GROUPING VIDEO SHOTS INTO SCENES BASED ON 1D MOSAIC DESCRIPTORS

H. Nicolas*, A. Manaury*, J. Benois-Pineau**, W. Dupuy***, D. Barba***

*IRISA/INRIA
Campus de Beaulieu
35042 RENNES Cedex
France
hnicolas@irisa.fr

**LABRI UMR n° 5800 CNRS
Université Bordeaux 1,
351 cours de la Libération
33405 TALENCE Cedex - France
jenny.benois@labri.fr

***IRCCyN UMR 6597 CNRS
Ecole Polytechnique de l'Université de Nantes
rue Christian Pauc - BP 50609
44306 NANTES Cedex 03- France
dominique.barba@polytech.univ-nantes.fr

ABSTRACT

This paper describes an original approach for structuring video documents into scenes by grouping video shots. The method is based on the construction of 1-D mosaics. 1-D mosaics are built based on X-ray projections of color video frames representing integration along vertical and horizontal axes. The mosaicing is realized by motion compensation in 1D domain. Grouping of shots in a scene is done by local and global matching of mosaics based on piecewise linear approximation and hierarchical clustering. The results obtained on feature documentaries are promising.

1. INTRODUCTION

The standard multimedia description interface MPEG-7 [1] requires a structural decomposition of video documents into segments which can be represented as semantic entities: objects, shots, scenes. Indexing should be done to characterize their content, movement, color, texture. The standard defines the video segment descriptors but not the way of performing the structural decomposition.

Most of the research in the field of video document structuring assumes a hierarchical decomposition of documents which shows such structural units as *scene*, *shot*, *object* or *event*. A set of techniques should be proposed to perform correctly these tasks. An example of such a system is the one developed in [2], which detects shot changes, extracts interesting objects and characterizes the camera motion from motion estimation. 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 can be described using a non linearly time dependent organized structure. For that purpose, consecutive and/or non-consecutive shots have to be merged according to a similarity criterion. The obtained clusters of shots are called here *hyper-scenes (HS)*.

Furthermore, this work has been designed in order to be used by a common user in his home environment where often only compressed video data streams are available. As a consequence, this paper proposes a complete method to define hyper-scenes using MPEG-2 compressed video data. This method performs as follows: the initial shot decomposition is assumed to be available. For each shot, descriptors based on 1-D mosaics are extracted using the method proposed in Section 2. Similarity criteria and a merging strategy are proposed in Section 3 to obtain the decomposition of content into hyper-scenes. The method has been validated on a corpus of artistic content as shown in Section 4.

2. VIDEO SHOT DESCRIPTORS

In the problem of grouping elementary video segments, such as video shots, into structural entities of higher order, such as hyper-scenes, the choice of descriptors expressing well the shots similarities is mandatory. In [3] we proposed a spatio-temporal signature which modeled a shot as a set of specific color vectors in an adapted space. This signature is based on a 1D discrete projective transform called "X-Ray image". The X-Ray image of a video frame was first introduced in Tonomura's VideoMap [4]. It is done by "projecting" an image (I (x,y) with $1 \le x \le w, 1 \le y \le h$) along straight lines either vertically or horizontally. This projection consists in the integration of image signal along these straight lines and in this sense, represents a discrete analogue of Radon transform used in tomography. Each projection is a vector with each element, or bin, being the mean of pixel values along the given direction. The value of an X-Ray vector element is:

$$I_x[I](y)=\frac{1}{w}\sum_{k=1}^{k=w} I(k,y) \text{ and } I_y[I](x)=\frac{1}{h}\sum_{k=1}^{k=h} I(x,k) \quad (1)$$

for horizontal and vertical projections, respectively. The spatial distribution of the image intensity in the direction of the axis orthogonal to the projection vector is preserved. This is an interesting property of this transform compared to usual histograms. Applying (1) to a color image, we will form a set of 3-component color bins, which represents a reduced (to a column or a line) version of a complete color video frame plane. Furthermore, mosaic descriptors are proposed in MPEG-7 to represent video segments such as shots [1]. Similarly, the 1-D mosaic descriptor can be proposed to represent the set of X-ray images obtained for each image of a shot. These mosaics then can be used to measure the similarity between shots. The problem to solve now is to define a good motion model and an adequate motion compensation method for the construction of such 1D mosaics.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

In case of 2D mosaics, methods for their construction from video frames such as in [5, 6], consist in compensating all video frames into the reference frame based on either optical flow or on a global motion model. In the 1D case of X-ray images the method would be similar. Therefore, the definition of an adequate 1D motion model is needed. 2D Affine models have proved to efficiently represent global motion. Many authors use a three parameters motion model [7] (with shift and zoom factors) despite its incompleteness, as it allows for characterization of most frequent situations in video such as pan, traveling, "zoom in" and "zoom out". Here the elementary displacement vector $(dx,dy)^T$ at each pixel location $(x,y)^T$ is expressed as

$$\begin{cases} dx = t_x + f(x - x_g) \\ dy = t_y + f(y - y_g) \end{cases} \quad (2)$$

where $(t_x,t_y)^T$ is a translation vector, f is a zoom factor, $(x_g,y_g)^T$ is a reference point, such as the center of a frame. This motion model is specifically in the focus of our attention since a simple relationship between the 2-D motion model in the image plane and the 1-D motion model in X-ray images can be obtained. In order to explain it we have to reply the general formulation of Radon transform [8]:

$$R\varphi[I](u) = \iint_D I(x, y)\delta(u - x\sin(\varphi) - y\cos(\varphi))dxdy \quad (3)$$

It represents an integration of a function $I(x,y)$ defined in R^2 along the straight line according to the projection angle φ, δ is the Dirac distribution. In case of video it is a Radon projection of each video frame. Considering a special case of discrete Radon transform, the so-called *Mojette* transform ($tan(\varphi) = -p/q$, where p and q are mutual prime integers), we showed in [3] the relation between model (2) and the 2-parameters model in Mojette domain. Denoting the coordinate of each bin of projections by m we have

$$d_m = t_m + f(m - m_g) \quad (4)$$

where d_m is an elementary displacement in transform domain, $t_m = -qt_x + pt_y$ is the transformed translation vector, and $m_g = -q.x_g + p.y_g$ is the transformed reference point. The zoom factor f remains the same as in 2-D case. Choosing $p=0$, $q=1$ and $q=1$, $p=0$ we obtain the relations between 2D and 1D motion models in vertical and horizontal directions. Thus, knowing the 3-parameter affine model in 2-D case, the 1-D motion models can be easily obtained. Various approaches for estimation of parametric motion models in video frames have been proposed [2, 6]. In the context of processing of already compressed (MPEG1,2,4) [9] content, it is advantageous to use the macro-block and block motion vectors to estimate global motion model. In [10] we developed a robust least-square estimator of global affine model from macro-block motion vectors. The model equation here is

$$Z = H\theta + V$$

with $\theta = (t_x,t_y,f)^T$, H the observation matrix according to the model (2), $Z = (dx_1,\ldots dx_N,dy_1,\ldots dy_N)^T$ the coordinates of macro-block motion vectors, and V the observation noise. Then the solution is obtained by weighted least square method as

$$\hat{\theta} = \left(H^T WH\right)^{-1} H^T WZ$$

Here W is a diagonal matrix of weights, where each diagonal element expresses the relevance of the current measure to the model. It is weak for outliers and allows for labeling macro-blocks in the image domain. The macro-blocks outliers will not be used for mosaic computation.

We can now built a mosaic for each direction by integrating all vertical and horizontal X-ray projections. To do this we have to calculate the coordinates of a bin m_i from the coordinate system of the current X-ray projection $I_x[I_i]$ and $I_y[I_i]$ to the coordinate system of the reference frame I_j (in which the bin will be called m_j). Based on Equation (4) and in the case of projection in the motion estimation direction, it will be computed as

$$m_j = m_i\left[1 + \sum_{k=i}^{j}\left[fm_k\prod_{l=i}^{k}\left(1 + fm_{n-l}\right)\right]\right] + \sum_{k=i}^{j}\left[\left(tm_k - fm_k * mg\right)\prod_{l=i}^{k}\left(1 + fm_{n-l}\right)\right]$$

4. MERGING PROCESS

Based on the two 1D mosaics associated to each shot, inter-shot distances have now to be defined in order to create hyper-scenes. This requires to define similarity criteria designed here according to the following principles:

- Detection of shots which have approximately the same color characteristics.
- Detection of shots corresponding to the same physical scene, but which may have been acquired with a moving camera.

Based on these principles, 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 n segments (the length of each segment is constant for a given mosaic, but differs from two mosaics if their lengths are not the same). The distance between two 1-D mosaics is then obtained by matching each segment of the first mosaic to a segment of the second one such as establishing a bijection, between the n segments of each mosaic, which minimizes the distance defined here as the difference between their average values $\overline{S}_i$. The global inter-mosaic distance is therefore expressed as:

$$GD(M_1,M_2) = \sum_{i=1}^{S}\left(\left|\overline{S}_1^{i,h} - \overline{S}_2^{i,h}\right| + \left|\overline{S}_1^{i,v} - \overline{S}_2^{i,v}\right|\right)$$

where 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 compensation of the camera displacement. The motion model takes into account zoom and shifting camera displacements in the 1-D space (see Eq. 2). 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:

$$ME(M_1, M_2) = \arg\min_{t,k} \frac{1}{Card[N_{1,2}]}$$
$$[\sum_{p \in N_{1,2}} |M_1^h(p) - M_2^h(p + d(t,f))| + \sum |M_1^v(p) - M_2^v(p + d(t,f))|] \quad (5)$$

where $N_{1,2}$ denotes the number of overlapped pixels after motion compensation. *d(t,f)* is the displacement vector computed with the model (4).

In practice the minimization is performed using a full search matching algorithm with a quarter pixel precision for translation vector *t*, and 0.005 precision on zoom *f*. A 1-D mosaic image represents an amount of data significantly lower than 2D mosaics. Nevertheless, since many comparisons have to be done, it is interesting to further reduce this amount of data. This is done here by representing a 1D mosaic as a set of linear segments computed by a piece-wise linear approximation of a mosaic function *M(m)*. With this approximation, the error terms in the equation (5) can be analytically computed which leads to a significantly faster 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:

$$D(M_1, M_2) = \min[GD(M_1, M_2), \alpha ME(M_1, M_2)]$$

where α is a weight coefficient. Its value depends on the user and the application constraints. Similarly, 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. The shot and/or *HS* merging process are based on a fine to coarse recursive approach. At each iteration, the couple of shot and/or *HS* which has the lowest distance is merged. The operator may fixed the final number of *HS* (which can be only 1), but it has the possibility to go to any intermediate levels if some of the obtained *HS* are not sufficiently coherent.

4. 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 shows the two 1-D mosaic images for six shots, and the breaking points of the linear segment representation. Figure 2 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. REFERENCES

[1] ISO/IEC JTC 1/SC 29/WG 11/M6156, MPEG-7 Multimedia Description Schemes WD (Version 3.1), Beijing, July 2000.

[2] P. Bouthemy, M. Gelgon, F. Ganausia. "A unified approach to shot change detection and camera motion characterization". IEEE Trans on CSVT, Vol. 9, n°7, pp. 1030-1044, October 99.

[3]. J.Benois-Pineau, W.Dupuy, D.Barba. "Re-covering of visual scenarios in movies by motion analysis and grouping spatio-temporal colour signatures of video shots". Invited paper at EUSFLAT'2001, special session on "Data Mining and Multimedia Systems", Leicester, pp. 385-389, September 5-7, 2001.

[4] Y. Tonomura, A. Akutsu, K. Otsui, T. Sadakata. "VideoMap and VideoSpaceIcon : tools for anatomizing video content". Proc. InterChi'93, ACM, 1993, pp131-136.

[5].Irani M., Anandan P. et al. "Efficient representation of video sequences and their application". Signal Processing: Image Communication, Vol. 8., 1996, pp. 327-351.

[6] H. Nicolas. "New Methods for Dynamic Mosaicing". IEEE Transactions on Image Processing, Vol. 10, No. 8, pp. 1239-1251, August 2001.

[7] P.Joly, H.-K.Kim. "Efficient automatic analysis of camera work and microsegmentation of video using spatio-temporal images ". *Signal Processing : Image Communication ,* 8, pp. 295-307, 1996.

[8] B.Jähne. "Spatio-temporal Image Processing. Theory and scientific applications", Lecture notes in Computer Science 751, pp. 92-93, Spronger – Verlag, 1993.
[9] ISO/IEC JTC1/SC29/WG11 N2202. Information technology-Coding of audio-visual objects: Visual. ISO/IEC 14496-2 Committee Draft (MPEG4: Visual). Tokyo, March 1998.
[10] M. Durik, J. Benois-Pineau. "Robust Global Motion Characterisation for Video Indexing Based on MEPG2 Optical Flow". CBMI, Brescia, Italy, pp. 57-64,19-21 September 2001

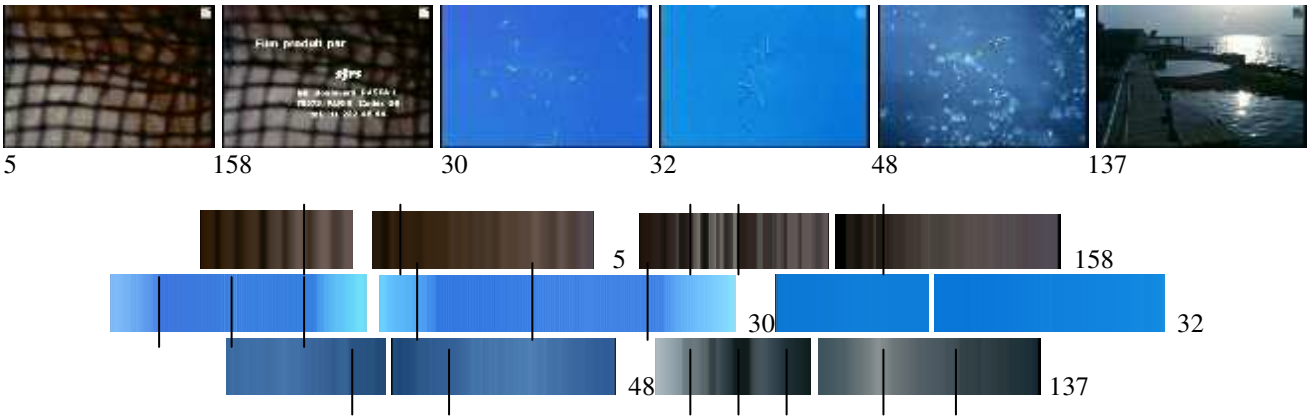

Figure 1: Key frames of some shots. For each shot: vertical and horizontal mosaics. Vertical black lines denote the breaking points of the linear segment decomposition.

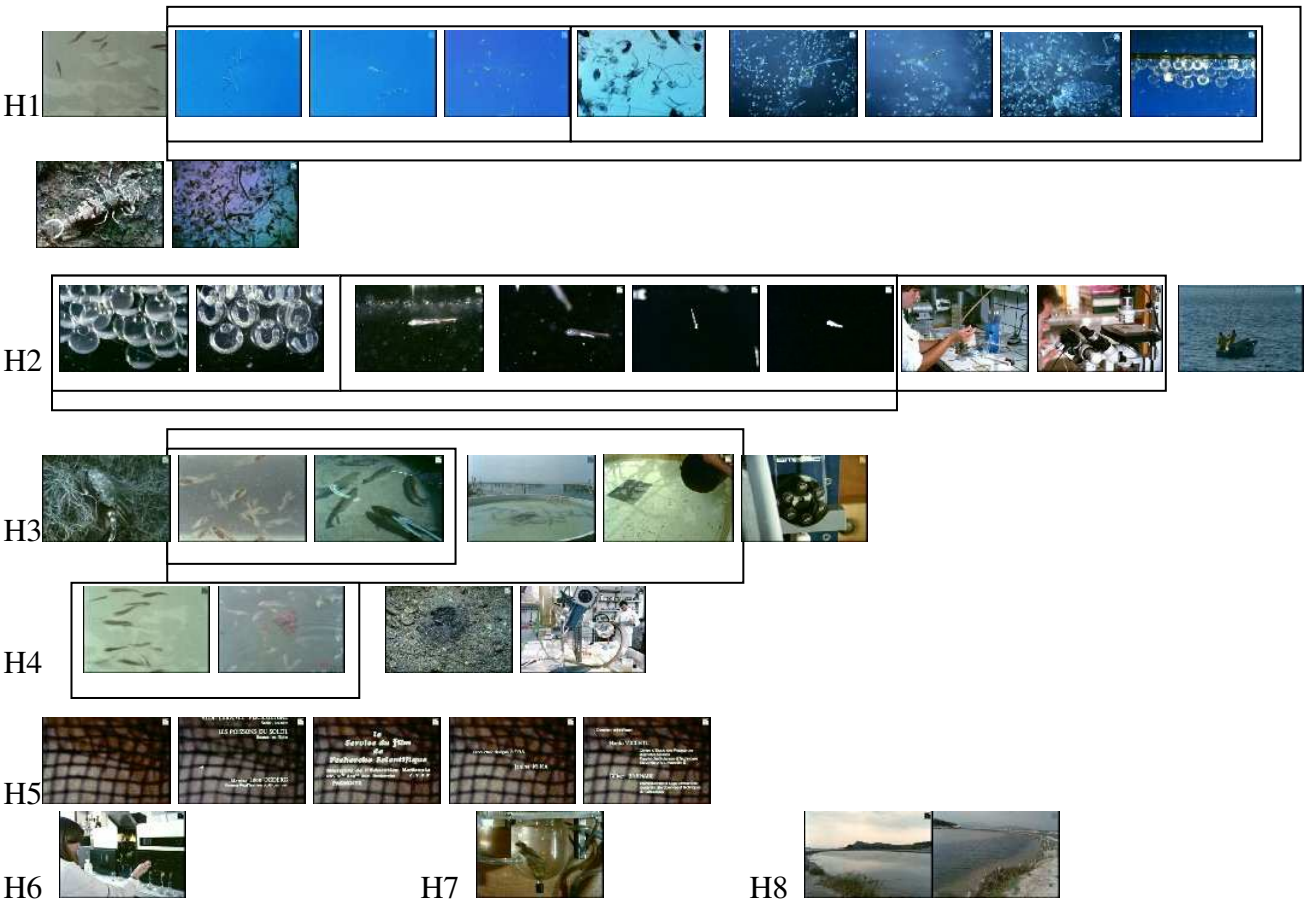

Figure 2: Some shots associated to 8 hyper-scenes created on sequence *Aquaculture*. The boxes indicate some hyper-scenes created at intermediate levels.