

P. Bouthemy¹, Y. Dufournaud², R. Fablet¹, R. Mohr², S. Peleg³ and A. Zomet³

¹IRISA/INRIA-CNRS
Campus de Beaulieu
35042 Rennes Cedex - France
e-mail : bouthemy@irisa.fr

²IMAG-GRAVIR/INRIA
ZIRST-655 av. de l'Europe
38330 Montbonnot - France
e-mail : mohr@imag.fr

³Institute of Computer Science
Hebrew University
91904 Jerusalem - Israel
e-mail : peleg@cs.huji.ac.il

in Proc. Workshop on Content-Based Multimedia Indexing, CBMI'99

Abstract This paper describes an original approach for automatically building a navigation tool within a video. It aims at determining efficient representations of a video, in the context of content-based video browsing, relying on the creation of hyper-links between elementary shots of a video. We exploit matching techniques applied either on extracted meaningful entities (mobile objects) or on panoramic views of the scene. We present experiments on a real image sequence and discuss the extension of the method to navigation in a video database.

1 Introduction and related work

Video structuring issues are of key importance for browsing, content-based retrieval or navigation. High level knowledge formalisms enable to tackle these issues, exploiting for instance semantic networks. For instance, [12] distinguishes five levels of codifications for cinematic levels to formalize a general system using video data for program synthesis, video search or content-based retrieval. The corresponding encoded information is far too complex to be extracted automatically from the video input and requires tedious manual encoding.

As a consequence, there is a real need for video structuring tools based on feature extraction and relying on visual content (see for instance [1, 2]). In this work, we consider this point of view with a particular emphasis on video browsing applications. We restrict the field to image sequence analysis and acknowledge that sound processing is a convenient additional tool for video processing, especially when considering applications such as video conferences ([18, 24]).

Video has led to its own direction in the research carried out for accessing content-based information, [1, 2, 30]. Issues concerned with video content representation benefit from studies devoted to temporal video segmentation, mosaic image construction or object tracking, to build iconic shot summaries, [6, 7, 10, 25], and these were evaluated to be comfortable [5]. Nevertheless, these schemes enable only to browse efficiently the content of a video through its shots. Besides, video structuring has also been addressed by means of shot clustering techniques with a view to identifying groups of shots which form a scene. However, in both cases, the representation of a video remains closely related to the linear fashion video are acquired along the time axis, and one needs that it can be represented in a way which enables the user to browse the information in the video more efficiently. For instance, it is desirable that the user can focus on a given object of interest, jump to shots where it appears, . . .

This paper describes how the combination of appropriate automatic video content analysis techniques can contribute to solve these issues. This work does not aim at developing a complete prototype but rather at demonstrating the feasibility of several pertinent functionalities related to video structuring, indexing and browsing. In fact, access to content-based information can be performed either via a content index, or through content-based hyper-links between shots. We consider the latter aspect which represents an important improvement in the indexing process to cope with high-level video structuring. As a consequence, we intend to create hyper-links between shots which present similarities in terms of spatio-temporal content. In order to define a set of meaningful content-based video index, our analysis relies on two types of entities : the background scene and moving objects.

Several processing steps are needed to extract such a representation of a video sequence automatically. First, the temporal structure of the video is recovered such that every elementary shot contains a continuous event captured by the camera. This requires to identify transitions in the video. Transitions may be abrupt (cuts) or progressive (fade, dissolve, wipe effects). Hence, each shot is analyzed separately and an iconic summary representing its spatio-temporal content the shot is created. On one hand, we attach to each shot a mosaic image of the background of the observed scene. This is achieved by merging the

background images into a single panoramic image, having a valid geometric and photometric quality. This background representation would allow not only to get a still image which condenses the shot content, but it can also be seen as a complementary modality for summarizing such a shot, particularly useful when the observer has to focus on details which are better observed in still images than in moving ones. On the other hand, moving objects in the scene are regarded as meaningful entities to cope with for indexing purpose. Therefore, a major task in the analysis of a shot is motion segmentation, *i.e.*, finding the image regions that belong to the background and the regions that belong to moving objects. The difficult point is that image motion is a result of both camera motion and changes in the scene. At this stage, we compute in a proper way the apparent dominant motion between successive images. If we assume that this dominant motion is due to the camera movement, we can detect moving regions in the scene, which can be further clustered into objects.

Once the indexing process is completed for each elementary shot, we extend it by linking two shots with similar moving objects or background parts. To this end, we exploit image matching techniques. Videos often display moving deformable objects on complex background, and may involve illumination changes. Image matching in such a general context is challenging. Furthermore, the matching scheme has to be performed on a large set of images, which is also challenging. Therefore, sub-linear searching mechanisms would be preferred. Image matching in the general case of complex scenes has often been achieved using global descriptors, especially color distribution, which proves to provide successful results as reported in [27]. Exploiting correlogram [9], some improvements can be obtained. Such descriptors can be split into subsets. This enables to speed up the search by looking for only in the potential relevant subsets. Such global methods can be applied either on entire images, or on segmented regions, and they are used in both cases in the results provided here. However, if the goal is to find non segmented objects, the results of such methods are very poor in presence of noise due to occlusion and clutter. As a consequence, in this case, we prefer to consider local methods. Finally, we can benefit from XML framework to encode our video indexing structure.

The subsequent is organized as follows. Section 2 describes the different stages concerned with video segmentation and representation. In Section 3, we present how these representations can be exploited to create video hyper-links. In Section 4, we conclude and discuss further developments.

2 Video segmentation and representation

2.1 Segmentation into shots and extraction of moving entities

The earlier phases of a content-based video indexing scheme are targeted at partitioning video into shots and extracting meaningful entities for each shot. Considering the first aspect, we exploit an original approach in order to determine the temporal segmentation of the video into shots. It relies on a robust, multi-resolution and incremental estimation of a 2D affine motion model between successive frames, accounting for the global dominant image motion, [14]. Therefore, it tolerates the presence of moving objects in the scene. Our scheme comes to study the temporal evolution of the size of the estimation support associated to this global dominant motion in order to detect shot changes. Our approach embraces the detection of both cuts and progressive transitions (*e.g.* dissolve, wipes,...) with the same technique and the same parameterization, [3, 4]. It was proven to be accurate and efficient even in quite complex situations as reported in [4].

Once the temporal segmentation into shots completed, attention can be paid to the representation of each shot. Since moving objects represent meaningful entities of the spatio-temporal content, we aim at extracting these mobile areas from the static background. To this end, we exploit a motion-based criterion, taking advantage of the dominant motion in the images computed in the video partitioning step. The problem is stated as the detection of regions non conforming to this global dominant motion. A multi-scale Markovian approach, [15], ensures a proper regularization and an efficient extraction of moving objects. This last point is of key importance when coping with content-based indexing in order to ensure the relevance of spatial index computed for each mobile entity.

To highlight the capability of our approach in the context of video hyper-link creation with a view to video browsing, we have performed experiments on several video sequences representative of various dynamic contents. We only report here one such experiment. The video sequence processed for illustration is composed of six different shots (see Fig. 1). Each shot involves a main mobile object, either a pedestrian or a car, tracked by the camera. Three objects are present in two shots :

- the same pedestrian appears in shots 2 and 4 ;
- shots 1 and 6 involves the “Renault” car ;

– the “Citroën” car is present in shots 3 and 5.

Besides, we have introduced a progressive transition between shots 5 and 6.

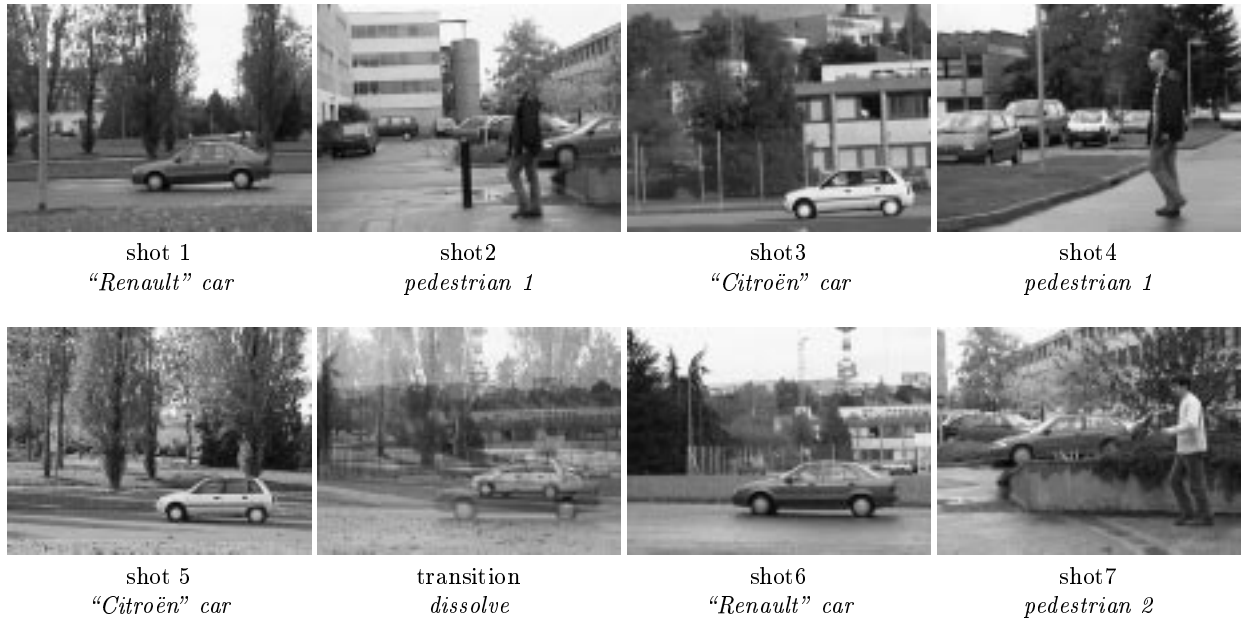


Figure1. Processed video sequence. For each extracted shot, we display the first frame. The middle frame of the detected progressive transition is also included.

We first perform the temporal segmentation of this video. All cuts are correctly detected and the bounds of the progressive transition are exactly recovered. This is of major importance in our case since all the subsequent developments depend on the quality of this first step. Hence, we apply for each determined shot the motion detection scheme. Results are displayed in Fig. 2. Although we tackle either non-rigid motions (pedestrian walk) or poorly textured mobile objects (cars), the motion detection reveals accurate enough to compute relevant spatial index as shown in Section 3.

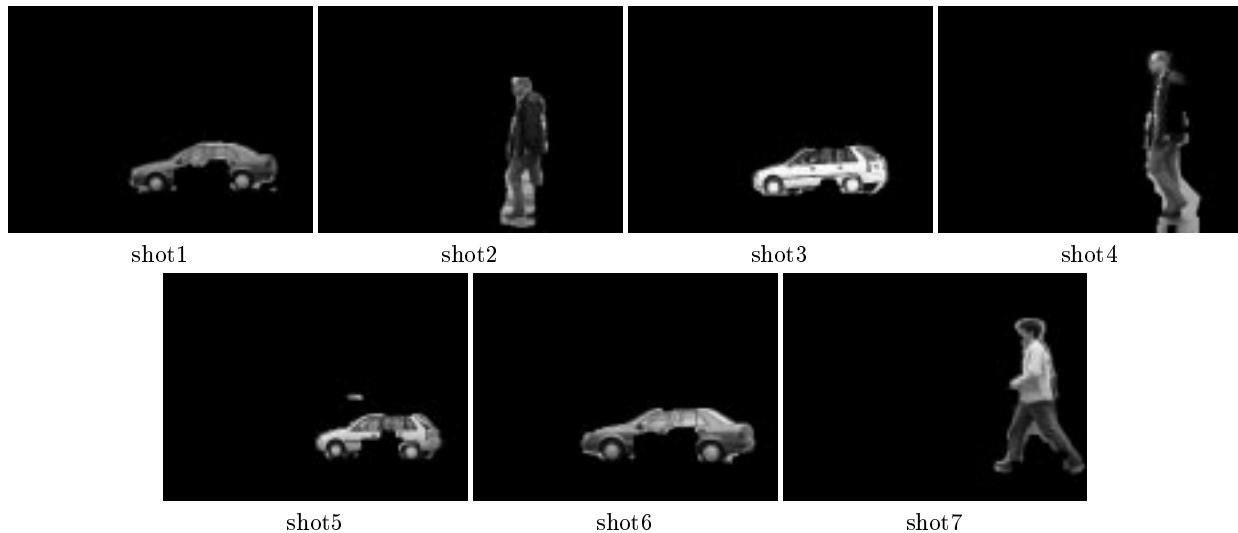


Figure2. Examples of detected moving objects for each shot

2.2 Representation of the spatio-temporal shot content

The creation of a relatively high-level representation of a video first requires to determine an appropriate characterization of the spatio-temporal content. First, the motion detection stage leads to create a

database of images of mobile objects which can be used to find similar entities in the video. Besides, video sequences of a scene can be compactly represented in a single static image using panoramic mosaicing. The images are projected onto a common manifold, [11, 17, 19, 20, 28], to create a single panoramic image.

As the first stage in the mosaicing process, the camera motion between each two successive frames is found, and the images are aligned accordingly. Only background regions, found in the previous segmentation stage, are used for alignment and mosaicing. Having the aligned background images, the panoramic image is created by pasting together image strips. Indeed, strip pasting for mosaicing avoids the blur associated with combining regions that are not aligned perfectly. Best mosaicing is obtained when the strips are perpendicular to the background image motion. Besides, since the resolution is higher in the middle of the image, selected strips are chosen close to the center of the image grid. Strip pasting can thus handle any camera motion, including rotations, forward/backward translations, and sideways motion. For further details, see [17, 19, 20].

In fact, we have exploited this technique to process each shot of the considered video example. The obtained panoramic views are displayed in Fig. 3. No misalignment can be visually detected, what ensures the reliability of the indexing of these mosaics relying on local properties of the intensity function as shown in the next section.

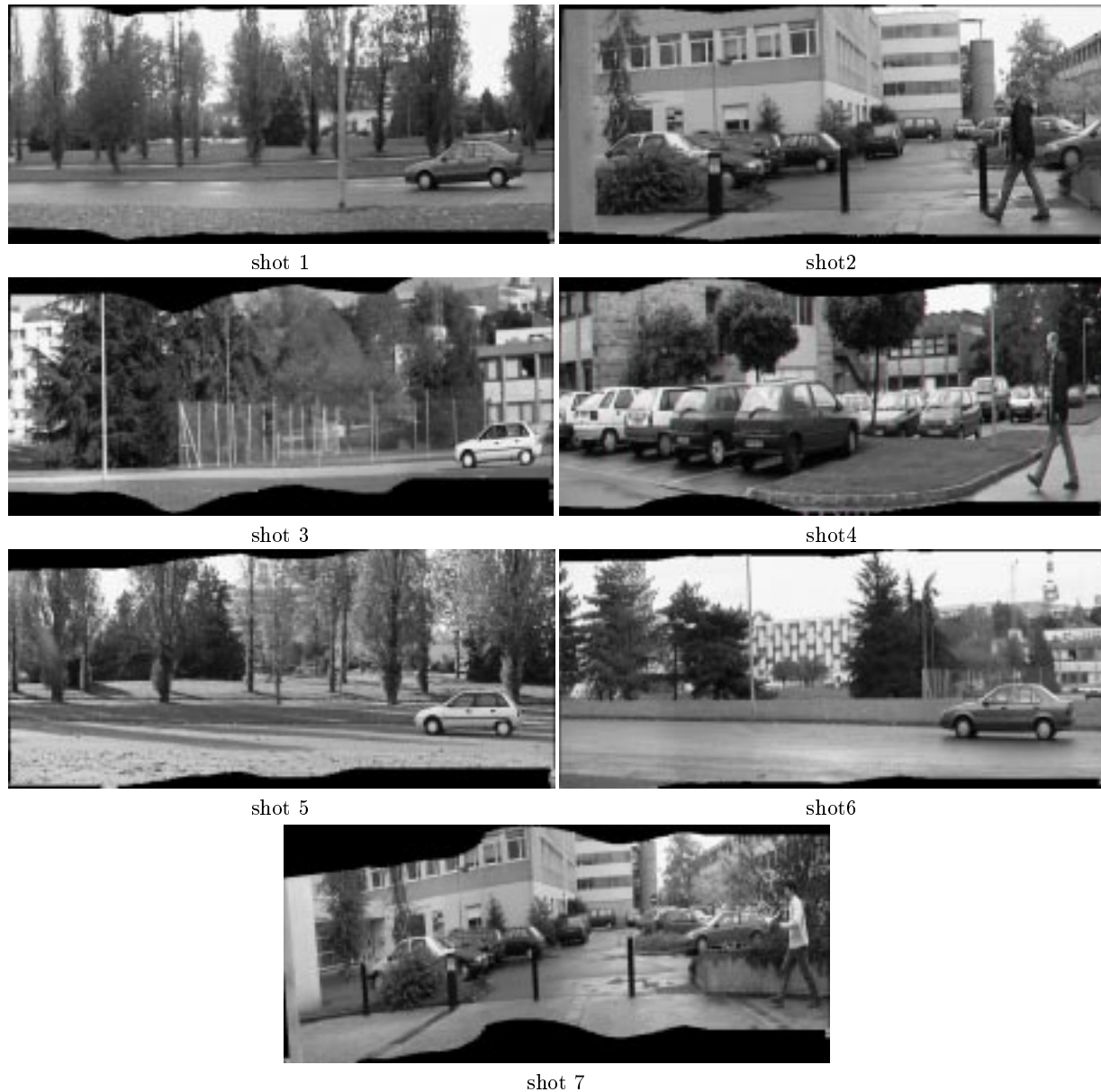


Figure3. Panoramic view built for each extracted shot

3 Finding similar objects

If a mosaic can provide useful abstract of a single shot, more high level structures have also to be extracted in order to supply a smart navigation. We present here just how visual information can be used to establish links between shots or objects, from where a jump backwards or forwards to “the other shots where this object has appeared” is easy to implement.

The key issue is here to match the corresponding objects. Two cases have to be considered : when objects are extracted using the tools considered in the previous section, or when they are not, as objects appearing in a mosaic without extraction. Two different kinds of tools are implemented for these two cases. In case of segmented objects, global descriptors are used; they conjugate easy implementation and speed. In the other case, we have considered the differential invariants presented in [23].

3.1 Global measures

The first global measure tested was color histogram [27], on which we have performed a χ^2 distance, since this has been proven to be the best relevant distance for image comparison [22]. Histograms are implemented in a straightforward way using bins derived from the standard splitting of each of the R, G and B axes.

Obviously, color histograms will fail when illumination strongly changes. A way to deal with this problem is to modify the set of histogram bins in order to take into account chromaticity as done in [25]. Moreover, when similar color distributions occur, any kind of color histogram refinement will fail. This can be overcome by using correlograms [9], where not only color distribution is taken into account, but also some kind of spatial distribution.

We have applied both techniques on the set of objects extracted using the motion detection scheme. On this small set of entities of the processed example described above both methods succeed perfectly. These tools have also been implemented in more difficult conditions. We have processed a larger database with four video sequences having together 109 shots, and 42 extracted objects, some of them quite similar (*i.e.* people similarly dressed). These video also involve complex occlusion and lighting changes. In such hard conditions, the best first match is considered, and the histogram technique recovers the correct object for 64% of the cases, and the correlogram technique for 80% of the cases.

3.2 Local measures

With non segmented objects, clutter and occlusion are killing the global methods described above. Therefore, local approaches are required. The approach we are considering here is taken from [23], it considers local image data around particular points. Alternative features could be considered like in contour matching issues. However, these methods should be made robust for considering the general context of this study : large clutter and occlusion, necessity for indexing the potential matches in order to reduce the search complexity. For instance, [21] provides an interesting robust alternative to conventional contour marching, but this tool is time consuming, and no indexing mechanism has yet been designed for this approach in order to speed up the matching process.

Therefore, following [23], interest points are extracted, and around these interest points local grey level signatures can be computed; these signatures are invariant to rotation and grey level affine shift. In order to capture potential changes in zooming aspects, they are computed at different scales. Locality implies robustness to occlusion; invariance implies robustness to geometry or illumination changes. On the top of that, a robust decision makes the whole process resilient to the presence of outliers.

The method can be briefly summarized as :

- extract feature points and compute signatures at different resolutions around these points;
- match these points with all possible signatures extracted on different objects;
- select the objects for which the number of matches is large enough and if they are geometrically coherent.

We exploit this approach to index the mosaic images with a view to matching common elements present in the scene relative to the different shots. The method provides for instance perfect matches for the feature points selected on the mosaic presented in Fig 4; this enables to identify the corresponding background in both views. On a medium size set of images (up to more than thousand object images), this method proved to be efficient if viewing conditions are not too difficult. Experiments have been performed on objects corresponding to a viewing direction changes lower than 20 degrees, and with object size in image larger than 20000 pixels, zooming scale less than 2, and the objects were at least slightly textured.



Figure 4. Examples of matching and correct correspondence (100% of correct matches in this case)

3.3 Building video hyper-links

Once matches between objects in video are performed, distance between objects can be computed. This leads to create classes representing for instance a person seen in different parts of the video, or an even more conceptual class like “car”. Hierarchical clustering has been used for this purpose. However, whatever method will be used, it has to be pointed out that classes will never be constructed in a fully automatic way : the same person, seen in a front close-up view and in a sided far view, gives rise to descriptors which are certainly too different. Hence, this two “entities” have to be put together by human interaction. Similarly a red car and a black car will never have the same descriptors at the information level we are considering. This is in accordance with matching results reported in the previous subsection.

Therefore, we have implemented an interface to supervise the class construction by interacting with the system : *e.g.* splitting suggested classes, or telling the system that these two objects have to be considered as similar. From there, links are constructed to browse the video in the following way. A video is displayed shot by shot (either continuously or with mosaics), a click on an indexed object will produce a move forward or backwards to find the next shot containing the object of the same class. This allows the user to browse the video by focusing on a particular object of interest.

This could be extended to inter-video links; there is not difficulty on the concepts to be implemented. However, such a construction will have to handle thousands of views. Building the links will need to search in a large amount of data descriptors; the relative accuracy of the descriptors and the high dimension of the vector data lead then to a data access problem presently not solved (see Section 4).

3.4 Encoding the video structure in XML

The resulting files that describe the data collected in the previous steps have been encoded in XML; for such a purpose, a Document Type Definition (DTD) had to be defined and the formal definition can be found in [16].

The encoding from the data structure files here described into XML was straightforward. Much more work was devoted to specify the DTD; this DTD contains also aspects that were not worked out here; In short, the DTD addresses:

- video structures: they comprise sequences and scenes that are not yet extracted in the present state of our work (see however [8] on how to proceed for these steps) and video shots with the corresponding transitions;
- object descriptions (called *class* in [16]) involving text associated with the description, type (person, etc.) and the list of occurrences of the corresponding occurrences;
- occurrences of objects within a shot (duration, which shot, descriptors), action to be performed if the object is clicked during interaction;
- events with duration, type and text descriptors;
- temporal and visual relations between the different entities described here above.

Such XML descriptors also enable then post processing like media synchronization, or interaction like displaying an object descriptor when the corresponding occurrence is designated by the mouse.

4 Discussion - Conclusion

4.1 On the work presented

We have presented in this paper how automatic content-based video analysis could be used to facilitate video structuring. In fact, the creation of hyper-links between shots extends the indexing process by

supplying an enriched representation of video content. Furthermore, it provides new functionalities which allow the user to browse a video in a way much more convenient than the usual display shot by shot, since the user can focus on a particular object. Our purpose was to prove the feasibility of the developments of such functionalities by combining different video processing stages with a view to introducing this kind of schemes in future video indexing prototypes. In order to build these links, the spatio-temporal content is represented by a panoramic view of the scene along with extracted moving objects. Hence, these entities are matched relying either on local invariant techniques or global description approaches. To illustrate our scheme, we have reported experiments on a real video sequence involving various dynamic contents.

Construction of hyper-links between objects in video has been implemented in two cases: with segmented objects and non segmented objects. In both cases, we obtained 80 to 100% of correct matches depending on how severe the experimental conditions were, relying either on global descriptors for segmented objects or local invariants for mosaic images. A human operator can validate and correct these links. At this stage, they permit to browse the video with a particular attention on an object : clicking on the object brings the system to jump to the next shot where it appears.

4.2 Extension and problems

There are several ways to extend the work presented here. First, we can enlarge the video characterization stage. In particular, we currently investigate feature-based techniques supplying global characterization of the motion content of a shot and leading to retrieval with query by example based on motion information. Besides, relying on shot clustering, we also aim at structuring video in groups of shots (*i.e.* sequences) which present sufficient similarities. Furthermore, we can enhance the resolution of the mosaic images by taking advantage of the overlap of images in the sequence.

In addition to the above mentioned aspects, we could use different invariant descriptors for matching : color invariants for instance. To cope with appearance changes, a solution is probably to capture in a shot object variability. To this end, we could combine the representations of a single object, which are relative to different matched entities. The result will no more be a single model, but more likely a combination of models as proposed in [26] for the face case.

The building of links between shots presented here can be extended towards inter-video links. We have then to tackle the problem of matching entities in a very large database of models. Handling the construction of the index database with a view to search of similar features in high dimensional space is a difficult problem that cannot be solved with standard database techniques [29] nor with clever data structures [13]. We did not address the problem here. We just point out that in the index space we were considering, when searching for the best match within a given distance (a problem slightly different from the one considered in [13, 29]) a good speed up could be obtained with tree structures. Nevertheless, the largest test we performed was limited to a set of 150000 descriptors, which is far below what the processing of a large set of video will supply. Therefore, more experimentations on larger data set and a deeper analysis have to be carried out. Search for alternative solution such as data reduction has also to be conducted in parallel.

Acknowledgments

The authors would like to thank R. Hammoud, C. Schmid and F. Spindler who made their codes available to us. This work was partially supported by AFIRST (Association Franco-Israélienne pour la Recherche Scientifique).

References

1. G. Ahanger and T.D.C. Little. A survey of technologies for parsing and indexing digital video. *Jal of Visual Communication and Image Representation*, 7(1):28–43, January 1996.
2. P. Aigrain, H.-J. Zhang, and D. Petrovic. Content-based representation and retrieval of visual media. *Multimedia Tools and Applications*, 3:179–202, 1996.
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. P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, 1999. To appear.
5. M.G. Christel, M.A. Smith, C.R. Taylor, and D.B. Wincler. Evolving video skims into useful multimedia abstractions. In *Proceedings of the CHI'98 Conference on Human Factors in Computing Systems*, 1998.

6. 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.
7. B. Günsel, 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. R. Hammoud, L. Chen, and F. Fontaine. An extensible spatial-temporal model for semantic video segmentation. In *First Int. Forum on Multimedia and Image Processing*, Anchorage, Alaska, 1998.
9. J. Huang, S. Ravi Kumar, M. Mitra, W.J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 762–768, June 1997.
10. 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.
11. R. Kumar, P. Anandan, M. Irani, J. Bergen, and K. Hanna. Representation of scenes from collections of images. In *IEEE Workshop on Representations of Visual Scenes*, pages 10–17, 1995.
12. C. A. Lindley and A.M. Vercoustre. A specification language for dynamic virtual video sequence generation. In *Int. Symposium Audio, Video, Image Processing and Intelligent Applications*, pages 17–21, Baden-Baden, Germany, August 1998.
13. S.A. Nene and S.K. Nayar. A simple algorithm for nearest neighbor search in high dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):989–1003, 1997.
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. 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 Academic Publisher edition, 1997.
16. Opera. Dtd for video. Inria Rhne-Alpes, <http://www.inrialpes.fr/opera/dtdvideo.txt>, 1999.
17. S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 338–343, June 1997.
18. R. Razman, R. Al-Halimi, W. Hun, and M. Mantei. Four paradigms for indexing video conferences. *IEEE MultiMedia*, 3(1):63–73, 1996.
19. B. Rousso, S. Peleg, and I. Finci. Mosaicing with generalized strips. In *DARPA Image Understanding Workshop*, pages 255–260, New Orleans, USA, May 1997.
20. B. Rousso, S. Peleg, I. Finci, and A. Rav-Acha. Universal mosaicing using pipe projection. In *International Conf. on Computer Vision*, pages 945–952, January 1998.
21. W.J. Rucklidge. Locating objects using the Hausdorff distance. In *Proceedings of the 5th International Conference on Computer Vision*, pages 457–464, Cambridge, USA, 1995.
22. B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proceedings of the 4th European Conference on Computer Vision*, pages 610–619, Cambridge, UK, 1996.
23. C. Schmid and R. Mohr. Local gray value invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.
24. M. Seck, F. Bimbot, D. Zugaĵ, and B. Delyon. Two-class audio signal segmentation for speech-music-noise detection. In *Proc. 6th Eur. Conf. on Speech Communication and Technology, EUROSPEECH '99*, Budapest, Hungary, September 1999.
25. M. Smith and T. Kanade. Video skimming for quick browsing based on audio and image characterization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1997.
26. K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.
27. M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.
28. R. Szeliski. Video mosaics for virtual environments. *IEEE Computer Graphics and Applications*, 16:22–30, March 1996.
29. R. Weber and P. Zezula. A quantitative analysis of performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24th VLDB Conf.*, 1998.
30. H.J. Zhang. Swim : A prototype environment for visual media retrieval. in *Recent Developments in Computer Vision*, pages 531–540, 1996. S.Z. Li, D.P. Mital, E.K. Teoh, H. Wang (Eds.), LNCS 1035, Springer.