VIDEO SEGMENTATION BASED ON 1D-MOSAICS REPRESENTATION

Anne Manoury, Henri Nicolas
IRISA/INRIA, Campus de Beaulieu
35042 Rennes Cedex, France
henri.nicolas@irisa.fr

ABSTRACT

In this paper, a new method is presented for video navigation based on a hyper-scene representation of MPEG-2 compressed videos. For that purpose, each shot of the video sequence is represented using two 1-D mosaic images obtained using the MPEG-2 motion vectors. The clustering process is based on the evaluation of the similarities between 1-D mosaics. Two similarity criteria, respectively based on a global comparison and a matching distance after motion compensation, have been designed. Furthermore, in order to reduce as much as possible the computational complexity, the mosaic images are represented using a Linear Piecewise Representation. Experimental results have been performed on documentary video sequences. Satisfactory hierarchical hyper-scene representations of these sequences have been obtained.

1. INTRODUCTION

With the rapid development of applications such as video broadcasting, interactive TV or numerical recorder, huge video database are now available. This creates difficulties to access and index the video database content. To solve this problem, video indexing techniques are under development in order to provide a description of the video content useful for video retrieval and archiving. The video sequences are usually decomposed into shots which represent the basic data unit manipulated by many video indexing algorithms. Several authors focused on this topic and a review was proposed in [1]. Since a video may contained many shots, it may be useful to merge successive shots when they represent a similar content. The similarity criterion is therefore defined according to the applications constraints. Furthermore, the user may want to navigate into a video sequence in order to visualize only scenes which are interesting for him, such as, for example, outdoor scenes in a documentary, or specific actions in a sport event. In order to be able to do that easily, the video content must be described using a non linearly time dependent organized structure. This segmentation process can logically be based on an initial video shot decomposition, and consecutive and/or non-consecutive shots may be merged in order to create *hyper-scenes*. The design of such decomposition requires defining relevant similarity criteria between shots and/or scenes. Furthermore, this structuring process cannot always be performed using original video sequences, since only the compressed data may be available. The method proposed in this paper is therefore based on the use of MPEG-2 compressed video data. The motion parameters available in the MPEG-2 stream are used which avoid a time consuming motion estimation phase.

1. GENERAL PRINCIPLE

The algorithm used to classify the shots according to the proposed structure is based on an unsupervised hierarchical classification algorithm [2]. The video content is described using three levels hierarchical structure based on shots, scenes and hyper-scenes. A video shot is defined as a group of images which appeared to have been continuously captured. This is the smallest element of the structure. The second one called *Scene* is composed by a group of consecutive shots. The last one, the *Hyper-Scene* level, contains scenes and/or shots which are not time-connected. The similarity measure is based on the definition of a distance between scenes and/or shot content. If N represents the initial number of shots, then N levels are defined in this hierarchy. At the lowest level, each shot represents a hyper-scene, while at the upper level the whole sequence is considered as a unique hyper-scene. A recursive fine to coarse merging process is performed where at each iteration; the two closest clusters are merged. The user can therefore navigate in the video using this hyper-scene decomposition. The hierarchical level can be adapted according to user whishes. In the context of this paper, the initial shot decomposition is considered to be available.

2. SHOT CHARACTERISTICS

For complexity and practical reasons, it is not reasonable to use all original images to evaluate the distance between two shots. A first solution consists in the selection in each shot of one image, considered as representative of the shot content, and to use it to define similarity criteria. Nevertheless, the selected image may not sufficiently reflect the shot content, mainly when the

camera is not fixed. It is also possible to use several images, but this solution makes the shot merged process more complex, and anyway, the use of several images does not fully guarantee that the shot content is entirely represented. The use of mosaic images appears therefore as a promising solution since it is by definition a summary of the global visual content of the shot background. The use of mosaic images in video representation and indexing is overviewed in [3,4]. Nevertheless, a classical 2-D mosaic representation may represent a huge image when the camera displacement is large. In order to further reduce the amount of data, 1-D mosaic representations can therefore be used (see [5]). Basically, such 1-D mosaic image is the projection, after camera motion compensation, of the successive images on a predefined direction. The use of two 1-D mosaic images allows generally a satisfactory description of the shot content. Horizontal and vertical projections are used in the context of this paper. The method proposed in [6] is used to create the 1-D mosaics. It basically performs as follows: the 2-D mosaic image is first built by aligning all the frames by motion compensation using the MPEG-2 motion vectors. Then, a blending process is used to remove the foreground objects by detecting the outliers, and the mosaic is projected on the retained directions.

The 1-D mosaic signals can be represented using a Piecewise Linear Representation (LPR) [6] to significantly reduce the amount of manipulated data. This technique refers to the approximation of a signal using straight lines. The LPR decomposition algorithm is based on a linear interpolation together with a top-down approach. It performs as follows: for each current mosaic segment, the breaking point corresponding to the highest extrema is detected. It is validated if the mean squared error remains higher than a predefined threshold. The LPR decomposition stops when the error is lower than the threshold for each segment.

3. INTER-SHOT DISTANCES

The definition of an inter-shot distance requires the definition of similarity criteria. These criteria are defined here according to the following principles:

- Evaluation of the global shot similarity in order to detect shots which have approximately the same color characteristics.
- Evaluation of the matching error after motion compensation in order to detect shots which correspond to the same scene but which have been acquired with a moving camera.

Based on these principles, the two merging criteria are defined as follows:

Global distance.
In order to evaluate the degree of homogeneity between two mosaics M_1 and M_2, they are split into S segments. The average value $\overline{S}$ is therefore computed for each segment. The distance between two 1-D mosaics is then obtained by matching each block of the first mosaic to a block of the second one such as establishing a bijection between the S segments of each mosaic which minimizes the distance. The global inter-mosaic distance is therefore expressed as:

$$GB(M_1, M_2) = \sum_{i=1}^{S} \left(\left| \overline{S}_{1,i}^{h} - \overline{S}_{2,i}^{h} \right| + \left| \overline{S}_{1,i}^{v} - \overline{S}_{2,i}^{v} \right| \right)$$

h and v denote the horizontal and vertical directions corresponding to the mosaics.

Matching distance after motion compensation.
This second distance is calculated as the matching error obtained after motion compensation of the camera displacement. The motion model takes into account zoom and shifting camera displacements in the 1-D space. The matching error between two shots i and j represented by their horizontal and vertical 1-D mosaics M_i^h, M_j^h, M_i^v and M_j^v is therefore computed as follows:

$$ME(M_1, M_2) = \arg\min_{t,k} \frac{1}{Card[N_{1,2}]} \left[\sum_{p \in N_{1,2}} \left| M_1^h(p) - M_2^h(p + d(t,f)) \right| + \sum \left| M_1^v(p) - M_2^v(p + d(t,f)) \right| \right]$$

where $N_{1,2}$ denotes the number of overlapped pixels after motion compensation. $D(T,k)$ is the displacement vector computed as:

$$d = T + k.(p - m)$$

where T is the shifting parameter, k the zoom coefficient, and m the middle of mosaic M. In practice the distance minimization is performed using a full search matching algorithm with a quarter pixel precision for T, and 0.005 precision on k. Furthermore, since the L1 norm is used, the error corresponds to the area between the two mosaics, and can therefore be analytically computed using the LPR mosaic representation, which leads to a fast error computation.
The error term ME is significant only if the overlap area is sufficiently large. This may not be the case if the minimization is performed using the previous equation. In order to avoid this problem, two modifications are introduced: 1) the overlap cannot represent less than

25% of the average mosaic length; 2) the error term is divided by the overlapped length in order to privilege a larger overlap area. Then we have:

$$ME(M_1, M_2) = \arg\min_{t,k} \frac{1}{Card[N_{1,2}]^2}$$
$$[\sum_{p \in N_{1,2}} |M_1^h(p) - M_2^h(p + d(t,f))| + \sum |M_1^v(p) - M_2^v(p + d(t,f))|]$$

Finally, the distance between two shots is defined as:

$$D(M_1, M_2) = \min[GB(M_1, M_2), aME(M_1, M_2)]$$

where a is a weight coefficient. It value depends on the the application constraints. The global distance between a shot and a hyper scene (or between two hyper-scenes) is defined as the average distance *D* between each shot composing the hyper-scene and the considered shot.

4. EXPERIMENTAL RESULTS

The method proposed in this paper is designed for a content for which it is difficult to propose an a priori model (such as for sport programs or news journals). In order to test its performance we collected a corpus of artistic content, namely feature documentaries produced by SFRS. The corpus is constituted of 12 feature documentaries and is of 6 hours duration. The proposed method was tested on various excerpts of the corpus and method assessment has been done by comparison of automatic scene grouping method and indexing of scenes by a professional on one entire film "Acquaculture in Mediteranée" which contains 158 shots. In the following figures, each shot is represented by a key-frame systematically chosen at the middle of the shot (it should be pointed out that the key images do not always represent the temporal variations of the shot content). Figure 1 and 2 shows the two 1-D mosaic images for six shots, and the breaking points of the linear segment representation. Figure 3 shows (partially) the obtained hierarchical hyper-scene representation. It can be observed that the merged process is generally performed in a logical order even if some shots have sometimes been erroneously merged. In order to have a fair evaluation of the performance and the utility of our system, we compared the hyper-scene decomposition with a manual decomposition performed independently by a professional archivist. 8 hyper-scenes have been obtained by the automatic classification, and 7 by the archivist. Three of them (HS 5, 7 and 8) are similar. HS 1 and 2 are identical at around 70 %, while H3 and 4 are less similar (these two last HS correspond more to a semantic clustering than the previous ones). It should be noticed that even manually, such decomposition on this kind of video data is highly subjective, and depends on the semantic perception of the user. The conclusion that we can derive from this comparison is that the proposed automatic method is able to provide with coherent hyper-scene decomposition.

5. CONCLUSIONS

In this paper, we present a new method for video segmentation which allows the creation of a hierarchical hyper-scene decomposition of the sequence useful for applications such as video navigation. Each shot is represented using two 1-D mosaic images represented using a linear piecewise representation in order to simplify their manipulation. Furthermore, only the compressed MPEG-2 streams are assumed to be available, and the MPEG-2 motion vectors are used to build the mosaics in order to reduce the computation time. Global and motion-based distances have been proposed and are incorporated into a fine to coarse merging process. Experimental results, obtained on TV documentary sequences, show that the proposed method can be efficiently used in the context of video navigation.

ACKNOWLEDGEMENT

The authors would like to thanks IVC team of IRCCyN of Nantes (France) and Professor J. Benois-Pineau from Labri (Uni. of Bordeaux, France) for their collaboration in the context of the RNTL DOMUS-VIDEUM project for which this work has been done.

REFERENCES

[1] I. Koprinska and S Carrato, "Temporal Video Segmentation: A Survey", *Signal Processing : Image Communication*. V.16(2001), pp.477-500.

[2] A.W. Smeulders, M. Worring, S. Santini,A Grupta and R. Jain "content based image retrieval at the end of the early years", *Pattern Analysis and Machine Intelligence*, 22(12):1349-1380,2000.

[3] H. Nicolas "New Methods for dynamic mosaicking". *IEEE Transactions on Image Processing*, Vol. 10, No. 8, pp. 1239-1251, August 2001

[4] M. Irani and P. Anandan, Video indexing based on mosaic representation. *IEEE Trans. PAMI*, 86(5):905--921, May 98.

[5] W. Dupuy, J. Benois-Pineau and D. Barba, "Outils pour l'analyse et l'indexation vidéo basée sur l'approche du signal 1D dans le domaine de la transformée Mojette," *RFIA'2002*, Angers, France, pp 337-386, 8-10 January 2002

[6] E. Keogh and M.Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *Proc. of ICK DD*, pp 239-24 -241, AAAI Press

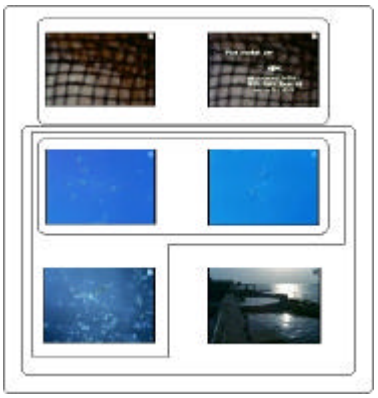

Figure 1: Merged order for shots 5, 158, 30, 32, 48 and 137

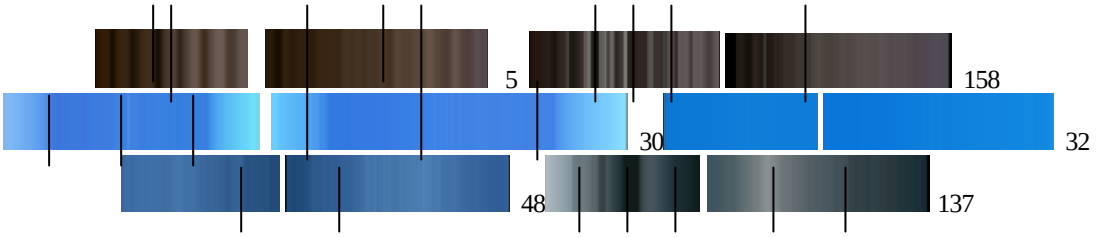

Figure 2: Vertical and horizontal mosaics, and shot number. Vertical black lines denote the breaking points of the LPR representation

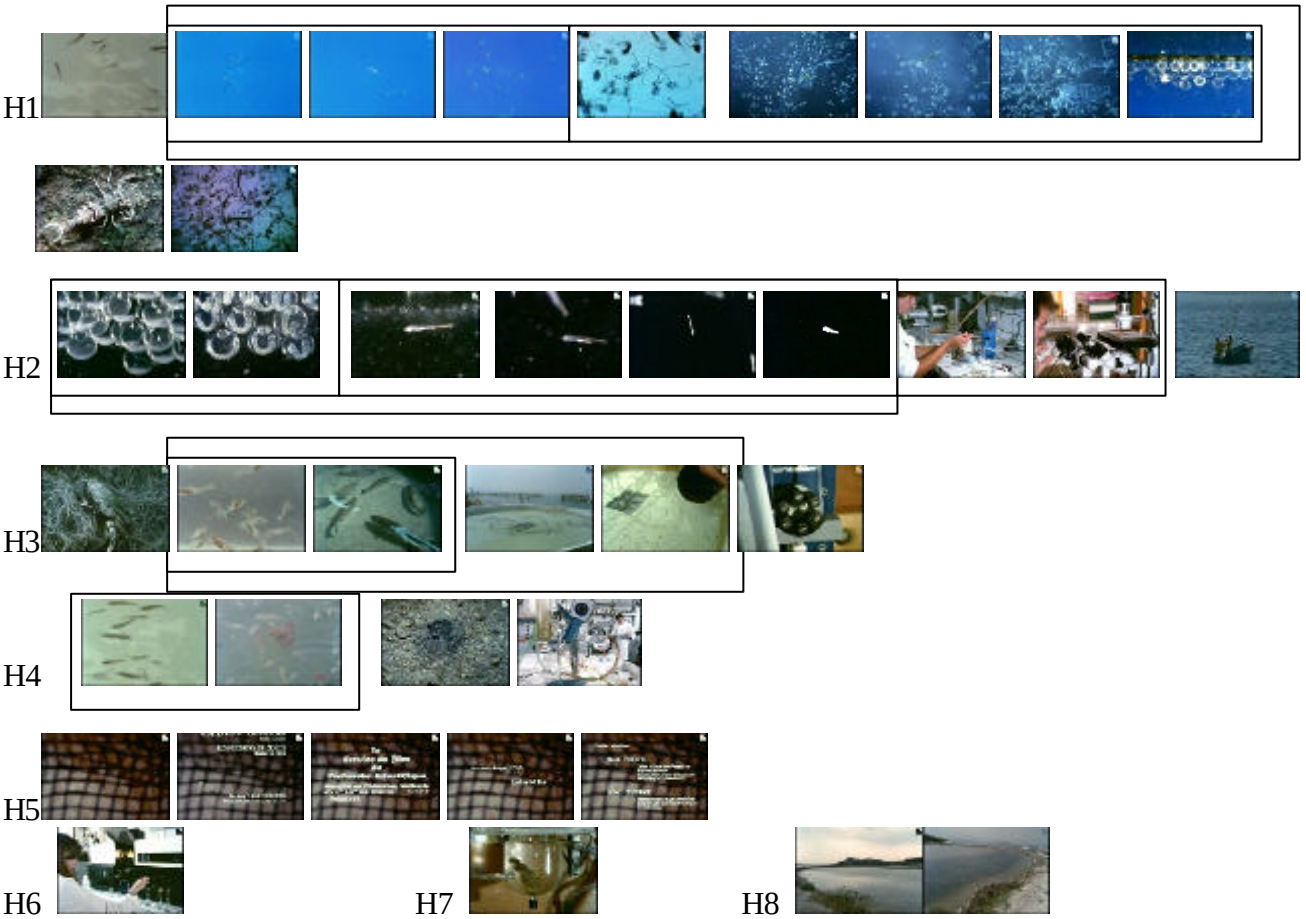

Figure 3. Some shots associated to 8 hyper-scenes created on sequence Aquaculture. The boxes indicate some hyper-scenes created at intermediate levels