{2}$$

where *obs* holds for the quantized version of v_{obs}, and $C_{d_t} = \big\{(r, s)$ at the same spatial position in the image grid $/ \exists t, r \in \text{image}(t)$ and $s \in \text{image}(t - d_t)\big\}$. From these cooccurrence matrices, global motion features similar as those defined in [9] are extracted :

$$\begin{cases} f^1 = \sum_{(i,j)} P_{d_t}(i,j) \log(P_{d_t}(i,j)) \\ f^2 = \sum_{(i,j)} P_{d_t}(i,j) / [1 + (i-j)^2] \\ f^3 = \sum_{(i,j)} (i-j)^2 P_{d_t}(i,j) \\ f^4 = \left[\sum_{(i,j)} i^4 P_{d_t}(i,j)\right] / \left[\sum_{(i,j)} i^2 P_{d_t}(i,j)\right] - 3 \\ f^5 = \left[\sum_{(i,j)} (i-j)^4 P_{d_t}(i,j)\right] / \left[\sum_{(i,j)} (i-j)^2 P_{d_t}(i,j)\right] - 3 \end{cases} \tag{3}$$

where f^1 is the entropy, f^2 the inverse difference moment, f^3 the acceleration, f^4 the kurtosis and f^5 the difference kurtosis.

This set of global motion features is in this work computed over all the image grid. In order to cope with non-stationarity in the spatial domain, we can easily obtain a region-based characterization of motion. Indeed, the extraction of the motion descriptors can also be achieved either on predefined blocks or on extracted regions resulting from a spatial segmentation, since we focus only on temporal interactions. In that case, the retrieval process will consist in determining regions of sequences of the database similar in terms of motion properties to those characterized for the processed query.

4 Motion-based indexing and retrieval

4.1 Motion-based indexing

Since we plan to design an efficient indexing and retrieval scheme based on the global motion features presented above, we are required to build an appropriate representation of the database. This will allow us to recover easily sequences similar in terms of motion properties to a given video query. Thus, we have to make use of a classification method in order to cluster video sequences into meaningful groups.

Among the numerous clustering algorithms, we have selected an iterative process called *ascendant hierarchical classification* (AHC), [5]. Due to its simplicity of computation and its hierarchical nature, it reveals efficient for image and video database management as shown in [11]. It comes to compute a binary decision tree expressing the hierarchy of similarities between image sequences according to some metrics.

Let us consider a set of motion-related feature vectors, $f_n = (f_n^1, \dots, f_n^5)$ where n refers to a sequence in the database. The AHC algorithm proceeds incrementally as follows. At a given level of the hierarchy, pairs are formed by merging the closest clusters in the feature space in order to minimize the within-class variance and maximize the between-class centered second-order moment. We will use the Euclidean norm. Moreover, if an element n represented by a feature vector f_n is too far from all the others one *i.e.* $\min_m \|f_n - f_m\|_2 > V_{max}$, where V_{max} is a predefined constant, it forms also a new cluster. This procedure is iterated from the lowest level to the upper one in the hierarchy. To initialize the algorithm at the lowest level, each cluster corresponds to a unique sequence.

In our experiments, we have extracted the motion-based descriptors presented in section 3.2 with a temporal distance $d_t = 1$. Nevertheless, we cannot directly use the Euclidean norm with such features of different nature. In order to exploit this norm to compare feature vectors, we compute for the feature f^3 its square root and we raise the features f^4 and f^5 to the one fourth power.

4.2 Retrieval with query by example

We are interested in retrieving sequences of the database the most similar to a given video query. More particularly, we focus on matching sequences according to global motion properties. Indeed, the index structure described above provides us with such an efficient hierarchical motion-based retrieving tool.

We compute first the hierarchical index structure over the video database. Second, to handle the submitted query, the proposed sequence is processed to extract the meaningful motion-based features. In the same manner as previously, we compute the square root of the feature f^3 and the power one fourth of features f^4 and f^5 in order to use the Euclidean norm as cost function. Then, we explore the hierarchy of sequences as follows.

At its upper level, the retrieval algorithm selects the closest cluster, according to the Euclidean distance to the center of gravity of the considered cluster in the feature space. Then, for each of the children nodes, the distance from the feature vector of the query video to the center of gravity of each cluster is computed, and the cluster with the shortest distance is selected. This procedure is iterated through the index structure until a given number of answers or a given similarity accuracy is reached.

5 Results and concluding remarks

We make use of the approach described above to process a database of image sequences. We have paid a particular attention to choose video representative of various motion situations. Indeed, the database includes temporal textures such as fire or moving crowds, examples with an important motion activity such as sport video (basket, horse riding,...), rigid motion situations (cars, train, ...), and sequences with a low motion activity. Finally, we consider a database of 25 video sequences (typically, each sequence is composed of 10 images).

First, AHC is applied to the database in the space (f^1, f^2, f^3, f^4, f^5). In Figure 2, the representation of the database in the feature space, restricted to (f^3, f^4, f^5) space for visualization convenience, is reported. The four sequence classes of level 4 in the hierarchy are really related to different types of motion situations : the class "o" involves temporal textures, the class"x" includes sport video motions, elements of the class "+" are related to rigid motion situations and the class "." is composed of low motion activity examples.

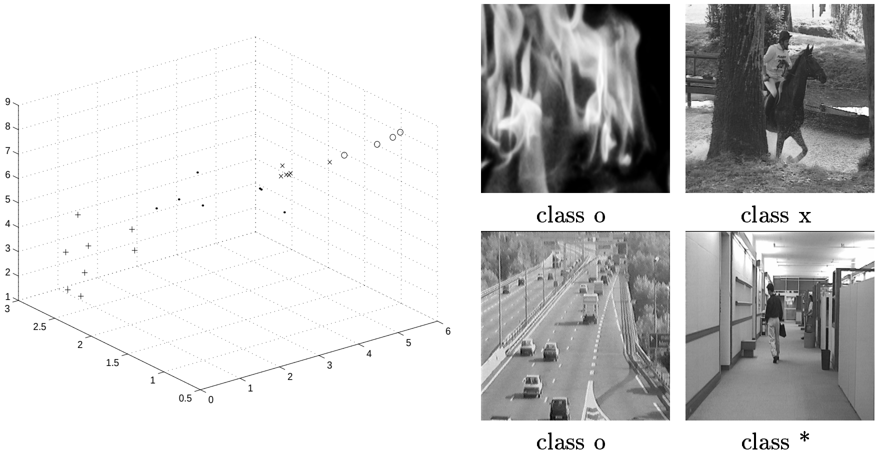
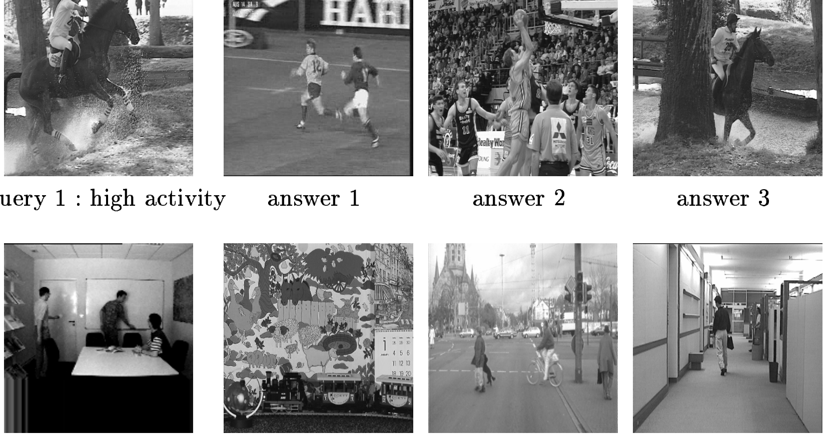


Figure 2.a Figure 2.b

Figure2. *Representation of the video database obtained with the AHC : a) Spreading of the sequences in the restricted feature space (f^3, f^4, f^5). Symbol (+,o,.,*) are indexes of classes at the level* 4 *in the AHC hierarchy. b) Examples representative of the extracted classes. We display the first image of the sequence which is the closest from the center of gravity of its class.*

Now, we deal with motion-based retrieval for query by example. Fig. 3 shows results obtained with two video queries. The maximum number of answers to



query 1 : high activity answer 1 answer 2 answer 3

query 2 : low activity answer 1 answer 2 answer 3

Figure3. *Results of motion-based retrieval operations with query by example for a maximum of three answers. We display for each selected sequence its first image.*

a given query is fixed to 3. The first example is a horse riding sequence. The retrieval process supplies accurate answers of sport shots which appear similar to the query in terms of global motion properties. The second video query is a static shot of a meeting. It is matched with other low motion activity sequences.

Let us proceed to a more quantitative evaluation of our approach. Since it seems difficult to directly analyze the accuracy of the classification scheme, we use the following procedure. First, we define a priori sequence classes among the dataset according to visual perception. Then, we analyze the three retrieved answers when considering each element of the base as a query. To evaluate the accuracy of our retrieval scheme, we consider two measures. We count the number of times that the query shot appears as the best answer, and, on the other hand, if the second retrieved sequence belongs to the same a priori class, we consider the retrieval process as correct. In practice, we have determined four a priori sequence classes : the first one with low motion activity, the second with rigid motions, important motion activity examples forms the third one, and temporal textures the fourth one. Even if this evaluation procedure remains somewhat subjective, it delivers a convincing validation of the indexing and retrieval process. Obtained results for the whole database are rather promising :

similar query and first retrieved answer (%)	80
correct classification rate according to a priori class (%)	75

Table1. *Evaluation of the motion-based indexing and retrieval process*

6 Conclusion

We have described an original method to extract global motion-related features and its application to video indexing and retrieval. Motion indexes rely on a second-order statistical analysis of temporal distributions of relevant local motion-related quantities. We exploit a hierarchical ascendant classification to infer a binary tree over the video database. Examples of retrieval using query by example have shown good results. In future work, we should determine optimal sets of global features adapted to different types of content in the video database, and evaluation over a still larger database should be performed.

Acknowledgments : This work is funded in part by AFIRST (Association Franco-Israelienne pour la Recherche Scientifique).

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, November 1996.
2. P. Bouthemy and R. Fablet. Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *Proc. Int. Conf. on Pattern Recognition, ICPR'98*, Brisbane, August 1998.
3. P. Bouthemy and F. Ganansia. Video partioning and camera motion characterization for content-based video indexing. In *Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, Lausanne, September 1996.
4. J.D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–625, April 1997.
5. E. Diday, G. Govaert, Y. Lechevallier, and J. Sidi. Clustering in pattern recognition. In *Digital Image Processing*, pages 19–58. J.-C. Simon, R. Haralick, eds, Kluwer edition, 1981.
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. B. Gunsel, A. Murat Tekalp, and P.J.L. van Beek. Content-based access to video objects : temporal segmentation, visual summarization and feature extraction. *Signal Processing*, 66:261–280, 1998.
8. A. Gupta, A. Hampaper, M. Gorkani, and R. Jain. On summarization of video. In *Proc. IEEE 4th Int Conf. on Image Processing, ICIP'97*, October 1997.
9. 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.
10. M. Irani and P. Anandan. Video indexing based on mosaic representation. *IEEE Trans. on PAMI*, 86(5):905–921, May 1998.
11. R. Milanese, D. Squire, and T. Pun. Correspondence analysis and hierarchical indexing for content-based image retrieval. In *Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, Lausanne, September 1996.
12. R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP : Image Understanding*, 56(1):78–99, July 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. K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii. Feature extraction of temporal texture based on spatiotemporal motion trajectory. In *Proc. Int. Conf. on Pattern Recognition, ICPR'98*, Brisbane, August 1998.
15. M. Szummer and R.W. Picard. Temporal texture modeling. In *Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, Lausanne, September 1996.
16. W. Xiong and J. C.H. Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, August 1998.
17. H.J. Zhang, J. Wu, D. Zhong, and S. Smolier. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 60(4), April 1997.