

Spatio-Temporal Segmentation and General Motion Characterization for Video Indexing and Retrieval

10th DELOS Workshop on Audio-Visual Digital Libraries, Santorini, Greece, June 1999

Ronan Fablet¹ and Patrick Bouthemy²

¹IRISA / CNRS

²IRISA / 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

1 Introduction

The use of video databases is of growing importance in various application fields concerned either with professional applications (remote sensing and meteorology from satellite images, road traffic surveillance from video sequences, medical imaging) or services targeted at a more general public (television archives including movies, documentaries, news ; multimedia publishing). Thus, the management of these databases demands fast, reliable and convenient access to visual information, which obviously involves content-based indexing. Surveys reviewing this issue can be found in [1, 12].

As far as video indexing is concerned, two main classes of applications can be distinguished. The first one includes various aspects related to browsing and navigation in large video databases. This requires to answer queries referring to the content of these videos, what implies a primary content-based indexing. The second aims at providing tools which facilitate video editing and enable to create interactive and structured video documents (hyper-video). This last application includes the generation of actions when designating a pre-indexed area in the image. This is of particular interest for education, tourism, teleshopping.

We have developed a scheme for structuring and indexing videos in relation to the dynamic content of image sequences. It includes, on one hand, spatio-temporal video structuring (Section 2), and on the other hand, the actual indexing aspect and the ability of retrieving video in terms of motion-based similarities (Section 3).

2 Spatio-temporal segmentation of in video sequences

2.1 Video partitioning and shot characterization

Content-based video indexing primarily requires to recover the temporal structure of video corresponding to the succession of its elementary shots, [1, 12]. We need to detect different types of transitions : cuts, progressive transitions such as fade-in, fade-out, dissolves or wipes. Substantial efforts have been devoted

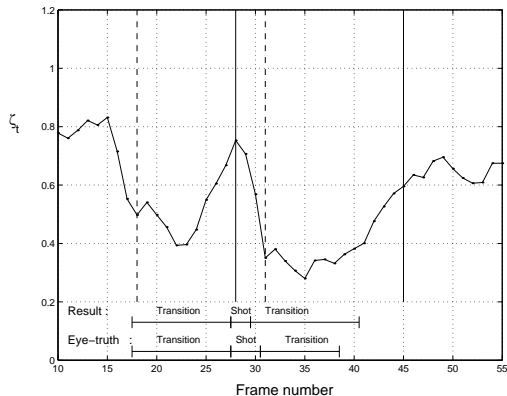
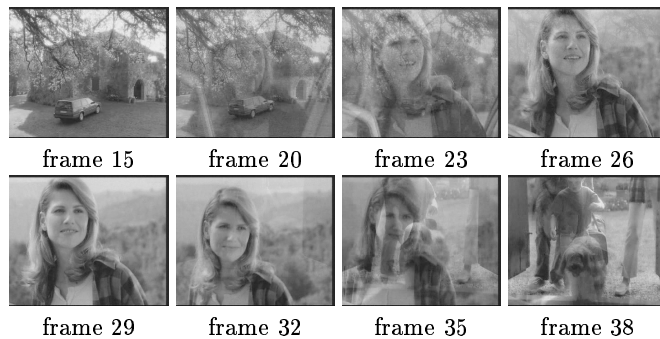


Figure 1. Excerpt of the sequence “Ajax” involving two dissolves separated by a shot containing few images. The plot displays the evolution of the quantity ζ_t w.r.t. the image number, the extracted transitions and shots, and the eye-determined ground truth.

to this issue, [9]. Since video databases are available in the compressed domain, direct processing of the MPEG bit-stream has also been considered, [15].

We have defined an original and efficient approach to handle this issue, [3]. Contrary to usual techniques directly relying on image information such as histogram comparison techniques, [9], our method exploits a more intrinsic information supplied by dominant motion between two successive images. The computation of this dominant motion is provided by the estimation of a 2D affine model using a robust method described in [14]. Besides, it also enables to easily characterize the type of each shot in terms of static shot, zooming, panning, tracking, from statistical likelihood tests applied to the estimated parameters of the dominant motion model assumed to account for camera motion.

The temporal segmentation of the video into shots relies on the temporal evolution of the size ζ_t of the support associated to the dominant motion estimated between two successive frames. When a cut occurs, no motion model can correctly describe the transformation between the two processed images, and the value of ζ_t suddenly falls down. When progressive transitions appear, this fall is less pronounced, but occurs over a longer interval. The detection of the relevant jumps of ζ_t is achieved using the accumulative Hinkley test, which is simple and efficient to compute. The main interest of our approach lies in the use of a unique framework for the detection of the different kinds of transitions. Furthermore, it involves only one parameter to be set by the user, and the same value enables

to process the different types of transitions. We report in Figure 1 an example of result on an advertisement sequence, involving many special effects.

2.2 Extraction and tracking of meaningful entities

Once temporal video partitioning is completed, we aim at characterizing the dynamic content of each extracted shot. To this end, the extraction of meaningful mobile entities in each shot is of key importance, [4, 10]. We have developed for this purpose several segmentation methods, [7, 10]. We only briefly outline hereafter one of them. The detection of moving entities is performed after estimating and cancelling the dominant image motion assumed to be due to the camera motion. A primary color-based segmentation allows us to better localize motion boundaries. The detection of moving objects relies on a Markovian modeling specified on an irregular graph relative to the adjacency graph of the spatial regions previously extracted, [7]. A limitation of this approach is that, if important perspective effects appear due to noticeable depth variations with regard to the distance camera-scene, the estimated representation of the 2D affine motion model representing the dominant motion can only embrace one part of the static scene (for example, the background). In that case, the extracted mobile regions can involve not only really moving objects in the scene, but also static parts of the scene which are not compensated (*e.g.* the foreground). Therefore, we have designed an additional step exploiting criteria based on projective geometry considerations, in order to distinguish these two classes of regions. Finally, we consider only regions corresponding to moving objects in the scene, [5].

In the context of interactive video environment, what is generally the relevant point of view for the types of applications considered here, we have proposed a region tracking technique which can be applied to any region predetermined by the user, [11]. One important point is that it can automatically detect tracking failure, and allows the user to reinitialize the procedure.

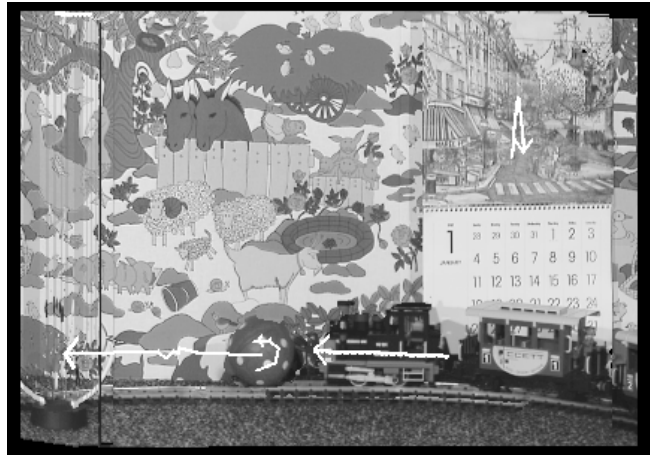


Figure 2. Iconic summary for the sequence “Mobi” formed by a mosaic image of the shot (involving a camera panning) with superimposition of computed trajectories of extracted and tracked moving objects

3 Motion information modeling and representation

3.1 Iconic summary and databases

Building a synthetic, structured and indexed representation of a video is of key importance for efficient browsing, navigation or retrieval in image sequence databases, [1, 8, 13]. To this end, we can supply different tools.

We can create video summaries for each extracted shots. One first straightforward possibility consists in selecting a key-frame properly representing the shot, [8]. In [10], we have proposed a richer representation of the content of each shot by means of a mosaic image which incorporates iconic annotations referring to moving entities, events (appearance/disappearance), and actions (trajectories). This is illustrated in Figure 2. Besides, an index database can be created using all these spatio-temporal features related to shot content and mobile entities extracted.

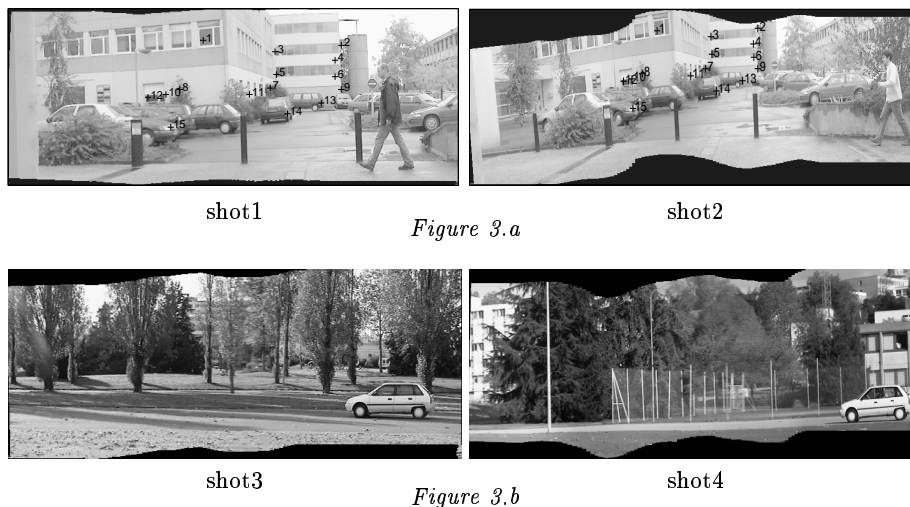


Figure3. Examples of creation of hyper-links between shots containing common parts in their background (buildings in shots 1 and 2 : figure 3.a) and the same mobile object (the car in shots 3 and 4 : figure 3.b). For each shot, we display the associated mosaic image.

It also appears necessary to create hyper-links between shots or videos to facilitate the navigation either in a single video or between different videos of a given database. In [2], we have described such a scheme for intra-video navigation, but it could be extended to video database management. We aim at matching elementary shots involving common entities. To this end, two types of entities are exploited : mosaic images and extracted mobile elements. Reconstructed panoramas will allow us to generate links between shots which contain common parts in the background, whereas, exploiting moving object matching, we can retrieve the different shots within which a given moving entity appears. Thus, through an appropriate interactive interface, the user can click on outlined objects displayed in an iconic summary of a shot (as for example, video sum-

maries previously described) with a view to recovering all the shots involving the selected pattern.

3.2 Global motion descriptors and similarity retrieval

The semantic characterization of shot content, [16], requires to define new motion descriptors similar as those exploited for color or texture characterization in still image analysis. We have developed an original motion-based feature extraction technique which handles a large variety of motion situations, [6]. It does not rely on parametric motion models, and does not require any primary image segmentation. Thus, we can characterize sequences involving for instance vehicles (rigid motion), persons (articulated motion), or rivers (fluid motions).

More precisely, we have adapted to global motion characterization techniques developed for texture analysis. In particular, our scheme relies on temporal cooccurrences of local motion-related measurements computed directly from the spatio-temporal derivatives of the intensity function. Then, from these cooccurrence distributions, we can extract global features which express properties of temporal coherence, acceleration, spreading out or complexity of the motion distribution in the considered shot.



Figure 4. 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.

We have exploited this approach for video indexing and retrieval with respect to the motion content, [6]. In order to reduce the search space when performing retrieval, we first determine a hierarchical classification of the considered video database. Then, we make use of this hierarchical structure to retrieve examples similar to a given video query in terms of dynamic content. It consists in exploring the classification tree and comparing features of the video query to those of tree nodes. Figure 4 displays results obtained with our retrieval scheme.

Acknowledgments : Gabriella Csurka, Thierry Dubois, Fabrice Ganansia and Marc Gelgon have contributed to the work reported in this paper. The authors

are thankful to INA, Département Innovation, Direction de la Recherche, for providing the MPEG-1 sequences Ajax, Avengers, Announcement, News, which are excerpts of the INA/GDR-ISIS video corpus. These studies were funded in part by French-Israeli scientific cooperation program AFIRST, Alcatel-CRC, the European project DiVAN EP24956, and DGA (student grant).

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, September 1996.
2. P. Bouthemy, Y. Dufournaud, R. Fablet, R. Mohr, S. Peleg, and A. Zomet. Video hyper-link creation for content-based browsing and navigation. submitted to Content-Based Multimedia Indexing workshop, CBMI99.
3. P. Bouthemy and F. Ganansia. Video partitioning 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. G. Csurka and P. Bouthemy. Direct identification of moving objects and background from 2D motion models. In *Proc. 7th IEEE int. Conf. on Computer Vision, ICCV'99*, Kerkyra, Greece, September 1999.
6. R. Fablet and P. Bouthemy. Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval. In *Proc. 3rd Int. Conf. on Visual Information and Information Systems, VISUAL'99*, Amsterdam, June 1999.
7. R. Fablet, P. Bouthemy, and M. Gelgon. Moving object detection in color image sequences using region-level graph labeling. In *Proc. 6th IEEE Int. Conf. on Image Processing, ICIP'99*, Kobe, Japan, October 1999.
8. A. Muffit Ferman and A. Murat Tekalp. Efficient filtering and clustering methods for temporal video segmentation and visual summarization. *Jal of Visual Communication and Image Representation*, 9(4):336–351, December 1998.
9. U. Gargi, R. Kasturi, and S. Antani. Performance characterization and comparison of video indexing algorithms. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'98*, pages 559–565, Santa-Barbara, June 1998.
10. M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. 5th Eur. Conf. on Computer Vision, ECCV'98*, Freiburg, June 1998.
11. M. Gelgon, P. Bouthemy, and T. Dubois. A region tracking technique with failure detection for an interactive video indexing environment. In *Proc. 3rd Int. Conf. on Visual Information and Information Systems, VISUAL'99*, Amsterdam, June 1999.
12. F. Idris and S. Pandranathan. Review of image and video indexing techniques. *Jal of Visual Communication and Image Representation*, 8(2):146–166, June 1997.
13. M. Irani and P. Anandan. Video indexing based on mosaic representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 86(5):905–921, May 1998.
14. J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.
15. N.V. Patel and I.K. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30(4):607–625, April 1997.
16. N. Vasconcelos and A. Lippman. A bayesian framework for semantic content characterization. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'98*, Santa-Barbara, June 1998.